# THE COMPARISON OF COMMON ITEM SELECTION METHODS IN VERTICAL SCALING UNDER MULTIDIMENSIONAL ITEM RESPONSE THEORY

By

Yang Lu

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Measurement and Quantitative Methods

2010

# ABSTRACT

THE COMPARISON OF COMMON ITEM SELECTION METHODS IN VERTICAL SCALING UNDER MULTIDIMENSIONAL ITEM RESPONSE THEORY

By

Yang Lu

The characteristics of common items are always considered as an important factor affecting the quality of scale linking between tests. Although many studies have focused on the common item selection via the Unidimensional Item Response Theory (UIRT) models, seldom did researchers investigate the selection in the vertical scaling under the framework of Multidimensional Item Response Theory (MIRT). This study examines different common item selection methods when the correlation among proficiencies varies at different levels and when the content structures in tests are either identical or different. With respect to the recoveries of the probability matrix, item parameters and effect sizes, the results show that (1) full content coverage in the common item set is important, no matter whether the content structures are identical or not, (2) high correlation among proficiencies could partly compensate for the adverse effect caused by the common items not covering all content domains, and (3) the common item set covering all content domains with medium difficulty items yields better linking results and common items with high item-total-test correlation also perform well when the content structures are identical for both tests.

*Dedicated to my beloved husband: Yu Fang*

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Dr. Mark Reckase, who is my academic advisor and chairperson of my dissertation committee. I am fortunate to have the privilege to be his student. I want to thank him for his excellent guidance and patience in my dissertation and research work. I also want to thank Dr. Sharon Senk, one of my committee members, for strengthening my background in math education field and giving me a warm hug when I felt frustrated in working on the project. Thanks also go to other members in my committee, Dr. Richard Houang and Dr. Sharif Shakrani, for their valuable insights and assistance on my dissertation research.

I would like express my appreciation to Dr. Maria Teresa Tatto for providing me with assistantship opportunities to work in the TEDS-M project for most of my graduate study. Through this experience, I have developed a deep understanding in measurement theory and improved my teamwork skills.

In addition, I want to thank the Measurement and Quantitative Methods program and the College of Education at the Michigan State University for providing me with an excellent atmosphere for my graduate study.

Finally, I owe my deepest appreciation to my husband, Yu Fang, for his continuous love, support and patience, and to my parents for their constant love and support in all aspects.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Standardized tests have been more and more important and widely applied in performance assessment. They are used to provide fair, reliable and objective information on examinees' abilities or skills that the tests are developed to measure. When educational achievement is assessed, it is important to estimate and track the extent to which examinees grow over time or over the course of their schooling.

One method to evaluate examinees' progress is to use a single set of test questions or equivalent forms for all the assessments across time. But this method could be problematic, since the shift in content and the challenge of materials may be too huge to be appropriate for all grade levels. For example, a test containing many items from higher grade levels could be too difficult or too advanced with brand new topics for the students at the early grade levels, so as to make them feel overwhelmed by the test. On the other hand, a test may also be too easy and lead to carelessness, inattention or creative but wrong thinking when items from lower grade levels are administered to the students in upper grades (Kolen & Brennan, 2004, p. 372).

The above problems could be avoided by administering the tests of different levels to students from different grades and using vertical scaling method to link these tests. Practically, vertical scaling is widely used in standardized tests. For example, for No Child Left Behind (NCLB), vertical scales are established to meet the requirement of evaluating the progress of children in attaining English proficiency as they grow from one year to the next. Vertical scales are also commonly used in several elementary test batteries, such as the Iowa Tests of Basic Skills (ITBS) (Hoover, Dunbar, & Frisbie, 2003) and the Cognitive Abilities Test (CogAT) (Lohnman &

Hagen, 2002). According to Harris (2007), the scores from tests in some testing systems, such as those from the EXPLORE, PLAN and ACT tests in the Educational Planning and Assessment System (EPAS), are also put on one vertical scale, and it is clearly stated that the target populations for these tests are students at different grade levels.

## 1.1 Introduction to Vertical Scaling

The primary reason to construct vertical scales is to "develop a conceptual definition of growth, especially for test areas that are closely related to the school curriculum" (Kolen & Brennan, 2004, p. 376). Among all tests, the mathematics achievement test is one of those covering several content areas that are closely related to the school curriculum. Some content domains are taught with different difficulty levels in different grades while others are not. As the grade level increases, curriculum contents become more advanced and new contents are added as well; consequently, test items become more difficult and items measuring new constructs are also included in the test.

With the increase of the depth or amount of the contents that students have been taught, it is natural to ask how much students gain in knowledge according to test scores from different tests administered in different years. How could these changes be measured by the scores from time to time? Could the scores from different grades be compared directly? Did the students grow as much in one content area as in another one?

Vertical scaling is one of the commonly used methods to answer these questions. There are many definitions for vertical scaling. According to Kolen and Brennan (2004) and Holland (2007), different from equating, vertical scaling tries to place tests with different difficulty but measuring similar constructs on the same scale. Kolen (2006) indicated that "vertical scaling procedures are used to relate scores on these multiple test levels to a developmental score scale that can be used to assess student growth over a range of educational levels". More specifically, vertical scaling focuses on the linking between tests with similar reliability and measuring similar construct, but with different difficulty and administered to different populations of examinees (Holland, 2007).

## 1.2 Construct Shift and Dimensionality in Vertical Scaling

There are many issues when the vertical scaling procedure is applied, one of which is the way to define the overlap of content structure across grades. Harris (2007) indicated that the relationship between the test content and the nature of growth plays a major role on the resulting score scales. It is generally easy to compare gains when the constructs measured by tests are fairly similar from year to year. In this situation, the linking of score scales can give sufficient and meaningful results. However, since the curriculum and instructions often change from grade to grade in practice, test designers need to modify the test content specifications to match the targets of instructions in order to accurately assess achievements at different grades.

Normally, the skills and knowledge included in instructions are not very simple. For example, the mathematics test could cover different content areas, such as number, algebra, arithmetic and geometry. Furthermore, a variety of skills, such as problem solving, logical thinking and reading and understanding are often required to solve some mathematics problems as well. The higher the grade level is, the more complex the covered contents and required skills in the test are.

This multidimensionality issue was addressed by Yen (1986), who pointed out that it is one of the major factors that are likely to affect the vertical scales. Several studies questioned the appropriateness of using only a single vertical scale to track students' growth from year to year since the instruction and curriculum change across years. Yen and Burket (1997) found that scales always vary by subjects and subtests within the subject. Martineau (2004, 2006) also showed the significant effect of construct shift when a single vertical scale was used in the value added model.

Li (2006) made a further study to capture the cross-grade content shift based on the MIRT, and found that the two constructs (vocabulary and problem solving) were overlapping and measured by both Grade 6 and Grade 7 Michigan Educational Assessment Program (MEAP) tests, while an additional construct (abstracting concept) was only measured by the Grade 7 test. Note that the tests in MEAP actually measure the students' learning of previous academic year. Patz and Yao (2007) also indicated that using the unidimensional IRT model is implausible when vertical scales are developed across grades. Specifically, they noticed that the construct measured by the seventh-grade mathematics achievement test was different from that by the fourth-grade test. They also pointed out that failure to account for the complexity of the large differences in test content and examinee skills could be an important reason that concurrent calibration using the unidimensional IRT model did not perform well in practical settings. What is more, some other studies (Braun, 2005; Doran & Cohen, 2005; Reckase & Li, 2007; Reckase & Martineau, 2004; Schmidt, Houang, & McKnight, 2005) also addressed the issues of content shift and the inappropriateness of using one single scale to track students' growth.

On the other hand, many studies discussed the effect of violation of the unidimensionality assumption in vertical scaling (e.g., Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985; Turhan, Tong, & Um, 2007; Yao & Mao, 2004). More specifically, using simulated test scores across grade levels, Yao and Mao (2004) found that the score distribution estimated under one-, two- or three-dimensional models did not differ significantly; however, they did not show how stable the actual dimensional structures of cross-grade tests were. Turhan et al. (2007) simulated data with the MIRT model and tested different ways of selecting common items in vertical scaling after the simulated data were calibrated with the commonly used unidimensional IRT model. They concluded that vertical scaling was robust to the types of slight violation of the unidimensionality assumption investigated in their paper, given the goodness of the content coverage by common items.

Because of these two factors, namely, the modification of content area and curriculum across grades and the multidimensionality in the tests, a single scale score for the tests often becomes less comparable across grades. These factors can influence the interpretation and alignment of vertical scales; hence, it is important to check the content shift and dimensionality for the constructs measured in the tests and select appropriate vertical scaling methods to make the scores more comparable. Since the MIRT model identifies the multidimensional content structure in the test and estimates all dimensions simultaneously, it is a promising method for vertical scaling among the tests measuring shifted and/or multiple constructs.

## 1.3 UIRT and MIRT

### 1.3.1 UIRT model

Birnbaum (1968) developed a three-parameter logistic unidimensional IRT model, which assumes that only one latent trait is necessary to account for variations in person-item responses and is widely used in test construction and equating. The formula for this model is

$$P(u_{ij} = 1 \mid a_i, b_i, c_i, \theta_j) = c_i + \frac{1 - c_i}{1 + \exp(-1.7 a_i (\theta_j - b_i))} \ , \qquad (1.1)$$

where $P(u_{ij} = 1 \mid a_i, b_i, c_i, \theta_j)$ is the probability of person *j* correctly answering item *i*, given that the person's ability level is $\theta_j$, and $a_i, b_i, c_i$ represent item discrimination, difficulty and guessing parameters, respectively.

### 1.3.2 MIRT model

Multidimensional abilities/traits are often required to get correct responses for items within one test, or even a correct response for one item (Reckase, 1985). Hence, theoretically, it is more appropriate to use the multidimensional IRT model instead of the unidimensional IRT model for the calibration and estimation of the aforementioned multidimensional data. Since “a set of test items can be sensitive to several traits, or a group of examinees might vary in several latent traits” (Li & Lissitz, 2000), the person-item interaction and parameter estimation can be quite complex in MIRT.

According to Reckase (1997a), MIRT is useful to (1) understand the proficiency structure needed to respond to test items, (2) describe the differential item functioning (DIF), and (3) choose items to fit the unidimensional IRT model. There are two types of MIRT models, namely, the compensatory MIRT model and the noncompensatory MIRT model. The main difference between these two models lies in the relationship among the multiple proficiencies that determine the probability of person-item responses. The compensatory model follows the logic of factor analysis in that the probability of a correct response is related to a linear combination of several proficiencies; therefore, proficiencies are additive so that high proficiency on one dimension can ‘make up’ the low proficiencies on other dimensions. The three-parameter compensatory MIRT model (Reckase, 1985, 1997b) is

$$P(u_{ij}=1\,|\,\mathbf{a_i},d_i,c_i,\boldsymbol{\theta}_\mathbf{j})=c_i+\frac{1-c_i}{1+\exp(-1.7(\mathbf{a_i}'\boldsymbol{\theta}_\mathbf{j}+d_i))}\ , \qquad (1.2)$$

where $P(u_{ij}=1\,|\,\mathbf{a_i},d_i,c_i,\boldsymbol{\theta}_\mathbf{j})$ is the probability of a correct response for person *j* on item *i*, $u_{ij}$ is the response for person *j* on item *i* (1 if correct and 0 otherwise), $\mathbf{a_i}$ is an *m*-element vector that specifies the discrimination power of item *i* on the *m* dimensions, $d_i$ is a scalar parameter that is related to the difficulty of item *i*, $c_i$ is a guessing parameter for item *i*, and $\boldsymbol{\theta}_{\mathbf{j}}$ is the person *j*'s proficiency vector in an *m*-dimensional space.

On the other hand, Sympson (1978) proposed a noncompensatory MIRT model,

$$P(u_{ij}=1\,|\,\mathbf{a_i},\mathbf{b_i},c_i,\boldsymbol{\theta}_{\mathbf{j}})=c_i+\prod_{k=1}^{m}\frac{1-c_i}{1+\exp(-1.7a_{ik}(\theta_{jk}-b_{ik}))}\ , \qquad (1.3)$$

where $a_{ik}$, $b_{ik}$ and $\theta_{jk}$ are the item discrimination, item difficulty and person proficiency on the *k*th dimension, and $c_i$ is the item guessing parameter. He argued that an increase in one proficiency could improve the overall probability of getting an item correct, but only to some extent. The probability of a correct response cannot exceed that defined by the dimension where the proficiency does not have a positive infinity value, even when all other proficiencies increase to positive infinity.

The compensatory and noncompensatory MIRT models are quite different, from either the mathematical formula or the assumption on how people use their skills and knowledge to answer items. However, Spray, Davey, Reckase, Ackerman and Carlson (1990) identified item parameters for the two models that can give similar classical item statistics, and found that when the correlation between proficiencies increases, the detectable difference between models decreases. Therefore, they concluded that the difference between these two models could be considered practically unimportant.

In practice, due to the comparative simplicity and easy estimation procedure, the compensatory model is the one commonly used in the MIRT calibration, scaling and equating. Thus, the compensatory MIRT model is used in this study, and for simplicity, no guessing parameter is assumed for items.

According to Reckase (1985), *MDISC* was developed to capture the discrimination power of an item in MIRT and its formula is given by

$$MDISC_i = (\sum_{k=1}^{m} a_{ik}^2)^{1/2} \ , \qquad (1.4)$$

where $MDISC_i$ denotes the item *i*'s multidimensional discrimination and $a_{ik}$ is the discrimination parameter of item *i* on the *k*th dimension. Also, the multidimensional difficulty of an item, *MDIFF*, is defined as

$$MDIFF_i = -\frac{d_i}{\sqrt{\mathbf{a_i}'\mathbf{a_i}}} = -\frac{d_i}{MDISC_i} \ . \qquad (1.5)$$

These two characteristics of an item can be represented graphically by an item vector in the multidimensional $\boldsymbol{\theta}$–space. In order to describe the most discriminating direction of an item in that space, Reckase (1985) proposed the direction cosine for the item vector as

$$\cos\alpha_{ik} = \frac{a_{ik}}{MDISC_i} \ , \qquad (1.6)$$

where $\alpha_{ik}$ is the angle between the vector of item *i* and the *k*th coordinate axis in an *m*-dimensional space.

Figure 1.1 shows these characteristics of item vectors using arrowed lines in a two-dimensional space. The length of the arrowed line represents *MDISC*, the distance from the origin to the base of the arrowed line is *MDIFF*, and the direction of the arrowed line is defined using angles from the direction cosines.

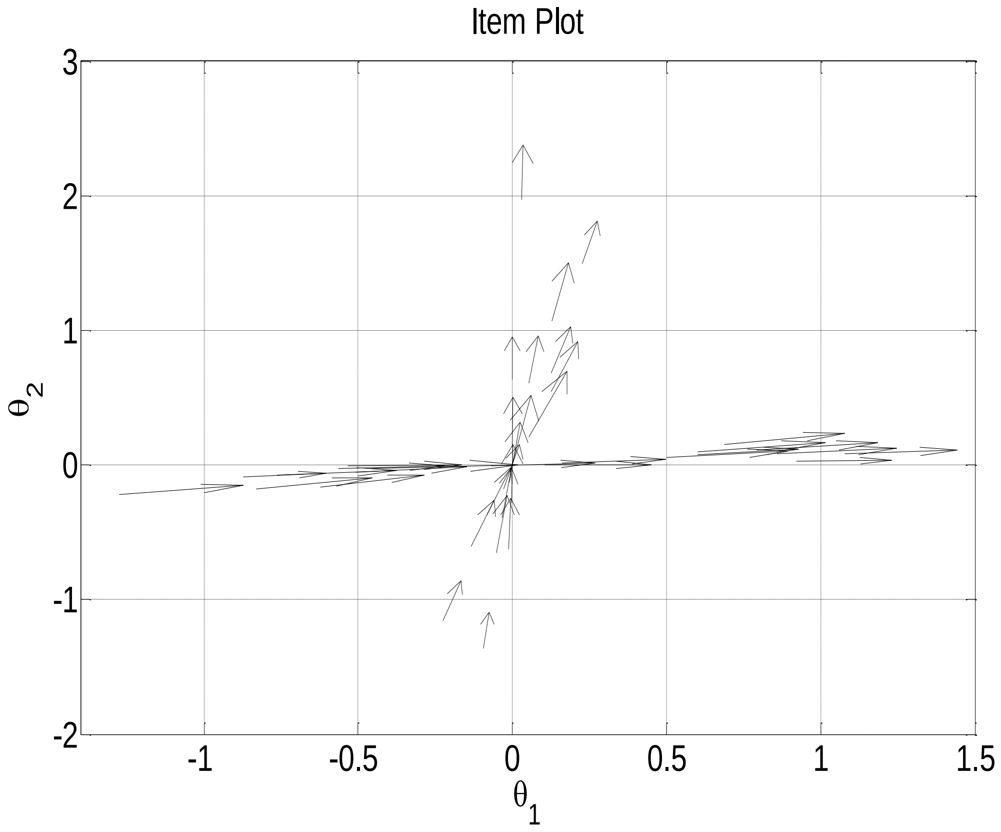

Figure 1.1. Representation of Item Vectors in a Two-Dimensional Space

### 1.3.3 Indeterminacies in MIRT

According to Reckase (1997a), the compensatory MIRT model could be considered as a special case of nonlinear factor analysis. More specifically, this model can be regarded as a combination of the factor analysis model and the unidimensional IRT model; therefore, it suffers from the indeterminacies inherent in either model, such as the orientation of coordinate axes relative to persons' locations, the units of measurement and the location of origin of the coordinate system. These three indeterminacies are named as rotational indeterminacy, unit indeterminacy and origin indeterminacy, which are often discussed in MIRT research (Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima, Davey, & Lee, 2000; Reckase, 2007; Reckase & Martineau, 2004).

Suppose $\mathbf{A} = (\mathbf{a_1}, \cdots, \mathbf{a_N})', \mathbf{d} = (d_1, \cdots, d_N)', \mathbf{\Theta} = (\mathbf{\theta_1}, \cdots, \mathbf{\theta_J})$ is one solution set for the MIRT calibration, where *N* is the total number of items and *J* is the total number of persons. There are always infinite sets of $\mathbf{A^*}, \mathbf{d^*}, \mathbf{\Theta^*}$ as defined in Equation 1.7, which satisfy the MIRT invariance property as shown in Equation 1.8.

$$\mathbf{\Theta^*} = \mathbf{T}^{-1}\mathbf{\Theta} - \mathbf{M1'}\ ,\ \mathbf{A^*} = \mathbf{AT}\ ,\ \mathbf{d^*} = \mathbf{d} + \mathbf{ATM}\ , \tag{1.7}$$

$$\mathbf{A^*\Theta^*} + \mathbf{d^*1'} = \mathbf{AT}(\mathbf{T}^{-1}\mathbf{\Theta} - \mathbf{M1'}) + (\mathbf{d} + \mathbf{ATM})\mathbf{1'} = \mathbf{A\Theta} + \mathbf{d1'}\ , \tag{1.8}$$

where **T** is a rotation matrix, **M** is a transformation vector and **1** is an *N*-element vector of 1s. Note that the rotational indeterminacy and unit indeterminacy are combined together as the **T** matrix in the above formulas.

Generally, for easy computation, the MIRT software packages provide one solution for the MIRT calibration by setting the constraints on the person proficiencies, or more strictly speaking, the coordinates of person locations in an *m*-dimensional space, as $E(\mathbf{\theta}) = \mathbf{0_{m \times 1}}$ and $\text{cov}(\mathbf{\theta}) = \mathbf{I_{m \times m}}$, although this zero correlation among proficiencies is implausible in practice. Besides these constraints, the software may also use the Varimax method to change the relative positions between item vectors and coordinate axes for a better interpretation of item characteristics.

### 1.3.4 Full information factor analysis in MIRT

TESTFACT is one of the software packages for MIRT calibration (Bock, Gibbons, Schilling, Muraki, Wilson, & Wood, 2003). In this package, full information factor analysis is used for the item and person parameter estimation, and this method uses all the information available in the matrix of dichotomously scored responses. Another software package, NOHARM (Fraser, 1988), uses the aggregated information, which is the product-moment matrix, for the MIRT parameter estimation.

Based on the local independence assumption that the examinee's responses to test items are statistically independent conditional on the examinee's ability, the probability for person *j* with the proficiency value $\boldsymbol{\theta}_\mathbf{j}$ to get a certain response pattern is

$$P(\mathbf{u_j} \mid \mathbf{A}, \mathbf{d}, \boldsymbol{\theta}_\mathbf{j}) = \prod_{i=1}^{N} P_{ij}^{u_{ij}} (1 - P_{ij})^{1-u_{ij}} \quad , \qquad (1.9)$$

where *N* is the total number of items in the test, $\mathbf{u_j}$ is the response vector of size *N* for person *j*, $u_{ij}$ is the response for person *j* on item *i* , and $P_{ij}$ is defined by the aforementioned compensatory MIRT model. Hence, by incorporating the prior assumption on the distribution of person proficiencies, the marginal probability for person *j*'s response pattern is

$$P(\mathbf{u_j} \mid \mathbf{A}, \mathbf{d}) = \int P(\mathbf{u_j} \mid \mathbf{A}, \mathbf{d}, \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta} \quad , \qquad (1.10)$$

where $g(\boldsymbol{\theta})$ is the pre-assumed distribution of person proficiencies.

With the similar local independence assumption among persons' response strings in the MIRT, the joint probability for the person-by-item data matrix can be obtained by multiplying the probability of each person's response string across persons. Thus the marginal likelihood function for all persons on all items is

$$L(\mathbf{U} \mid \mathbf{A}, \mathbf{d}) = \prod_{j=1}^{J} P(\mathbf{u_j} \mid \mathbf{A}, \mathbf{d}) \quad . \qquad (1.11)$$

Then the TESTFACT software package sets the constraint as $E(\boldsymbol{\theta}) = \mathbf{0}_{\mathbf{m}\times\mathbf{1}}$ and $\operatorname{cov}(\boldsymbol{\theta}) = \mathbf{I}_{\mathbf{m}\times\mathbf{m}}$, and applies the Expectation-Maximization (EM) algorithm to maximize this marginal likelihood function (Bock, Gibbons, & Muraki, 1988).

The initial values of slopes for the EM algorithm are calculated from the factor loadings, which result from the principal factor analysis on the tetrachoric correlation matrix among item responses. Then they are rotated orthogonally with the Varimax criterion to serve as starting values if the Varimax or Promax rotation option exists in the TESTFACT command syntax. There is concern that the starting values for slopes may be negative for any dimension due to the rotational indeterminacy in factor analysis.

With these starting values, the item parameter estimates can be obtained after the EM cycle converges, which generally means that the change in parameter estimates between adjacent cycles is less than some predefined value. Then these estimates are regarded as fixed parameters and person proficiency estimates are calculated under the Bayesian framework by incorporating prior information on its distribution. There might be many or all negative *a*-parameter estimates on certain dimensions in the calibration result, which is most likely due to the defaults used in TESTFACT for the MIRT calibration. It is reasonable and legitimate to change the sign of all estimates on these dimensions for a better interpretation. For the scoring option in the TESTFACT software package, both MAP (Maximum A Posteriori) and EAP (Expected A Posteriori) scores can be requested, and only the latter is used in this study. The reason to use the EAP score is that compared with MAP and MLE (Maximum Likelihood Estimate) scores, the EAP score is not only more stable and easy to compute without any iterative procedure, but also has "smaller mean square error over the population for which the distribution of ability is specified by the prior" (Bock & Mislevy, 1982). According to Muraki and Engelhard (1985), the EAP score is calculated with the posterior distribution of person proficiency by

$$\tilde{\boldsymbol{\theta}}_{\mathbf{j}} = E(\boldsymbol{\theta}_{\mathbf{j}} \,|\, \mathbf{u}_{\mathbf{j}})$$

$$= \int_{\boldsymbol{\theta}} \boldsymbol{\theta} P(\boldsymbol{\theta} \mid \mathbf{u_j}) d\boldsymbol{\theta}$$

$$= \int_{\boldsymbol{\theta}} \frac{\boldsymbol{\theta} P(\mathbf{u_j} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta})}{h(\mathbf{u_j})} d\boldsymbol{\theta}, \qquad (1.12)$$

and
$$h(\mathbf{u_j}) = \int_{\boldsymbol{\theta}} P(\mathbf{u_j} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $P(\mathbf{u_j} \mid \boldsymbol{\theta})$ is the probability function defined in Equation 1.9, $g(.)$ is the prior distribution of person proficiencies, $h(.)$ is the marginal probability for the response string $\mathbf{u_j}$, and $P(\boldsymbol{\theta} \mid \mathbf{u_j})$ is the posterior distribution, which is the conditional density for $\boldsymbol{\theta}$ given the response vector $\mathbf{u_j}$.

# CHAPTER 2

# LINKING AND COMMON ITEM SELECTION

## 2.1 Linking Designs

Vertical scaling is well known to be very important in tracking students’ growth over time. The purpose for vertical scale construction is to develop a common score scale across grades. When the vertical scale is to be constructed, numerous decisions need to be made, one of which is the design for data collection (Yen & Burket, 1997). Since the scaling design greatly determines the quality of the collected data, no matter how properly the measurement model, the calibration or linking method is used, if the data collection design is not appropriate, the resulting vertical scales will not be correctly constructed among multiple grades (Kolen & Brennan, 2004).

As is well known in the educational measurement field, horizontal equating is used to adjust the difference in difficulty among forms that are built to be similar in difficulty and content (Kolen & Brennan, 2004), while vertical scaling aims to link scales among forms that are built to be different in difficulty and to be administered to students of different levels. Different from those in horizontal equating, the test forms used in vertical scaling, which are most likely administered to adjacent grade levels, are not parallel or equivalent. Therefore, an equating of forms is not requested in the vertical scaling procedure; instead, the test linkage is more crucial and the strength of the link would affect the validity of the inferences based on such linkages (Patz & Yao, 2007).

According to Kolen and Brennan (2004), there are several ways to design the linkage for the tests from different grade levels. For example, in the equivalent groups design, through spiraling of test forms, examinees are randomly chosen to be administered the test designed for their own grade or for the adjacent lower grade.

In the test scaling design, the scaling test consists of items covering the contents across all grade levels. Since off-grade items may be too easy or too difficult for students in different grades, special instructions are needed to advise students to do their best. This design is not often used in practical settings, because it requires the construction of a complicated scaling test with appropriate length and the content areas covered by the scaling test may be too broad to be appropriate for all students in different grades.

The common item design is different from the above two designs and has its own characteristics. This design identifies the overlapping structure of the tests between adjacent grades. In addition to the appropriately designed items for the test at each grade, a set of common items, also called anchor items, are included in the tests of both grades in order to link the scale from one grade level to the next. These common items could provide basic statistical inferences for linking tests with similar construct and reliability so that scaled scores from both tests are comparable without the assumption of group equivalence.

Although this common item design is not difficult to implement, there are many practical issues that need to be considered when this design is used in vertical scaling, especially under the MIRT framework. For example, there is no general rule on the number of common items needed for an adequate vertical scaling. Also, the characteristics of items have not been clarified when these items are selected as common items for vertical scaling. Although there are some rules for the above questions in horizontal equating, the relationship between the common items and the tests to be linked has not been fully examined yet in vertical scaling, let alone under the MIRT framework.

## 2.2 Linking Methods

For the linking under the common item design, item parameters estimated from different test forms can be put onto the same scale with either separate calibration or concurrent calibration. When the concurrent calibration is used, item responses for all grade levels are formatted for a single computer run, with missing item responses coded as “not presented”. For the separate calibration, numerous research studies have been conducted in both UIRT and MIRT fields.

### 2.2.1 Linking methods in UIRT

In the unidimensional IRT model, the probability of a correct answer mainly depends on the linear combination of item discrimination parameter, item difficulty parameter and person proficiency parameter in the exponent of the model. When the UIRT model holds, given the probability of a correct response, a proper linear transformation of the proficiency scale can result in a consistent transformation of item parameter scale. That is to say, if the proficiency is linearly transformed from *Y* scale to *X* scale, which is $\theta_X = A\theta_Y + B$, item parameters can then be transformed as following so that the model can produce exactly the same fitted probabilities.

$$a_X = \frac{a_Y}{A} \text{ , } b_X = Ab_Y + B \text{, and } c_X = c_Y \text{ .} \tag{2.1}$$

Note that guessing parameters are independent from the scale transformation.

The above is called the unit and origin indeterminacies in the UIRT model. Due to these indeterminacies, the software packages for the UIRT calibration often provide item and person parameter estimates based on the constraints that $E(\theta) = 0$ and $\text{var}(\theta) = 1$.

Therefore, if two forms under the common item design are separately calibrated, a linear transformation is needed to put the two sets of estimates onto the same scale using the assumption that the common items in both forms have the same item parameters, so as to capture the proficiency difference between groups. In practice, this is often done by putting item parameters estimated from the new form on the scale of the old form or base form.

Generally speaking, there are four methods for the scale transformation in UIRT: the mean/mean method and mean/sigma method that belong to the moments methods, and the Haebara method and Stocking-Lord method that belong to the characteristic curve methods.

In the mean/mean method (Loyd & Hoover, 1980), the *A* and *B* parameters for the scale transformation are computed as

$$A = \frac{\mu(\hat{a}_Y)}{\mu(\hat{a}_X)} \text{ and } B = \mu(\hat{b}_X) - A\mu(\hat{b}_Y) \ . \qquad (2.2)$$

On the other hand, in the mean/sigma method (Marco, 1977), the *A* parameter is estimated via the standard deviations of difficulty in both forms by

$$A = \frac{\sigma(\hat{b}_X)}{\sigma(\hat{b}_Y)} \text{ and } B = \mu(\hat{b}_X) - A\mu(\hat{b}_Y) \ . \qquad (2.3)$$

The mean/mean and mean/sigma methods described above do not consider all item parameters simultaneously (Kolen & Brennan, 2004). Haebara (1980) and Stocking and Lord (1983) avoided it by using the item or test characteristic curves to estimate the transformation.

The Haebara method considered the difference between the item characteristic curves of common items, and for examinees of a particular proficiency level $\theta_j$, the sum of the squared difference between the curves of each item is expressed as

$$Hdiff(\theta_j) = \sum_i [p_{ij}(\theta_j; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - p_{ij}(\theta_j; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi})]^2 \ . \quad (2.4)$$

Then this method finds *A* and *B* by minimizing the summation across all proficiency levels as

$$Hcrit = \sum_j Hdiff(\theta_j) \ . \qquad (2.5)$$

Comparatively, the Stocking-Lord method minimizes the sum of the squared differences between the two test characteristic curves across all proficiency levels as

$$SLcrit = \sum_{j}[\sum_{i} p_{ij}(\theta_j; \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - \sum_{i} p_{ij}(\theta_j; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi})]^2 \quad . (2.6)$$

### 2.2.2 Linking methods in MIRT

Several researchers (Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Reckase, 2007; Reckase & Martineau, 2004) proposed the scaling methods based on separate multidimensional IRT calibrations. Oshima et al. (2000) developed several methods, which are extensions from the unidimensional IRT linking methods (e.g. the Haebara method and the Stocking-Lord method), to obtain (1) the rotation matrix to simultaneously adjust the orientation of coordinate axes and the variances of proficiencies, and (2) the translation vector to adjust the means of proficiencies. The scaling procedure in Li and Lissitz (2000) was carried out through an orthogonal Procrustes rotation matrix and a central dilation constant obtained by minimizing the sum of squared errors between the estimated item discrimination matrix from the base form and the transformed one from the alternate form, and a translation vector obtained by the least squares method of minimizing differences between the difficulty estimates from the base form and the transformed ones from the alternate form. Min (2003) improved this procedure by using a dilation matrix in lieu of the central dilation constant to take different unit scales for different dimensions into account. Later, Reckase and Martineau (2004) proposed an oblique Procrustes rotation method to match the discrimination estimates of common items, which are obtained from separate calibrations on linking tests.

With the MIRT indeterminacy formula in Equation 1.7, the oblique Procrustes rotation method is shown below in details. First, the transformation between discrimination estimates of common items from the alternate form to the base form is obtained by

$$\mathbf{T} = (\mathbf{A}_\mathbf{a}^{'}\mathbf{A}_\mathbf{a})^{-1}\mathbf{A}_\mathbf{a}^{'}\mathbf{A}_\mathbf{b} \ , \tag{2.7}$$

where $\mathbf{T}$ is the *m* by *m* rotation matrix, $\mathbf{A}_\mathbf{a}$ is the *n* by *m* matrix of discrimination estimates for the alternative form, and $\mathbf{A}_\mathbf{b}$ is the matrix of the same size for the base form that is the target for the transformation. Then $\widehat{\mathbf{A}}_\mathbf{b}$, the **a**-parameter estimates from alternate form on the metric of base form, is

$$\widehat{\mathbf{A}}_\mathbf{b} = \mathbf{A}_\mathbf{a}\mathbf{T} \ . \tag{2.8}$$

Transformation of the *d*-parameter estimates from the alternate form to the metric of the base form is obtained by

$$\mathbf{M}' = (\mathbf{d}_\mathbf{b} - \mathbf{d}_\mathbf{a})'\widehat{\mathbf{A}}_\mathbf{b}(\widehat{\mathbf{A}}_\mathbf{b}^{'}\widehat{\mathbf{A}}_\mathbf{b})^{-1} \ , \tag{2.9}$$

where $\mathbf{d}_\mathbf{b}$ is the *n*-element vector of *d*-parameter estimates from the base form, and $\mathbf{d}_\mathbf{a}$ is the vector of the same size for the alternate form. Then the *d*-parameter estimates of the alternate form on the base form metric is

$$\widehat{\mathbf{d}}_\mathbf{b} = \mathbf{d}_\mathbf{a} + \mathbf{A}_\mathbf{a}\mathbf{T}\mathbf{M} \ . \tag{2.10}$$

Accordingly, the $\boldsymbol{\theta}$ estimates from the alternate form on the base form metric is

$$\widehat{\boldsymbol{\theta}}_\mathbf{b} = \mathbf{T}^{-\mathbf{1}}\boldsymbol{\theta}_\mathbf{a} - \mathbf{M} \ . \tag{2.11}$$

Most studies of MIRT linking methods focused on separate calibrations. One of the few examples for the concurrent calibration for the MIRT linking could be found in Reckase and Li

(2007), where they discussed the issue of using concurrent calibration to link the tests from adjacent grades.

### 2.2.3 Linking through separate and concurrent calibrations

Numerous studies discussed and compared the linking results by using separate and concurrent calibrations. The findings for which method performs better are mixed. The study of Kim and Cohen (1998) compared separate and concurrent linking methods in three ways (separate calibration with the characteristic curve linking, concurrent calibration with the marginal maximum a posteriori estimation, and concurrent calibration with the marginal maximum likelihood estimation) using the simulated unidimensional data. They found that the three methods could yield similar results when the number of common items was large; however, when the number of common items was small, they noticed that the separate calibration could provide more accurate results. Contrary to that, some other studies found that concurrent calibration could produce more stable linking results even when the number of common items was not large. For example, in the study by Hanson and Béguin (2002), concurrent calibration was found to result in lower errors than separate calibration by using the BILOG-MG software package when the groups were non-equivalent. One possible reason given by them was that the parameter estimates for the common items are based on larger samples. But they also reported the effect of different software packages when comparing concurrent and separate calibrations. When the MULTILOG software package was used, the concurrent estimation also performed well, except when the two groups had a mean difference of one standard deviation.

Furthermore, Kolen and Brennan (2004, p. 391) pointed out that concurrent estimation might be more preferable, since it is less time consuming and is supposed to generate more stable results given the IRT model holds. However, they also indicated that in practice the separate estimation might be more popular. The reason was that the two sets of item parameters estimated from different tests could be compared to check their behaviors and identify potential problems under the common item design. More importantly, according to them, the violation of unidimensionality assumption could cause problems in the concurrent calibration for vertical scaling, since this approach assumes a single proficiency estimated across all grades. Béguin, Hanson and Glas (2000) also examined the accuracy of equating with separate and concurrent calibration using the unidimensional IRT model when the data were actually generated with a two-dimensional model. It was found that under the equivalent groups design, the separate calibration performs consistently better than the concurrent calibration. In the non-equivalent groups design, both methods gave unsatisfactory results when their results were compared to those from a correctly specified two-dimensional model; however, when the proficiency correlation was high, the separate calibration still performed better than the concurrent calibration. Therefore, separate calibration seemed to show comparative robustness to the violation of unidimensionality and this may be due to the fact that the parameter estimation was carried out for only one grade level at each computer run. This finding was confirmed by many studies (e.g., Hoskens, Lewis, & Patz, 2003; Karkee, Lewis, Hoskens, Yao, & Haug, 2003; Kim & Cohen, 1998). However, according to the studies by Hanson and Béguin (2002), Patz and Yao (2007) and Yao and Mao (2004), there is some evidence that given the model is correctly specified, the concurrent calibration method might produce more stable results.

There are also several discussions which argued that concurrent calibration should be conducted with the software packages that allow the multiple group estimation, such as the BILOG-MG. However, when the MIRT model is assumed, currently, there is no efficient software package for the MIRT calibration for multiple groups. On the other hand, according to

Simon (2008), in the MIRT linking, concurrent calibration, which was conducted without additional parameters for multiple groups, generally performs better than the linking methods with separate calibration even when the mean difference between proficiencies of two groups was 0.5 standard deviation and the correlation among proficiency dimensions was high.

According to the results of above research, the concurrent calibration method should be used for the multidimensional IRT estimation in this study so that the MIRT estimates from different tests are automatically aligned to the same coordinate system.

## 2.3 Research on Common Items in Equating

With the common item non-equivalent groups design, there are numerous research studies on the effect of different common items on the equating results. Some studies (Haertel, 2004; Michaelides & Haertel, 2004) investigated the behavior of linking items on the test equating results and found that the error caused by the selection of common items has been overlooked in the error calculation process. Some other studies (Raju, Edwards, & Osberg, 1983; Wingersky & Lord, 1984) investigated the minimum adequate number of common items and suggested that as few as five or six carefully chosen items could serve as satisfactory anchors in IRT equating when the item parameters of both tests are estimated in one single analysis. However, the rule of thumb for the minimum number of common items was given by Angoff (1984, p. 107), who suggested that 20 items or 20% of the total number of items in the test are more appropriate for linking.

More importantly, in order to reflect group differences accurately, the set of common items should be proportionally representative of the total test in content and statistical characteristics (Kolen & Brennan, 2004, p. 19; Petersen, Kolen, & Hoover, 1989, p. 246). This is a commonly accepted rule for the equating in either research or operational work. Sinharay and Holland

(2006a) reexamined discussions about the correlation between common items and the total test in equating. Based on the previous studies, they asserted that the miditest, which consists of medium difficulty items, has higher reliability and also higher correlation with the total test than the commonly used minitest. Since it is long believed that higher anchor-test-to-total-test correlation could lead to better equating (Angoff, 1971, p. 577; Petersen et al., 1989, p. 246), Sinharay and Holland (2006b) doubted the necessity of selecting common items to form a mini-version of the total test and pointed out that the anchor test with a spread of item difficulties less than that of a total test seems to perform as well as or even better than a minitest. In that study, they also discussed the issue of composing an anchor test with different spread of difficulties using the data simulated from multidimensional IRT model. They found that the content representativeness of anchor items is very crucial in the equating but there seems to be no practically significant difference in the equating performances using either minitest or miditest as the anchor test. Note that all the equatings in their study were conducted using the classical methods, although the data were simulated with the IRT models.

Nevertheless, all the above research focused on the principles and suggestions on common item selection in horizontal equating. Different from that in equating, tests in vertical scaling cannot be designed as parallel forms, since items with different difficulty levels should be selected to be consistent with the curriculum and instructions for different grades; therefore, vertical scaling can only be called linking instead of equating, which requires more restricted conditions on the characteristics of equated tests. Additionally, the validity of inferences is highly influenced by the strength of the linkage, which is determined by the characteristics of common items under the common item design.

## 2.4 Research on Common Items in Vertical Scaling

Several studies have also been conducted on the common item selection in vertical scaling. Two simulation studies (Jiao & Wang, 2006; Wang, Jiao, Young, & Jin, 2006) explored the effects of the linking items when they come from different sources: using only below-grade items, using only above-grade items, or using both below-grade and above-grade items. They showed inconsistent recovered growth patterns and variability of scale scores when different off-grade items were used for vertical linking in a common person design.

Jiao and Wang (2007) tested the effect of anchor items with respect to the source of linking items, target test difficulty and the percentage of linking items relative to the total test. They found that separate calibration method using both below-grade and above-grade items for linking would lead to the best recovery. In addition, they concluded that more linking items could yield less mean bias in proficiency estimation and higher classification accuracy. The simulation study by Turhan et al. (2007) tested the effect of anchor items on vertical scaling according to the item difficulty level, the proficiency distribution and the dimensionality of the constructs. They concluded that with appropriate content coverage, any item from upper or lower grade tests could be selected as an anchor item, and slight violations of the unidimensionality assumption did not distort the vertical scale, given the good content coverage by the anchor items. However, from their design, anchor items were only selected according to the difficulty and grade level instead of different content domains; therefore, it was insufficient to draw the conclusion on the effect of content coverage by anchor items, which was not examined in their study.

## 2.5 Research Objectives and Questions

Most of the above studies were based on the results from the unidimensional IRT calibration, even for the data simulated from a MIRT model. As is well known, the unidimensional IRT model assumes that only one latent trait is necessary to account for variations in examinees' response strings (Lord, 1980); however, in practical settings, multiple abilities/traits are often required to get correct responses in standardized tests. Since the MIRT model is specially designed for the multidimensional data, it is very important to choose appropriate common items for vertical scaling under the framework of multidimensional IRT instead of the misspecified unidimensional IRT.

Based on the previous studies on the common item selection methods, although some criteria were set up to select common items via unidimensional IRT, hardly could I find any guidelines in the MIRT vertical scaling literature. Nor did any research evaluate different common item selection methods when constructs measured by the tests from different grades are not identical. Therefore, it is very important to further examine these issues and this study aims to answer the following three research questions.

First, in Part I, a design is used to evaluate different ways to select common items for vertical scaling in MIRT when both lower and upper grade tests measure the same constructs. In this part, items in the tests of both grades are manipulated to mainly differ in difficulty. Common items used for linking can be selected according to different content coverage and item difficulty level from the MIRT framework. In addition to these MIRT methods, one classical correlation method is also applied to select common items to examine their influence on the scale linking.

Second, in Part II, a design is used to evaluate different ways to select common items for vertical scaling in MIRT when the upper grade test measures more constructs than the lower grade test. This part is designed to evaluate whether it is useful to include in the anchor test items that measure the constructs only in the upper grade test, or it is sufficient to include items which measure the constructs in both tests. More specifically, this design is to test the effectiveness of using items measuring common constructs to replace those measuring the unique constructs in the anchor test, especially when the proficiency correlation between these constructs is high.

Third, in both Part I and Part II, designs are also used to evaluate the effect of different proficiency correlation levels on the linking results. One special interest is to examine whether the correlation between the construct only measured by the upper grade test and the construct measured by both tests has any impact on the linking strength in the Part II design.

Different ways of selecting common items are evaluated and compared by checking the accuracy of item and person parameter recovery. Since parameters are not available in real data, a simulation study is used to answer the aforementioned research questions. Results of this study can provide practitioners with guidance on which common item selection method should be used in vertical scaling under the MIRT framework.

# CHAPTER 3

# DESIGNS AND METHODS

This chapter first describes the design and data generation method for the two parts that deal with different content structures. Then the MIRT calibration and the evaluation criteria for parameter recovery are discussed in details.

## 3.1 Parameter Simulation

In order to make parameters more realistic, efforts were made to match the generating parameters to those estimated from the real data with respect to the distribution, structure and complexity. The parameters in this study were either from those in the study by Reckase and Li (2007) or generated from the estimated distributions from that study, and some adjustments were also made to these parameters to match the research interest of this study. The parameters used in their study were revised from the research result of Li (2006), who analyzed the data of the 2005 mathematics tests from the Michigan Educational Assessment Program (MEAP) (2005) in details. For the content structure defined in that study, the Grade 6 test is considered to measure two constructs named as ‘Problem solving’ and ‘Arithmetic’, while besides these two constructs, the Grade 7 test measures one more construct ‘Algebra’. Note that these constructs were determined from item clustering rather than the content specifications for the test.

All items were generated to be approximate simple structure items (Roussos, Stout , & Marden, 1998) so that they have high discrimination values on one dimension and low values on all other dimensions. These approximate simple structure items were used for simplicity and clarification in that the correlation between item responses is assumed to be only determined by person proficiencies, instead of the correlated composite effects caused by items (Fang, 2008).

The idea of item cluster, which was first proposed by Miller and Hirsch (1992), was also used for the generation of item discrimination parameters. The item cluster is defined as a set of items whose vectors roughly point to the same direction in the multidimensional space. All items within the same cluster are supposed to measure the same proficiency that can be put onto one continuum scale. Roussos et al. (1998) gave an example of using item clusters to define the dimensionality based on the inter-cluster proximity matrices. According to Reckase (2009, p. 221), for all items within one cluster, the angle between each pair of item vectors should be small. As is well known, the angle between any two item vectors in a multidimensional space can be computed through the following formula:

$$\alpha_{i1,i2} = \arccos(\cos \boldsymbol{\alpha_{i1}}' \cos \boldsymbol{\alpha_{i2}}) = \arccos\left( \frac{\sum_{k=1}^{m} a_{i1k} a_{i2k}}{\sqrt{\sum_{k=1}^{m} a_{i1k}^2} \sqrt{\sum_{k=1}^{m} a_{i2k}^2}} \right), \quad (3.1)$$

where $\alpha_{i1,i2}$ is the angle between vectors for items $i_1$ and $i_2$, $\boldsymbol{\alpha_{i1}}$ and $\boldsymbol{\alpha_{i2}}$ are the vectors of direction angles for item $i_1$ and $i_2$, and $a_{i1k}$ and $a_{i2k}$ are the *k*th dimensional discrimination parameters for item $i_1$ and $i_2$, respectively. The angle between item vectors can range from 0° to 90°. An angle of 0° means that the two item vectors point in exactly the same direction and the underlying proficiencies measured by these two items are perfectly correlated, while an angle of 90° indicates that there is no correlation between the two underlying proficiencies.

The multidimensional discrimination parameter was simulated from a lognormal distribution and $\log(MDISC)$ had a mean of 0 and a standard deviation of 0.2. The simulation of within-cluster angles followed the idea of approximate simple structure, where item vectors measuring a certain dimension randomly fall within 15° around that dimensional axis. Therefore, the angle on the dominating dimension, $\alpha$, was simulated to follow a uniform distribution with the range between $0°$ and $15°$. In a two-dimensional situation, the angle on the other dimension was then calculated by $90^{\circ} - \alpha$; however, in a three-dimensional situation, the angle for a second dimension was simulated from a uniform distribution with the range between $90° - \alpha$ and $90°$, and the third angle was obtained by the mathematical fact that the sum of the squared cosines of all angle degrees should be equal to one.

The *MDIFF* was simulated according to the normal distributions with a mean of -0.2 for the lower grade, a mean of 0 for the upper grade, and standard deviations of 0.75 for both grades. Since the mean of these simulated *MDIFF*s may not be close to the proposed mean due to the small sample, the *MDIFF*s of all unique items were adjusted according to the difference between the proposed and sample means, and this was done for each dimension and for each grade. Finally, $d_i$ was computed by $-MDIFF_i * MDISC_i$ as in Equation 1.5.

This study was divided into two parts that address different research questions. In Part I, both lower and upper grade tests measure the same two constructs, while in Part II, besides the same two constructs measured by both tests, one more construct is measured by the upper grade test. In each part, 3000 person and 40 unique item parameters were simulated for each grade. The same unique item parameters were employed for the simulation for each part, while the person proficiency parameters were manipulated to vary according to different correlation levels. Additionally, one item pool with 100 items was created according to the generation distribution of these unique items, and 10 common items were selected from the pool according to different criteria, such as content coverage, item difficulty and classical point-biserial correlation. All the parameters were generated using Matlab (The MathWorks., 2008).

With the item and person parameters, the probability matrix was computed using the MIRT model in Equation 1.2 by assuming the guessing parameters to be 0. Then the dichotomous response matrix was created by comparing the true probability matrix with a matrix where elements were randomly simulated from a standard uniform distribution. In this study, in order to make the results more comparable, the full response matrix was generated for each replication. This full response matrix is a matrix of 6000 examinees by 180 items. The first 100 columns in the matrix include responses on all items in the item pool and for all examinees, while the remaining 80 columns contain responses on unique items for examinees from either lower or upper grade and each of the two sets of 40 unique items was answered only by 3000 examinees in the corresponding grade. Since unique items designed for lower grade were not administered to upper grade examinees and vice versa for lower grade examinees, those responses were missing by design and coded as not presented in the response matrix.

Based on this full response matrix, the response matrix for analysis on different common item selection method was created by combining responses to the 10 common items selected from the item pool and responses to all the unique items. The layout for this response matrix is shown in Table 3.1. Therefore, for each replication, the difference among the data matrices for different item selection methods only lies in the responses to different sets of common items.

Table 3.1. Layout of Person by Item Response Matrix

|  | Item |  |  |
|---|---|---|---|
| Person | Common Items (10) | Lower grade items (40) | Upper grade items (40) |
| Lower grade examinees (3000) | lower grade item responses | lower grade item responses | not presented |
| Upper grade examinees (3000) | upper grade item responses | not presented | upper grade item responses |

## 3.2 Parameter Estimation and Vertical Scaling

In this study, concurrent calibration with common items as links was used for scaling so that the item estimates for tests in both grades were automatically put on the same coordinate system. The MIRT calibration on the response matrix was conducted via the TESTFACT software package. As mentioned in Section 1.3.4, the TESTFACT software estimates item parameters by applying the EM algorithm to maximize the constructed marginal maximum likelihood, where person proficiency parameters are integrated out via some pre-assumed proficiency distribution. These item estimates are then regarded as fixed parameters for the proficiency estimation. The convergence criterion of the EM algorithm was set to a maximum of 200 cycles and the precision of 0.005, and the options of nine quadrature points and EAP scoring method were specified in the TESTFACT syntax for the proficiency estimation.

For the concurrent calibration in TESTFACT, item and person parameter estimates were supposed to be ready for further analysis; however, one small error was found for the person parameter estimation in this software package when the concurrent calibration was used under the common item design. It seemed that person proficiencies for the second group were incorrectly estimated by using the item estimates from the first group; therefore, instead of one computer run to obtain both item and person estimates simultaneously, person proficiencies were separately estimated for each grade using additional computer runs by fixing the related item (10 common items + 40 unique items) estimates as parameters.

Due to the aforementioned three indeterminacies in the MIRT model (Li & Lissitz, 2000), TESTFACT provides one solution of item and person parameter estimates using some convenient constraints. These estimates are subject to rotational, unit and origin transformations; however, the core part $\mathbf{a}'\boldsymbol{\theta} + d$ should be invariant to the transformation to ensure the invariance property of the MIRT model, which means that the probabilities of item responses remain unchanged through the transformation (Reckase, 2009, p. 235).

## 3.3 Evaluation Criteria

In order to reduce the impact of different common items on the evaluation, these common items were not included in the calculation of evaluation indices, although they were used to link the scales across the two grades and estimate person proficiencies for each grade. In this study, three indices, including Pearson's correlation, bias and Root Mean Squared Error (RMSE), were employed to evaluate the parameter recovery. Among these indices, correlation explains the linear trend between the estimated and true parameters, bias represents the average of the differences between estimated and true parameters across replications and RMSE refers to the square root of the sum of squares of those differences. In this study, high correlation, zero bias and low RMSE indicate a good recovery of parameters. The formulas of bias and RMSE for parameter $\eta$ are shown in Equations 3.2 and 3.3:

$$Bias = \frac{\sum_r (\hat{\eta}_r - \eta)}{R} \quad , \tag{3.2}$$

$$RMSE = \sqrt{\frac{\sum_r (\hat{\eta}_r - \eta)^2}{R}} \quad , \tag{3.3}$$

where $\hat{\eta}_r$ is the estimate for $\eta$ in the *r*th replication, and *R* is the total number of replications.

The linking performance of different item selection methods was evaluated with four parameter recovery criteria, including the probability matrix recovery, the item **a**-parameter recovery, the item *d*-parameter recovery and the effect size recovery. Although the estimated and true probability matrices can be compared directly for different item selection methods, the recoveries related to item and person parameters can only be evaluated after the estimates from the TESTFACT software are put onto the same coordinate system as the parameters. In the study by Reckase and Li (2007), before the evaluation on the parameter recovery, the item discrimination estimates were rotated to a simple structure through the oblique Procrustes rotation method described in Section 2.2.2 to match the content dimensions measured by these items. The target matrix in their study was defined as the 0/1 matrix with 1s for the measured dimension and 0s elsewhere; rather, in this study, the true discrimination parameter matrix was used as the target to solve the rotational and unit indeterminacies as in Simon (2008), and the *d* estimates were then adjusted accordingly by matching with the *d*-parameters. Finally, the evaluation on the effect size recovery was conducted after the person parameter estimates were transformed based on the transformations of item estimates. The computation details for each evaluation criteria are described as follows:

1. The recovery of probability matrix. The estimated probability matrix for correct responses was computed using the item and person parameter estimates. Meanwhile, the true probability matrix could also be obtained with the parameters for the simulation. The correlation was calculated between the two vectors based on the vectorization of the true and estimated probability matrices for each replication and then averaged across all replications; however, the bias and RMSE were computed for each entry in the probability matrix, and then averaged across items and examinees.

2. The recovery of item **a**-parameters. With the oblique Procrustes rotation method, the item estimates from the TESTFACT software were first transformed to match the generating parameters and then compared with the parameters. The correlation for the recovery on each dimension was calculated between the estimated and true *a*-parameters for each replication and

then averaged across replications; however, the bias and RMSE were computed for each *a*-parameter and then averaged across items for each dimension.

3. The recovery of item *d*-parameters. The *d* estimate is affected not only by the rotation and unit indeterminacies but also by the origin indeterminacy. Reckase (2009, p. 242) mentioned that "the change in *d*-parameter is the addition of a term that is the shift in origin weighted by the *a*-parameter corresponding to the coordinate axis." Hence, it is the cumulative effect from the change of the coordinate system that results in the change of the *d*-parameters. After the *d* estimates from the TESTFACT software were transformed, the correlation was calculated between the adjusted estimates and the true *d*-parameters for each replication and then averaged across replications; however, the bias and RMSE were computed for each item and then averaged across items.

4. The recovery of effect sizes. The use of effect size, which is of great interest to policy makers, shows how sensitive each common item selection method is to detect the differences in achievements across the two grades. In this study, the effect size was computed by dividing the difference in means of proficiencies for examinees in different grades by the pooled standard deviation of proficiencies for these examinees, based on the true parameters as well as those transformed estimates. The formula is shown as following:

$$ES_k = \frac{\bar{\theta}_{k,upper} - \bar{\theta}_{k,lower}}{s_k} \ , \qquad (3.4)$$

and

$$s_k = \sqrt{\frac{s_{k,upper}^2 + s_{k,lower}^2}{2}} \ ,$$

where $\bar{\theta}_k$ is the mean of proficiencies on the *k*th dimension and $s_k$ is the standard deviation of proficiencies on the same dimension. Since the effect size is an aggregated statistics, the three indices for parameter recovery are not appropriate for the evaluation; instead, the magnitudes of the estimated and true effect sizes were used for the comparison. Therefore, for each dimension, the estimated effect size was calculated for each replication and then averaged across replications for the comparison with the true effect size.

# CHAPTER 4

# PART I: SAME CONSTRUCTS

## 4.1 Parameters and Designs

In this part, both lower and upper grade tests are assumed to measure the same two constructs. There are 40 unique items in the test of each grade; besides, in both tests, there are 10 common items that are selected according to nine different methods. For different grades, person proficiencies are simulated from the multivariate normal distributions with different mean vectors but the same variance-covariance matrix.

### 4.1.1 Unique items

For each grade, 20 unique items measure the first construct and the other 20 unique items measure the second one. The cluster number, item parameters, multidimensional difficulty, multidimensional discrimination and direction angles of unique items for tests in both grades are listed in Tables 4.1 and 4.2. Items in Cluster 1 with large loadings on $a_1$ mainly measure the proficiency on Dimension 1, while those in Cluster 2 measure the proficiency on Dimension 2. The means of *MDISC*s for different grade levels are quite similar; however, the mean and standard deviation of *MDIFF*s are -0.2 and 0.83 for the lower grade, while they are 0 and 0.63 for the upper grade. Therefore, the mean of *MDIFF*s for unique items is smaller for the lower grade than for the upper grade.

### 4.1.2 Common items

After unique items for each grade level were simulated, an item pool with 100 items was generated with half of the items simulated using the generation distribution of unique items for each grade level. Items in this pool were then divided into several categories according to the

Table 4.1. Unique Item Parameters and Statistics for Lower Grade in Part I

| Cluster | $a_1$ | $a_2$ | $d$ | *MDIFF* | *MDISC* | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.02 | 0.07 | 1.00 | -0.97 | 1.02 | 4 | 86 |
| 1 | 1.07 | 0.25 | 0.65 | -0.59 | 1.10 | 13 | 77 |
| 1 | 0.83 | 0.03 | 0.52 | -0.63 | 0.83 | 2 | 88 |
| 1 | 0.97 | 0.08 | -0.27 | 0.28 | 0.98 | 5 | 85 |
| 1 | 0.89 | 0.20 | -1.22 | 1.34 | 0.91 | 13 | 77 |
| 1 | 0.90 | 0.06 | -0.47 | 0.52 | 0.90 | 4 | 86 |
| 1 | 1.37 | 0.26 | 0.39 | -0.28 | 1.39 | 11 | 79 |
| 1 | 1.18 | 0.09 | -0.38 | 0.32 | 1.18 | 5 | 85 |
| 1 | 1.12 | 0.14 | 0.35 | -0.30 | 1.13 | 7 | 83 |
| 1 | 1.08 | 0.23 | 2.00 | -1.81 | 1.10 | 12 | 78 |
| 1 | 1.04 | 0.23 | 0.23 | -0.22 | 1.06 | 13 | 77 |
| 1 | 1.27 | 0.23 | 0.60 | -0.46 | 1.29 | 10 | 80 |
| 1 | 1.13 | 0.14 | 0.38 | -0.33 | 1.14 | 7 | 83 |
| 1 | 1.04 | 0.21 | 0.38 | -0.35 | 1.06 | 12 | 78 |
| 1 | 0.96 | 0.08 | -1.94 | 2.03 | 0.96 | 5 | 85 |
| 1 | 1.27 | 0.31 | 1.68 | -1.28 | 1.31 | 14 | 76 |
| 1 | 1.13 | 0.06 | 1.34 | -1.19 | 1.13 | 3 | 87 |
| 1 | 1.08 | 0.09 | 1.04 | -0.96 | 1.09 | 5 | 85 |
| 1 | 0.75 | 0.13 | -0.71 | 0.94 | 0.76 | 10 | 80 |
| 1 | 1.00 | 0.08 | 0.04 | -0.04 | 1.00 | 4 | 86 |
| 2 | 0.12 | 1.26 | -0.76 | 0.60 | 1.27 | 85 | 5 |
| 2 | 0.17 | 0.78 | -0.39 | 0.49 | 0.80 | 78 | 12 |
| 2 | 0.09 | 0.93 | 1.39 | -1.48 | 0.94 | 84 | 6 |
| 2 | 0.24 | 0.90 | 0.25 | -0.27 | 0.93 | 75 | 15 |
| 2 | 0.13 | 0.97 | 0.60 | -0.62 | 0.98 | 82 | 8 |
| 2 | 0.14 | 0.88 | 0.41 | -0.46 | 0.89 | 81 | 9 |
| 2 | 0.00 | 0.80 | 0.38 | -0.47 | 0.80 | 90 | 0 |
| 2 | 0.19 | 0.98 | 0.36 | -0.36 | 1.00 | 79 | 11 |
| 2 | 0.16 | 1.09 | 0.82 | -0.75 | 1.10 | 82 | 8 |
| 2 | 0.22 | 1.01 | 0.85 | -0.82 | 1.03 | 78 | 12 |
| 2 | 0.19 | 0.81 | 0.19 | -0.23 | 0.83 | 77 | 13 |
| 2 | 0.11 | 0.83 | -1.10 | 1.31 | 0.84 | 82 | 8 |
| 2 | 0.13 | 1.00 | 0.23 | -0.22 | 1.01 | 82 | 8 |
| 2 | 0.12 | 0.86 | 0.47 | -0.54 | 0.87 | 82 | 8 |
| 2 | 0.26 | 1.21 | 0.17 | -0.14 | 1.23 | 78 | 12 |
| 2 | 0.02 | 0.88 | 1.49 | -1.69 | 0.88 | 89 | 1 |
| 2 | 0.12 | 1.07 | -1.29 | 1.21 | 1.07 | 84 | 6 |
| 2 | 0.14 | 0.90 | -0.08 | 0.09 | 0.91 | 81 | 9 |
| 2 | 0.02 | 0.67 | 0.15 | -0.22 | 0.67 | 88 | 2 |
| 2 | 0.23 | 0.98 | -0.60 | 0.60 | 1.01 | 77 | 13 |
| Mean | 0.60 | 0.55 | 0.23 | -0.20 | 1.01 | | |
| Std | 0.48 | 0.42 | 0.84 | 0.83 | 0.16 | | |

Table 4.2. Unique Item Parameters and Statistics for Upper Grade in Part I

| Cluster | $a_1$ | $a_2$ | $d$ | *MDIFF* | *MDISC* | $\alpha_1$ | $\alpha_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1.46 | 0.31 | -0.91 | 0.61 | 1.49 | 12 | 78 |
| 1 | 1.01 | 0.22 | 0.11 | -0.11 | 1.04 | 12 | 78 |
| 1 | 0.68 | 0.13 | -0.03 | 0.04 | 0.69 | 11 | 79 |
| 1 | 0.80 | 0.05 | -0.58 | 0.72 | 0.80 | 4 | 86 |
| 1 | 0.55 | 0.12 | -0.10 | 0.17 | 0.56 | 12 | 78 |
| 1 | 1.07 | 0.28 | 0.54 | -0.49 | 1.11 | 15 | 75 |
| 1 | 1.08 | 0.15 | -0.04 | 0.04 | 1.09 | 8 | 82 |
| 1 | 1.25 | 0.02 | -2.22 | 1.78 | 1.25 | 1 | 89 |
| 1 | 1.08 | 0.20 | 0.72 | -0.65 | 1.10 | 10 | 80 |
| 1 | 0.60 | 0.11 | -0.08 | 0.13 | 0.61 | 10 | 80 |
| 1 | 1.35 | 0.19 | -0.90 | 0.66 | 1.36 | 8 | 82 |
| 1 | 0.94 | 0.24 | -0.01 | 0.01 | 0.97 | 14 | 76 |
| 1 | 1.19 | 0.16 | -0.77 | 0.64 | 1.21 | 8 | 82 |
| 1 | 1.12 | 0.16 | 0.60 | -0.53 | 1.13 | 8 | 82 |
| 1 | 0.80 | 0.02 | 0.47 | -0.58 | 0.80 | 2 | 88 |
| 1 | 1.18 | 0.29 | 0.36 | -0.29 | 1.22 | 14 | 76 |
| 1 | 1.03 | 0.25 | 0.00 | 0.00 | 1.06 | 14 | 76 |
| 1 | 1.05 | 0.15 | 0.58 | -0.55 | 1.06 | 8 | 82 |
| 1 | 0.91 | 0.19 | 0.58 | -0.63 | 0.93 | 12 | 78 |
| 1 | 1.19 | 0.28 | 1.18 | -0.96 | 1.23 | 13 | 77 |
| 2 | 0.19 | 0.80 | -0.33 | 0.41 | 0.82 | 76 | 14 |
| 2 | 0.07 | 0.96 | 1.52 | -1.59 | 0.96 | 86 | 4 |
| 2 | 0.16 | 0.98 | 1.50 | -1.52 | 0.99 | 81 | 9 |
| 2 | 0.25 | 1.03 | -0.62 | 0.59 | 1.06 | 76 | 14 |
| 2 | 0.21 | 0.86 | 0.38 | -0.43 | 0.89 | 76 | 14 |
| 2 | 0.07 | 1.03 | -0.63 | 0.61 | 1.03 | 86 | 4 |
| 2 | 0.24 | 0.90 | 0.03 | -0.03 | 0.93 | 75 | 15 |
| 2 | 0.11 | 0.92 | -0.11 | 0.12 | 0.92 | 83 | 7 |
| 2 | 0.15 | 0.73 | -0.39 | 0.53 | 0.74 | 79 | 11 |
| 2 | 0.22 | 1.22 | 0.74 | -0.60 | 1.23 | 80 | 10 |
| 2 | 0.09 | 0.79 | 0.01 | -0.01 | 0.79 | 83 | 7 |
| 2 | 0.10 | 0.93 | -0.26 | 0.27 | 0.93 | 84 | 6 |
| 2 | 0.27 | 1.22 | -0.20 | 0.16 | 1.25 | 77 | 13 |
| 2 | 0.12 | 0.90 | -0.72 | 0.79 | 0.91 | 83 | 7 |
| 2 | 0.16 | 0.89 | -0.12 | 0.13 | 0.90 | 80 | 10 |
| 2 | 0.10 | 1.04 | 0.23 | -0.22 | 1.04 | 85 | 5 |
| 2 | 0.21 | 1.18 | -0.05 | 0.05 | 1.20 | 80 | 10 |
| 2 | 0.07 | 0.84 | -0.46 | 0.55 | 0.84 | 85 | 5 |
| 2 | 0.20 | 1.31 | -0.35 | 0.26 | 1.33 | 81 | 9 |
| 2 | 0.16 | 0.92 | 0.06 | -0.06 | 0.93 | 80 | 10 |
| Mean | 0.59 | 0.57 | -0.01 | 0.00 | 1.01 | | |
| Std | 0.47 | 0.42 | 0.69 | 0.63 | 0.21 | | |

two content dimensions and three difficulty levels. The numbers of items for different combinations of these two factors are shown in Table 4.3.

Table 4.3. Number of Items for Different Content and Difficulty Categories in Item Pool of Part I

| | Low | Medium | High | All |
|---|---|---|---|---|
| Dimension 1 | 15 | 20 | 15 | 50 |
| Dimension 2 | 15 | 20 | 15 | 50 |

Ten common items for tests in both grades were selected from the item pool according to the MIRT methods and the classical correlation method. The details for each method are shown below.

(1) The MIRT methods consist of eight methods according to different combinations of content and difficulty coverage.

- Content coverage. Common items are selected (a) only from items in Cluster 1, or (b) evenly from items in Cluster 1 and Cluster 2 to achieve full content coverage.
- Difficulty coverage. Items in the pool are grouped into the low, medium and high difficulty levels. Common items are selected from (a) only the low level, (b) only the medium level, (c) only the high level, or (d) all three levels. Note that the item difficulty is confounded with the grade level, because most likely, difficult items come from the upper grade and easy items from the lower grade.

(2) The classical method selects 10 items with high item-total-test correlation in both lower and upper grade tests.

In Methods 1-4 where partial content coverage is achieved, 10 common items were selected from items in Cluster 1 according to different difficulty coverage. For Methods 1-3, a simple random sample of items was selected from the low, medium and high difficulty levels, respectively. For Method 4 with full difficulty coverage, three low, four medium and three high difficulty items were randomly chosen from the corresponding categories, according to their proportions in the item pool.

In Methods 5-8, five items were chosen from each of the two item clusters to achieve the full content coverage. Common items in Methods 5-7 were selected from each of the three difficulty categories, respectively. Note that the common item set in Method 6 can be regarded as a miditest with full content coverage. For Method 8, the set of common items covers all three difficulty levels and two content domains, and can be considered as a mini-version of the whole test.

Method 9 is a post-hoc method with common items selected based on the analysis on some generated response matrices. Correlations between the scores of items in the item pool and the total score for unique items in each grade test were calculated and ranked from high to low. These correlations were computed for two samples of response matrix at each proficiency correlation level. Ten items with high item-total-test correlation in tests of both lower and upper grades were selected as common items for this method.

The statistics for different common item sets are listed in Table 4.4. The means of *MDIFF* values range from -1.05 to 0.92. The average *MDIFF* values for common items in Methods 2, 4, 6 and 8 are all close to 0. This is reasonable since items in these methods are either from medium difficulty level with a small spread of difficulty values or from all three difficulty levels with a large spread. From the perspective of MIRT, common items in the classical correlation method seemed to be selected from the medium difficulty level but with an unbalanced coverage for the two content domains. This also indicated that items from the medium difficulty level are more highly correlated with the total test than other items.

Table 4.4. Statistics of Common Items for Different Selection Methods in Part I

| Selection Method | MDIFF Mean | MDIFF Std | # of items in Dim 1 | # of items in Dim 2 | MDIFF Mean in Dim 1 | MDIFF Mean in Dim 2 |
|---|---|---|---|---|---|---|
| (1) 1D, Low | -0.95 | 0.30 | 10 | 0 | -0.95 | NA |
| (2) 1D, Medium | 0.00 | 0.30 | 10 | 0 | 0.00 | NA |
| (3) 1D, High | 0.61 | 0.23 | 10 | 0 | 0.61 | NA |
| (4) 1D, All | 0.10 | 0.89 | 10 | 0 | 0.10 | NA |
| (5) 2D, Low | -1.05 | 0.44 | 5 | 5 | -1.05 | -1.05 |
| (6) 2D, Medium | -0.09 | 0.18 | 5 | 5 | -0.09 | -0.09 |
| (7) 2D, High | 0.92 | 0.47 | 5 | 5 | 0.90 | 0.95 |
| (8) 2D, All | -0.16 | 0.87 | 5 | 5 | -0.37 | 0.05 |
| (9) Classical | -0.26 | 0.30 | 2 | 8 | -0.42 | -0.22 |

### 4.1.3 Person parameters

For different grades, person proficiency parameters were simulated according to the multivariate normal distributions with different mean vectors but the same variance-covariance matrix. According to the analyses on MEAP mathematics test by Li (2006), the increases on mathematical skills from the lower grade to the adjacent upper grade were about 0.2 standard deviation units. Therefore, the mean vector of normal distribution for the person proficiency generation was set to be (-0.2, -0.2) for the lower grade and (0, 0) for the upper grade. The variances of person proficiencies on both dimensions were set to one; however, the correlation between proficiencies was manipulated to vary from 0, 0.4, 0.6 to 0.8 for both grades. Responses were simulated for 3000 examinees and 40 unique items in each grade, while those for the common items were simulated for all 6000 examinees in both grades.

## 4.2 Estimation

In order to reduce sampling errors, 30 response matrices were simulated based on the same probability matrix for each proficiency correlation level and the MIRT calibration was conducted on the data matrix for the selected common items and all unique items. This resulted in a total of 1080 computer runs (4 proficiency correlation levels x 30 replications x 9 item selection methods). It took about 20 minutes for each calibration with the TESTFACT software package. Dimensionality of the simulated data was checked with the MIRT calibration using one more dimension. The results showed that the item discrimination estimates on the extra dimension were very small, while the estimates were large on at least one of the other dimensions. This justified the use of the two-dimensional solution.

Due to the rotational indeterminacy, the TESTFACT software may not orient the axes of the coordinate system for the parameter estimates in a proper direction as those for the generating parameters. One common problem is that the *a*-parameter estimates on different dimensions may be switched or the estimates from some dimension may be negated, with respect to the generating parameters. However, some additional TESTFACT runs confirmed that the proficiency estimates do not change, no matter whether the *a*-parameter estimates on any dimension are negated or not. This phenomenon was also observed in the study by Fang (2008) and explained in the TESTFACT help file, which says that "it may therefore happen that negative scores are associated with above average percent responses and vice versa for below average responses. TESTFACT software attempts to reverse the signs in such a way that scores above zero are usually assigned with above average achievement."

These item and person parameter estimates were rotated to match the generating parameters before the evaluation. Thus, the order of these estimates does not lead to any problem; however, it is problematic if proficiency estimates are not correctly paired with item estimates. Therefore, the signs of item and person parameter estimates may need to be corrected to make these estimates a valid pair as one solution for the MIRT calibration.

In order to correct the sign of item discrimination estimates, the mean of these estimates was computed for each dimension. Based on the assumption that item discrimination parameters should be positive in the MIRT model, if the mean of the estimates from the TESTFACT software was negative on any dimension, the negated estimates were regarded as the correct item discrimination estimates on that dimension; otherwise, these estimates were kept the same.

Although the proficiency estimates seemed to be automatically corrected in the TESTFACT software, a double check was still conducted to examine the signs of these proficiency estimates separately for each grade. First, the percentage of correct responses was obtained for each examinee. Then, examinees with the highest and lowest percentages were picked and their proficiency estimates were examined. It was expected that all proficiency values should be positive for the examinee with the highest percentage value and negative for the examinee with the lowest percentage. If both criteria failed for the proficiency estimates on any dimension, it was very likely that the estimates provided by the TESTFACT software were incorrect and all proficiency estimates on that dimension should be negated to be paired with the item discrimination estimates on that dimension. If one of the two criteria failed, that replication would be picked for further checking. From the checking results, no proficiency estimates were found to have the sign problem; therefore, the TESTFACT software package seems to align the proficiency estimates in a correct direction as the percentage of correct responses predicts when both proficiencies contribute significantly to the percentage.

## 4.3 Results

### 4.3.1 Recovery of probability matrix

The probability matrix of correct responses is obtained by applying item and person parameters in the MIRT model as shown in Equation 1.2. Since the value of $\mathbf{a}'\boldsymbol{\theta}+d$ is not affected by the MIRT indeterminacies, so is the case with the probabilities of correct responses. The recovery of probability matrix, which is indicated by the similarity between the estimated and true probability matrices, is evaluated via the correlation, bias and RMSE indices.

For each replication, the correlation was computed using all corresponding elements in the estimated and true probability matrices. The correlation values averaged across replications for different conditions are listed in Table 4.5. All correlation values are above 0.96, which indicates that the ordering of estimated probabilities is very similar to that of true ones. As the proficiency correlation increases, the correlation between the estimated and true probabilities also increases. Among all nine selection methods, Method 9 gives the highest correlation values for all four proficiency correlation levels and the correlation values for Method 6 are the second highest.

Figure 4.1 gives the plot between the proficiency correlation level and the correlation for the probability matrix recovery for each item selection method. It is easy to observe that the points representing Methods 9 (classical correlation) and 6 (full content coverage with medium difficulty items) are above all the other points. On the other hand, the points representing Methods 1 (partial content coverage with low difficulty items) and 3 (partial content coverage with high difficulty items) are at the bottom.

Table 4.5. Correlation for the Recovery of Probability Matrix in Part I

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 1D, Low | 0.9651 | 0.9681 | 0.9700 | 0.9732 |
| (2) 1D, Medium | 0.9664 | 0.9691 | 0.9710 | 0.9741 |
| (3) 1D, High | 0.9650 | 0.9680 | 0.9700 | 0.9733 |
| (4) 1D, All | 0.9660 | 0.9689 | 0.9707 | 0.9738 |
| (5) 2D, Low | 0.9665 | 0.9694 | 0.9712 | 0.9741 |
| (6) 2D, Medium | 0.9675 | 0.9701 | 0.9719 | 0.9749 |
| (7) 2D, High | 0.9656 | 0.9685 | 0.9705 | 0.9735 |
| (8) 2D, All | 0.9664 | 0.9693 | 0.9711 | 0.9741 |
| (9) Classical | 0.9687 | 0.9712 | 0.9729 | 0.9757 |

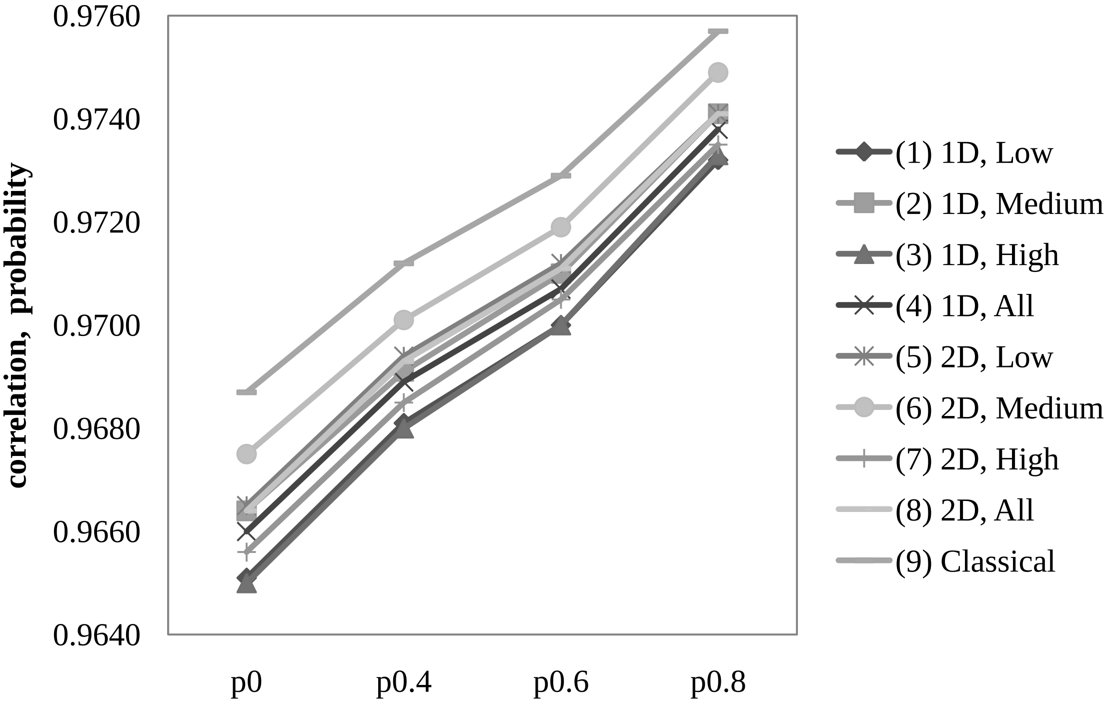

Figure 4.1. Correlation for the Recovery of Probability Matrix in Part I

Table 4.6 provides the result of bias for the recovery of probability matrix and Figure 4.2 shows the plot between the proficiency correlation level and the bias for the probability matrix recovery for each item selection method. Note that the bias values in the table were averaged across all items and examinees. One observation is that the selection methods with low difficulty items always yield positive values for bias and those with high difficulty items give negative values, while other MIRT methods give comparatively small absolute values. It is interesting to observe that Methods 2 and 4 also give good results, which seems to indicate that the bias is not influenced by the content coverage. All the absolute values of bias are less than 0.001, which tends to suggest that no bias exists in the probability estimation; however, the values in the table are not sufficient for this judgment, since parameters at different value levels may have different degrees of bias in the estimates and the bias values with different signs can be cancelled out when averaged across items and examinees.

Figure 4.3 shows the plot between the probability parameters and their estimation bias for Method 1 under the condition of zero proficiency correlation. As can be observed from the figure, the probabilities of large values tend to be underestimated and those of small values tend to be overestimated. The underestimation and overestimation of parameters seem to be less severe for the probabilities of medium values. A further analysis showed that the underestimation is mostly on the probabilities for difficult items answered by examinees with extremely high proficiency on the dimension, which is dominantly measured by each of these items. This is because the estimated proficiencies yielded from the EAP scoring method tend to be smaller than the true proficiencies with large values and this makes the estimated probabilities much smaller than the true probabilities for these difficult items. The overestimation can be explained in a similar way. The plots for all other methods and conditions are similar to this one.

Table 4.6. Bias for the Recovery of Probability Matrix in Part I

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---:|---:|---:|---:|
| (1) 1D, Low | 0.0005 | 0.0001 | 0.0006 | 0.0002 |
| (2) 1D, Medium | 0.0001 | -0.0004 | 0.0001 | -0.0001 |
| (3) 1D, High | -0.0001 | -0.0006 | -0.0001 | -0.0002 |
| (4) 1D, All | 0.0001 | -0.0004 | 0.0002 | 0.0000 |
| (5) 2D, Low | 0.0008 | 0.0004 | 0.0008 | 0.0003 |
| (6) 2D, Medium | 0.0002 | -0.0002 | 0.0003 | 0.0000 |
| (7) 2D, High | -0.0003 | -0.0007 | -0.0002 | -0.0002 |
| (8) 2D, All | 0.0002 | -0.0002 | 0.0002 | 0.0000 |
| (9) Classical | 0.0004 | 0.0000 | 0.0004 | 0.0001 |

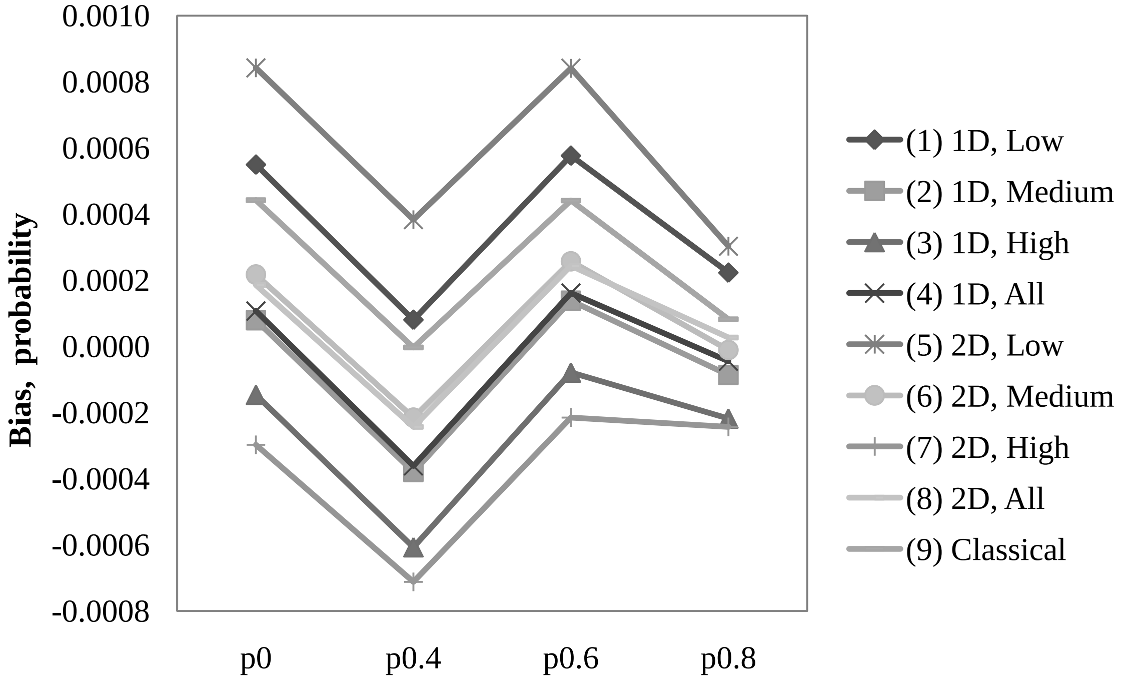

Figure 4.2. Bias for the Recovery of Probability Matrix in Part I

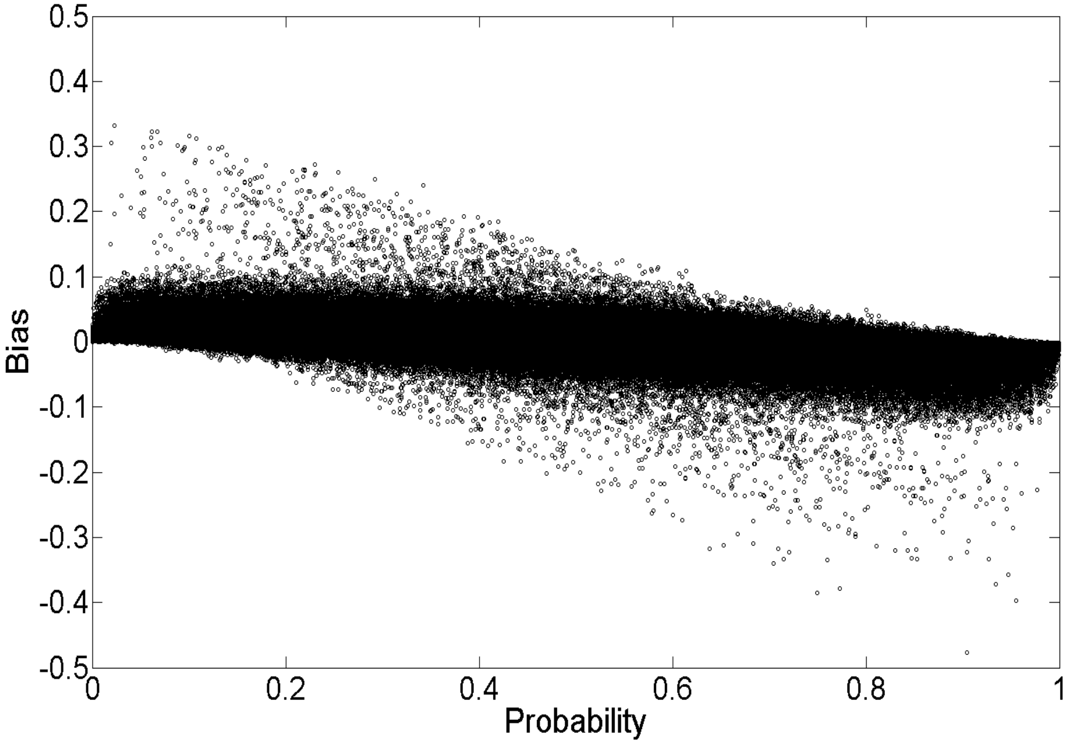

Figure 4.3. Bias for the Recovery of Probability Matrix for Method 1 at Zero Proficiency Correlation Level in Part I

The RMSE values listed in Table 4.7 also give information on the recovery of the probability matrix for correct responses. There is a clear pattern that as the correlation between proficiencies increases, the value of RMSE consistently decreases for all selection methods. For the MIRT methods selecting items from the same difficulty level, the full content coverage is more important than the partial content coverage. Also, for the same content coverage, the method of including medium difficulty items is the best among the four methods based on different difficulty levels. In particular, Methods 1 and 3 give comparatively larger RMSE values while Methods 6 and 9 provide smaller RMSE values for all four proficiency correlation levels. This can also be observed from Figure 4.4, which gives the plot between the proficiency correlation level and the RMSE value for the probability matrix recovery for each common item selection method.

In conclusion, the higher the correlation between proficiencies is, the better the estimated probabilities could match the true values. The classical correlation method gives the highest correlation and lowest RMSE for the recovery of probability matrix and the method of full content coverage with medium difficulty items is the second best.

Table 4.7. RMSE for the Recovery of Probability Matrix in Part I

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 1D, Low | 0.0752 | 0.0727 | 0.0707 | 0.0668 |
| (2) 1D, Medium | 0.0737 | 0.0714 | 0.0694 | 0.0656 |
| (3) 1D, High | 0.0753 | 0.0727 | 0.0707 | 0.0667 |
| (4) 1D, All | 0.0742 | 0.0718 | 0.0699 | 0.0661 |
| (5) 2D, Low | 0.0738 | 0.0713 | 0.0693 | 0.0657 |
| (6) 2D, Medium | 0.0726 | 0.0704 | 0.0685 | 0.0648 |
| (7) 2D, High | 0.0748 | 0.0723 | 0.0702 | 0.0665 |
| (8) 2D, All | 0.0739 | 0.0715 | 0.0695 | 0.0657 |
| (9) Classical | 0.0711 | 0.0689 | 0.0671 | 0.0635 |

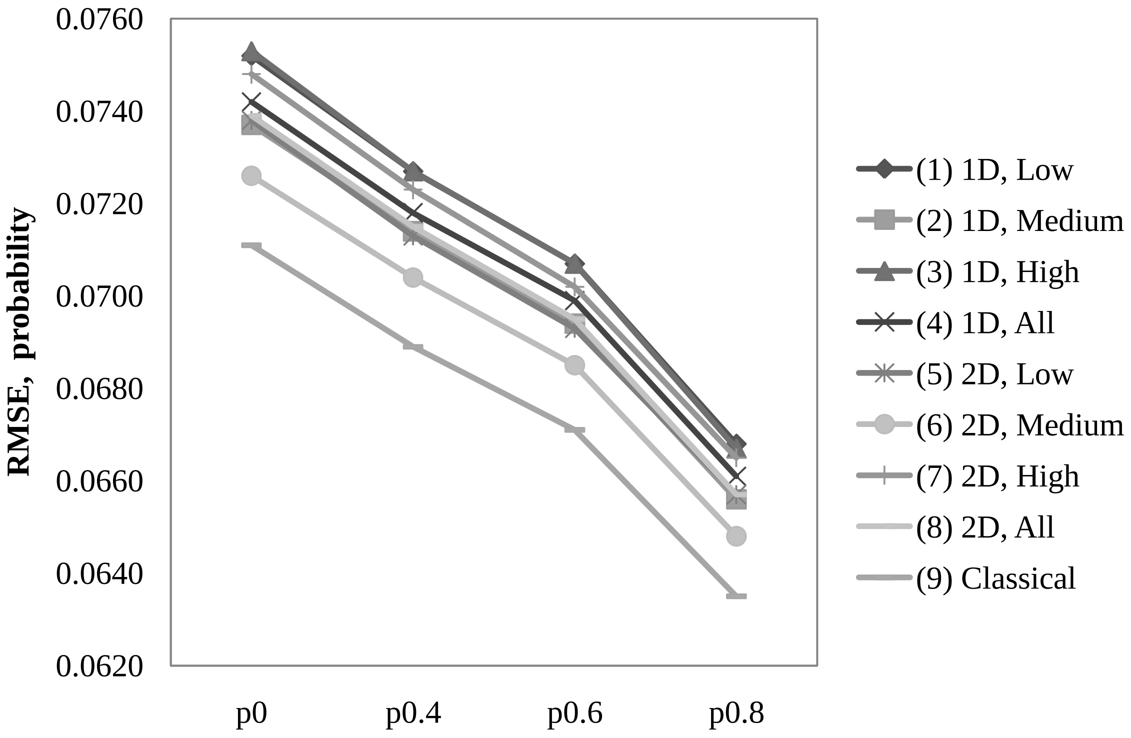

Figure 4.4. RMSE for the Recovery of Probability Matrix in Part I

### 4.3.2 Recovery of a-parameters

As is well known, the default in the TESTFACT software package for the MIRT calibration uses the zero mean vector and identity variance-covariance matrix for the distribution of proficiency coordinates, which may not match the real situation for proficiencies, and employs no rotation or the Varimax rotation for the orientation of coordinate axes to solve the MIRT indeterminacies. Therefore, the **a**-parameter estimates obtained from the TESTFACT software cannot be directly compared with generating parameters since they are not in the same coordinate system. These discrimination estimates were rotated to match the parameters using Equations 2.7 and 2.8 for the oblique Procrustes rotation before they were compared with generating parameters.

Table 4.8 shows the average correlations between the rotated estimates and true parameters. The result shows that as the correlation between proficiencies increases, the correlation between the estimated and true **a**-parameters decreases slightly but consistently. Compared with all other selection methods, Methods 6, 8 and 9 could give a little higher correlation values for the recovery on both dimensions and for all four proficiency correlation levels.

Figures 4.5 and 4.6 also plot the correlation for the recovery of $a$-parameters on each dimension. It is clear that when the correlation between proficiencies is small, the differences among methods are also small. However, with the increase of the correlation between proficiencies, the difference between the $a$-parameters and their rotated estimates tends to become larger for the recovery on any dimension.

Table 4.9 shows the bias for the recovery of **a**-parameters. The value of bias is negative for each dimension and for each item selection method, which indicates that the estimates for **a**-parameters are most likely negatively biased. Also, it seems that the magnitude of bias is larger for the lowest and highest proficiency correlation levels than for the two middle correlation levels. Figure 4.7 shows the plot between the bias values and $a_1$-parameters for Method 1 under the zero proficiency correlation condition. Points in each of the two clusters represent items that

Table 4.8. Correlation for the Recovery of **a**-parameters in Part I

| | Dimension 1 | | | | Dimension 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | p0 | p0.4 | p0.6 | p0.8 | p0 | p0.4 | p0.6 | p0.8 |
| (1) 1D, Low | 0.9944 | 0.9942 | 0.9912 | 0.9841 | 0.9949 | 0.9939 | 0.9917 | 0.9847 |
| (2) 1D, Medium | 0.9947 | 0.9944 | 0.9925 | 0.9867 | 0.9950 | 0.9940 | 0.9919 | 0.9852 |
| (3) 1D, High | 0.9946 | 0.9940 | 0.9928 | 0.9868 | 0.9948 | 0.9940 | 0.9917 | 0.9844 |
| (4) 1D, All | 0.9947 | 0.9942 | 0.9928 | 0.9867 | 0.9949 | 0.9940 | 0.9919 | 0.9850 |
| (5) 2D, Low | 0.9944 | 0.9943 | 0.9920 | 0.9861 | 0.9948 | 0.9939 | 0.9920 | 0.9857 |
| (6) 2D, Medium | 0.9947 | 0.9944 | 0.9927 | 0.9871 | 0.9950 | 0.9940 | 0.9920 | 0.9858 |
| (7) 2D, High | 0.9946 | 0.9940 | 0.9928 | 0.9867 | 0.9948 | 0.9936 | 0.9918 | 0.9846 |
| (8) 2D, All | 0.9947 | 0.9944 | 0.9926 | 0.9867 | 0.9951 | 0.9940 | 0.9920 | 0.9856 |
| (9) Classical | 0.9946 | 0.9942 | 0.9927 | 0.9870 | 0.9950 | 0.9940 | 0.9920 | 0.9861 |

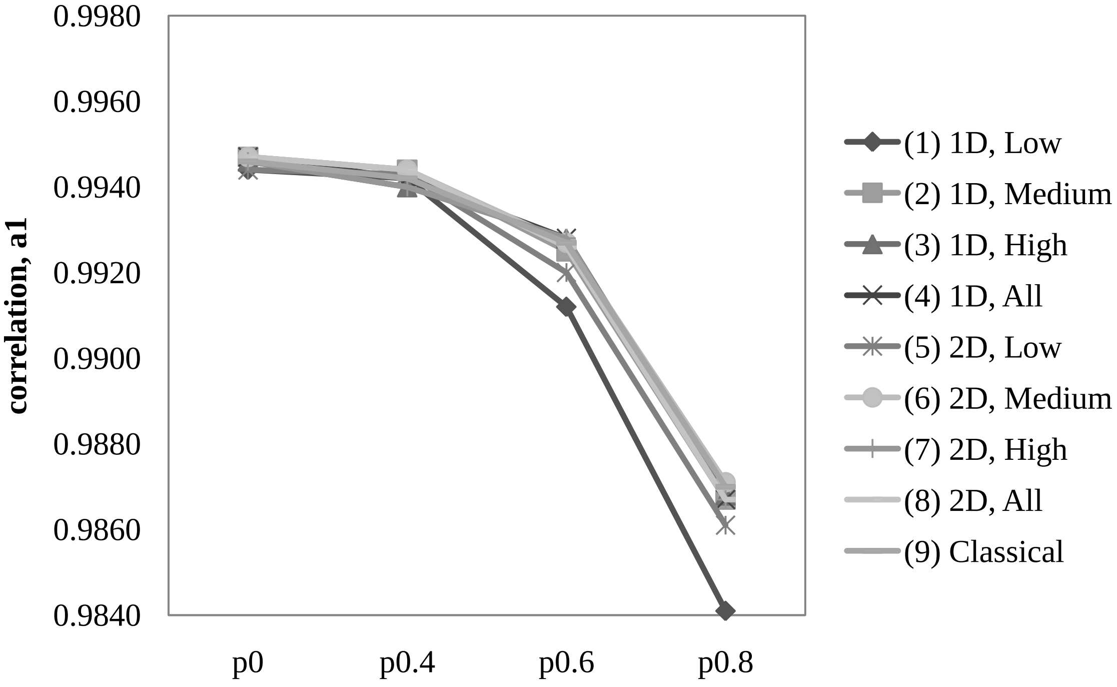

Figure 4.5. Correlation for the Recovery of $a_1$-parameters in Part I

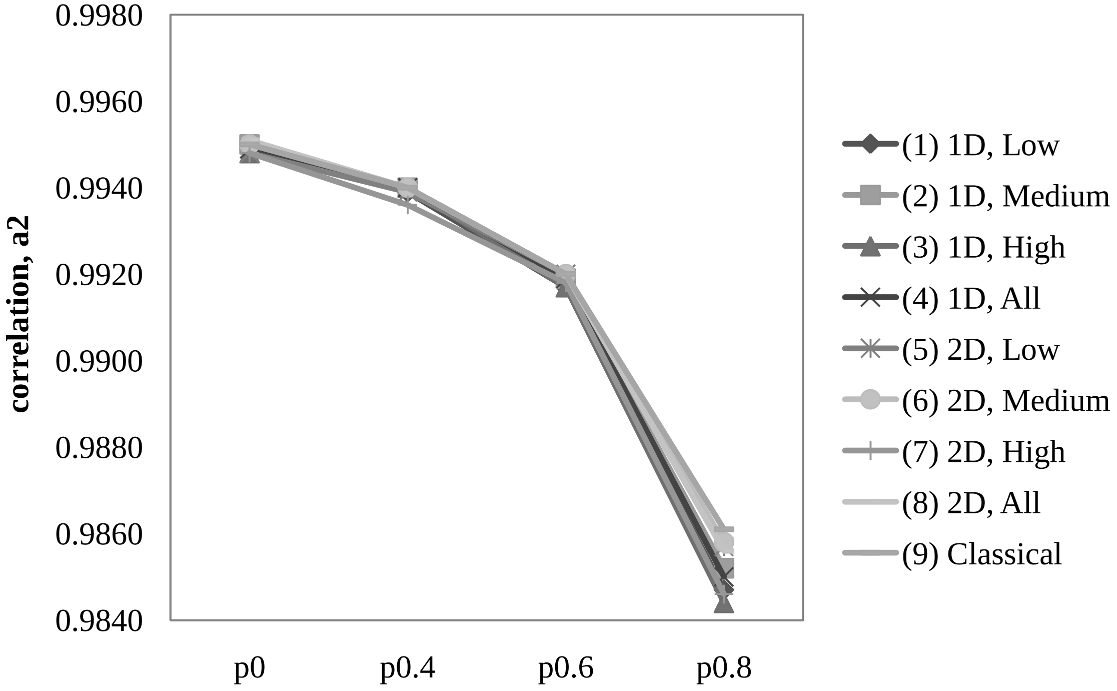

Figure 4.6. Correlation for the Recovery of $a_2$-parameters in Part I

Table 4.9. Bias for the Recovery of **a**-parameters in Part I

| | Dimension 1 | | | | Dimension 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | p0 | p0.4 | p0.6 | p0.8 | p0 | p0.4 | p0.6 | p0.8 |
| (1) 1D, Low | -0.0016 | -0.0004 | -0.0012 | -0.0014 | -0.0011 | -0.0007 | -0.0003 | -0.0012 |
| (2) 1D, Medium | -0.0013 | -0.0003 | -0.0007 | -0.0006 | -0.0010 | -0.0007 | -0.0005 | -0.0015 |
| (3) 1D, High | -0.0014 | -0.0005 | -0.0006 | -0.0005 | -0.0011 | -0.0008 | -0.0006 | -0.0015 |
| (4) 1D, All | -0.0013 | -0.0003 | -0.0007 | -0.0006 | -0.0011 | -0.0007 | -0.0005 | -0.0015 |
| (5) 2D, Low | -0.0016 | -0.0006 | -0.0010 | -0.0010 | -0.0013 | -0.0008 | -0.0006 | -0.0017 |
| (6) 2D, Medium | -0.0013 | -0.0004 | -0.0004 | -0.0006 | -0.0011 | -0.0006 | -0.0005 | -0.0013 |
| (7) 2D, High | -0.0012 | -0.0007 | -0.0005 | -0.0006 | -0.0011 | -0.0008 | -0.0005 | -0.0015 |
| (8) 2D, All | -0.0013 | -0.0005 | -0.0006 | -0.0007 | -0.0011 | -0.0007 | -0.0004 | -0.0013 |
| (9) Classical | -0.0014 | -0.0006 | -0.0005 | -0.0007 | -0.0012 | -0.0005 | -0.0006 | -0.0015 |

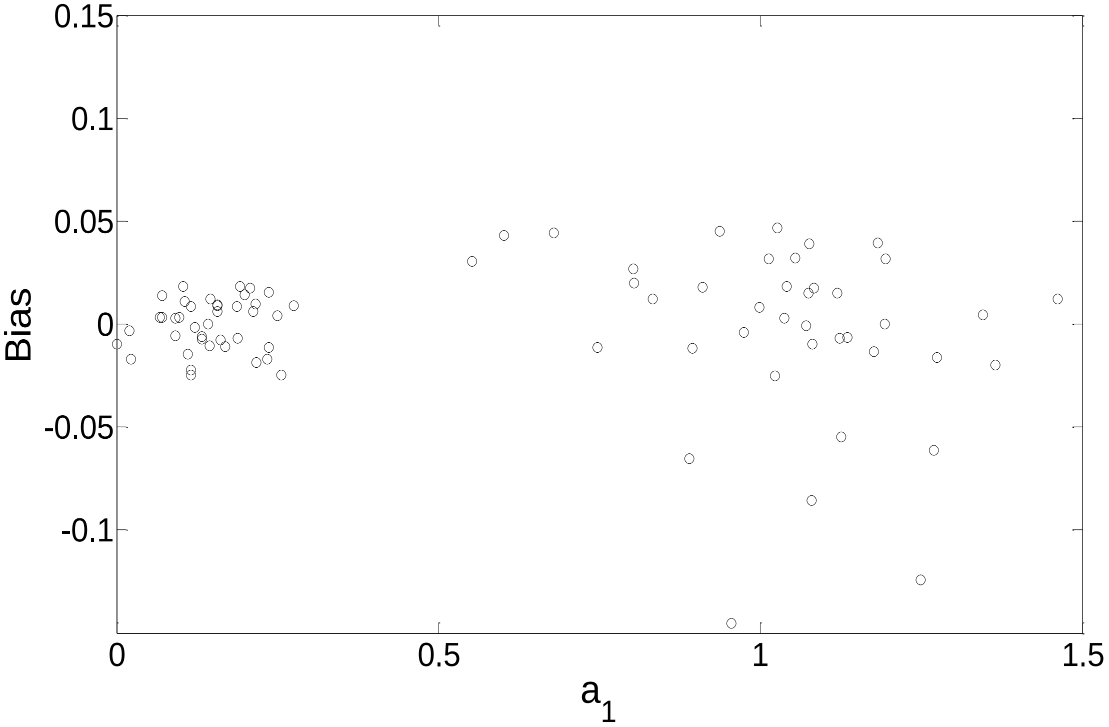

Figure 4.7. Bias for the Recovery of $a_1$-parameters for Method 1 at Zero Proficiency Correlation Level in Part I

dominantly measure each of the two dimensions. It seems that there is no clear pattern for the bias in the estimation for the $a_1$-parameters of small values, while the $a_1$-parameters of large values tend to be underestimated. The plots for all other dimensions, methods and conditions are similar to this one.

The RMSE values for the **a**-parameter recovery are listed in Table 4.10. According to the table, the value of RMSE increases as the proficiency correlation increases. Since "the observed correlations among the item scores will be accounted for solely by the *a*-parameters (when the proficiency correlation is forced to zero, for example, in the TESTFACT software)" (Reckase, 1997b, p. 275), it seems that as the proficiency correlation increases, it becomes more difficult to separate the effect of proficiency correlation from the estimation of **a**-parameters even with the Procrustes rotation method to match generating parameters. It can be observed that Method 1 (partial content coverage with low difficulty items) gives the largest RMSE values, while Methods 6, 8 and 9 generally provide comparatively lower RMSE values for the recovery of *a*-parameters on any dimension than all other methods. Figures 4.8 and 4.9 show the plots of RMSE for the *a*-parameter recovery for each item selection method and for each dimension.

In a word, as the proficiency correlation increases, the correlation between the rotated estimates and parameters decreases and the deviation increases for the recovery of *a*-parameters on both dimensions. Comparatively, Methods 6, 8 and 9 give slightly higher correlation and lower RMSE than other methods. A little negative bias was found in the estimation of *a*-parameters of large values.

Table 4.10. RMSE for the Recovery of **a**-parameters in Part I

| | Dimension 1 | | | | Dimension 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | p0 | p0.4 | p0.6 | p0.8 | p0 | p0.4 | p0.6 | p0.8 |
| (1) 1D, Low | 0.0445 | 0.0468 | 0.0585 | 0.0793 | 0.0400 | 0.0438 | 0.0515 | 0.0708 |
| (2) 1D, Medium | 0.0436 | 0.0461 | 0.0535 | 0.0727 | 0.0398 | 0.0434 | 0.0510 | 0.0696 |
| (3) 1D, High | 0.0437 | 0.0479 | 0.0524 | 0.0725 | 0.0405 | 0.0435 | 0.0515 | 0.0714 |
| (4) 1D, All | 0.0432 | 0.0470 | 0.0526 | 0.0730 | 0.0401 | 0.0435 | 0.0510 | 0.0701 |
| (5) 2D, Low | 0.0440 | 0.0461 | 0.0552 | 0.0741 | 0.0402 | 0.0439 | 0.0503 | 0.0682 |
| (6) 2D, Medium | 0.0427 | 0.0455 | 0.0525 | 0.0719 | 0.0397 | 0.0437 | 0.0504 | 0.0681 |
| (7) 2D, High | 0.0426 | 0.0472 | 0.0524 | 0.0731 | 0.0404 | 0.0449 | 0.0510 | 0.0710 |
| (8) 2D, All | 0.0428 | 0.0458 | 0.0530 | 0.0729 | 0.0394 | 0.0438 | 0.0505 | 0.0686 |
| (9) Classical | 0.0428 | 0.0463 | 0.0525 | 0.0717 | 0.0398 | 0.0438 | 0.0505 | 0.0672 |

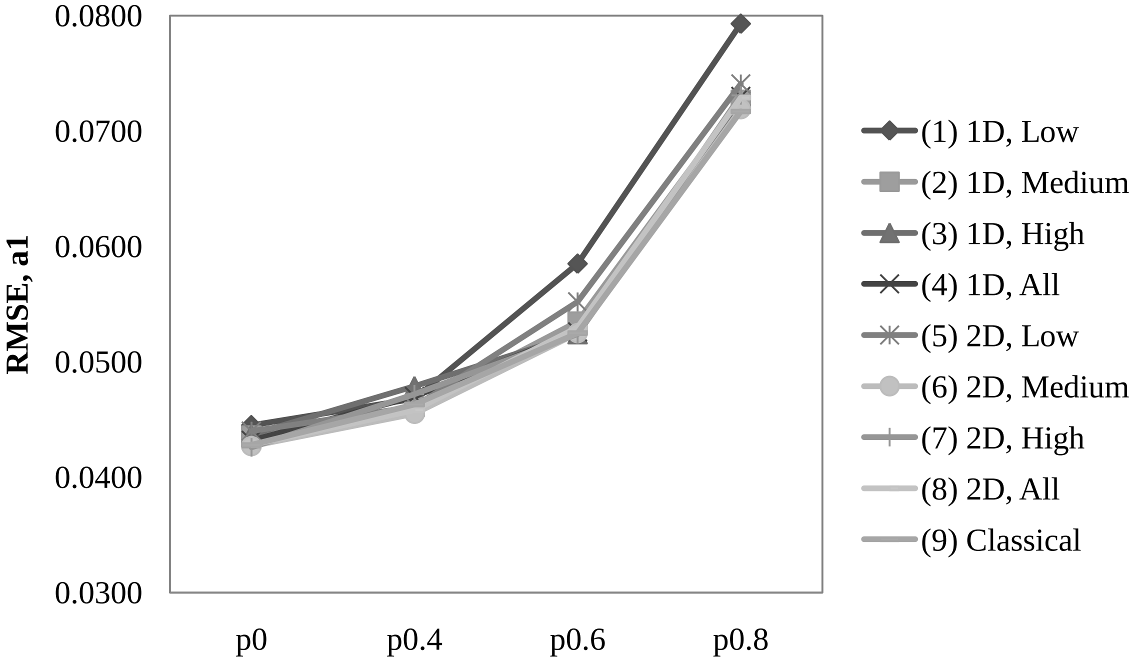

Figure 4.8. RMSE for the Recovery of $a_1$-parameters in Part I

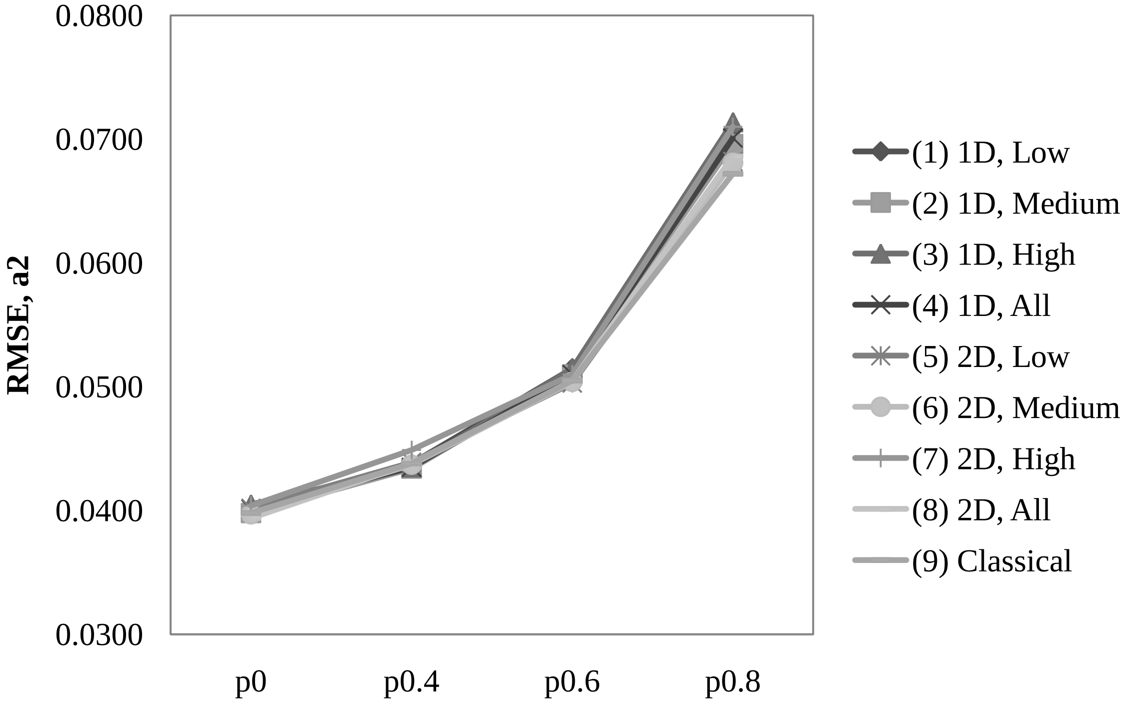

Figure 4.9. RMSE for the Recovery of $a_2$-parameters in Part I

### 4.3.3 Recovery of *d*-parameters

The person proficiency parameters were separately simulated from different multivariate normal distributions for the two grades. More specifically, the mean vector of the normal distribution is (-0.2, -0.2) for the lower grade and (0, 0) for the upper grade. However, the item estimates from the TESTFACT software were obtained by assuming that person proficiency coordinates follow a standard multivariate normal distribution with a zero mean vector, in order to facilitate the calibration and solve the indeterminacies in the MIRT model. Thus, it was expected that the item *MDIFF* estimates would be inflated by around 0.1, which would also directly lead to a negative bias for the *d* estimates from the TESTFACT software. The effect incurred by the inconsistency between proficiency distributions assumed for the generation and the estimation was minimized by matching the raw estimates with generating parameters via the aforementioned oblique Procrustes rotation method.

The correlation values between the adjusted estimates and the true values of *d*-parameters are listed in Table 4.11. From the table, all correlation values are close to one. Also, as the correlation between proficiencies increases, the correlation between the estimated and true *d*-parameters also increases, although sometimes an opposite pattern may occur between the correlation levels of 0.4 and 0.6. The correlation values for Method 6 are the largest among all the methods; besides, the values for Methods 8 and 9 also seem to be slightly larger than those for other methods.

Figure 4.10 shows the plot between the proficiency correlation level and the correlation value for the recovery of *d*-parameters for each item selection method. The points representing Methods 6, 8 and 9 are above all other points, while the points for the methods with partial content coverage are at the bottom. However, the difference between these values becomes smaller as the proficiency correlation increases.

Table 4.11. Correlation for the Recovery of *d*-parameters in Part I

| | p0 | P0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 1D, Low | 0.9938 | 0.9965 | 0.9962 | 0.9981 |
| (2) 1D, Medium | 0.9944 | 0.9966 | 0.9971 | 0.9983 |
| (3) 1D, High | 0.9943 | 0.9965 | 0.9967 | 0.9982 |
| (4) 1D, All | 0.9938 | 0.9964 | 0.9966 | 0.9983 |
| (5) 2D, Low | 0.9965 | 0.9979 | 0.9976 | 0.9983 |
| (6) 2D, Medium | 0.9976 | 0.9983 | 0.9981 | 0.9986 |
| (7) 2D, High | 0.9963 | 0.9975 | 0.9973 | 0.9985 |
| (8) 2D, All | 0.9971 | 0.9980 | 0.9978 | 0.9985 |
| (9) Classical | 0.9970 | 0.9983 | 0.9981 | 0.9986 |

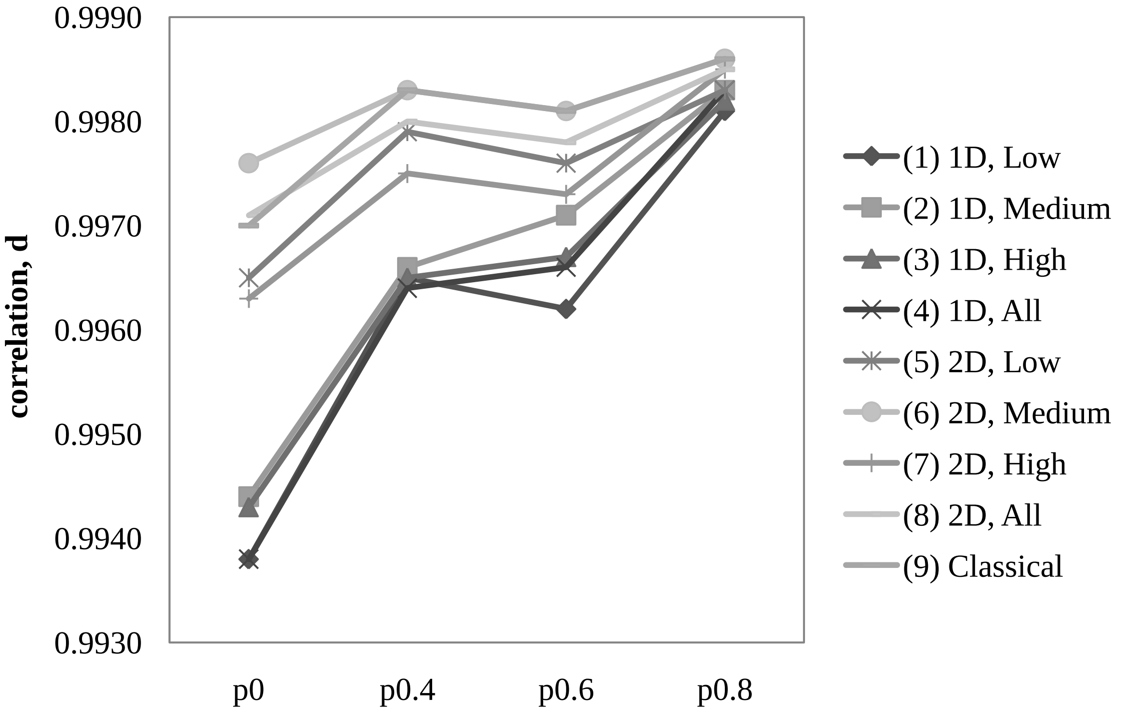

Figure 4.10. Correlation for the Recovery of *d*-parameters in Part I

The bias values for *d*-parameter estimates are listed in Table 4.12, and Figure 4.11 gives the plot between the proficiency correlation level and the bias for the *d*-parameter recovery for each item selection method. All the bias values are close to zero, which seems to indicate that there is no systematic error in the estimate. In order to examine the bias for parameters of different values, Figure 4.12 shows the plot between the bias and *d*-parameters for Method 1 under the zero proficiency correlation condition. From the figure, the *d*-parameters of large values tend to be underestimated and those of small values tend to be overestimated. The plots for all other methods and conditions are similar to this one.

Table 4.12. Bias for the Recovery of *d*-parameters in Part I

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 1D, Low | -0.0014 | -0.0003 | -0.0012 | -0.0004 |
| (2) 1D, Medium | -0.0009 | -0.0001 | -0.0005 | -0.0001 |
| (3) 1D, High | -0.0007 | 0.0002 | -0.0003 | 0.0001 |
| (4) 1D, All | -0.0009 | -0.0001 | -0.0006 | 0.0000 |
| (5) 2D, Low | -0.0010 | -0.0002 | -0.0007 | -0.0004 |
| (6) 2D, Medium | 0.0000 | 0.0003 | 0.0000 | 0.0001 |
| (7) 2D, High | 0.0004 | 0.0007 | 0.0003 | 0.0002 |
| (8) 2D, All | -0.0002 | 0.0002 | -0.0002 | 0.0000 |
| (9) Classical | -0.0003 | 0.0004 | -0.0001 | 0.0002 |

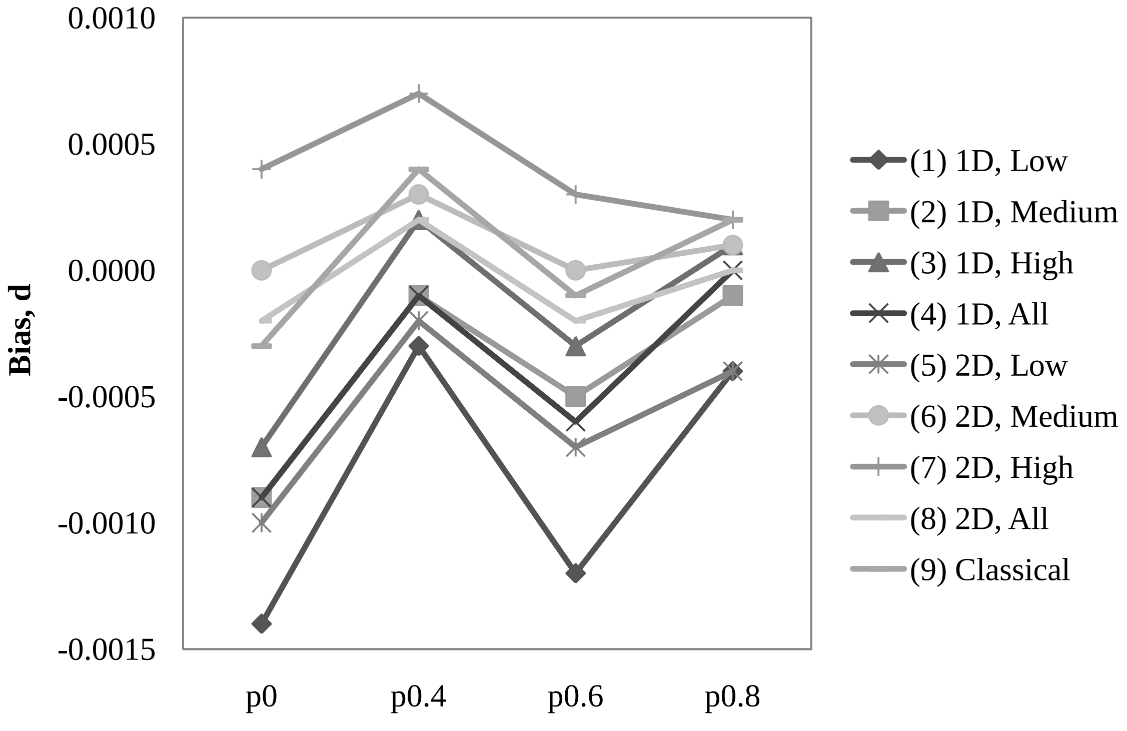

Figure 4.11. Bias for the Recovery of *d*-parameters in Part I

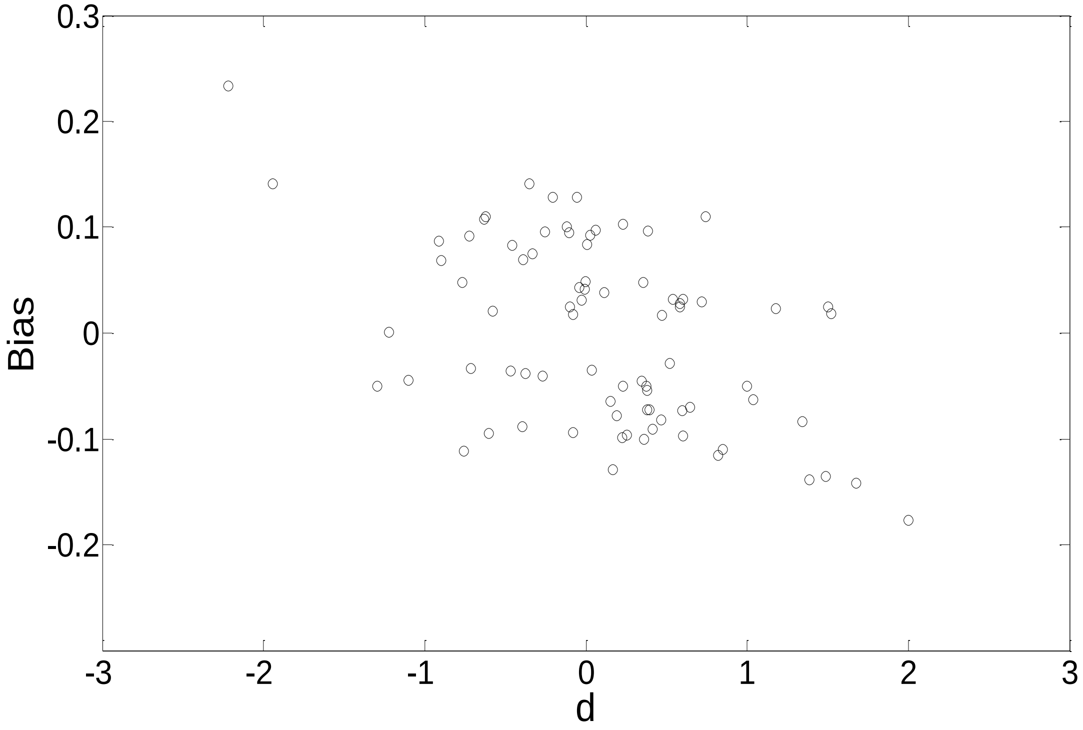

Figure 4.12. Bias for the Recovery of *d*-parameters for Method 1 at Zero Proficiency Correlation Level in Part I

Table 4.13 gives the RMSE values between the adjusted estimates and true values of *d*-parameters, and Figure 4.13 shows the plot of RMSE for the *d*-parameter recovery. As the proficiency correlation increases, the value of RMSE decreases, although sometimes an opposite pattern may occur between the correlation levels of 0.4 and 0.6. One reason to explain this non-monotonicity pattern may be that the mean differences between the simulated examinee proficiencies of the two grades at the 0.4 correlation level are the smallest among the four proficiency correlation levels, which may result in more accurate estimates with the single group MIRT calibration in TESTFACT. Also, this effect is more obvious when the common items cover all content domains.

The RMSE values for Methods 6, 8 and 9 are smaller, which indicates that these methods can give a better match between the estimated and true *d*-parameters. For all proficiency correlation levels, the methods with full content coverage yield lower RMSE than those with partial content coverage. However, the difference becomes smaller when the correlation between proficiencies increases.

Table 4.13. RMSE for the Recovery of *d*-parameters in Part I

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 1D, Low | 0.0826 | 0.0647 | 0.0680 | 0.0490 |
| (2) 1D, Medium | 0.0770 | 0.0627 | 0.0596 | 0.0453 |
| (3) 1D, High | 0.0786 | 0.0643 | 0.0630 | 0.0468 |
| (4) 1D, All | 0.0806 | 0.0648 | 0.0637 | 0.0462 |
| (5) 2D, Low | 0.0635 | 0.0524 | 0.0562 | 0.0463 |
| (6) 2D, Medium | 0.0540 | 0.0468 | 0.0496 | 0.0415 |
| (7) 2D, High | 0.0661 | 0.0564 | 0.0587 | 0.0440 |
| (8) 2D, All | 0.0590 | 0.0505 | 0.0528 | 0.0428 |
| (9) Classical | 0.0568 | 0.0461 | 0.0484 | 0.0406 |

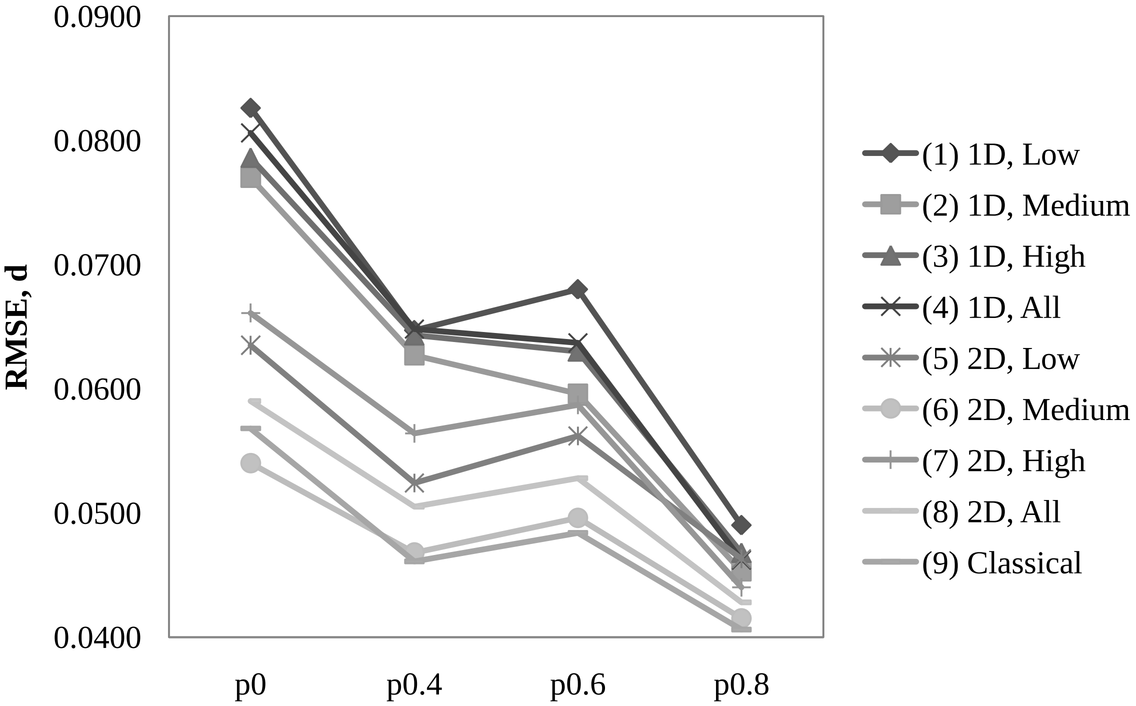

Figure 4.13. RMSE for the Recovery of *d*-parameters in Part I

### 4.3.4 Recovery of effect sizes

The person proficiency was estimated with the EAP scoring method that incorporates the prior distribution into the estimation. The observed covariance matrix for estimated proficiencies on different dimensions was compared with the true matrix. It was found that the variances were underestimated and the correlations were overestimated. This may be due to the EAP scoring method, which yields estimates biased towards the prior mean.

Table 4.14 shows the true effect sizes as well as the means and standard deviations of estimated effect sizes on both dimensions for all methods and for all proficiency correlation levels. From the table, the effect size is underestimated for any dimension and for any method. For Methods 1-4 with partial content coverage, when the proficiency correlation is low, the effect size estimates on Dimension 1 are slightly better than all other methods, while those on Dimension 2 are highly negatively biased. However, both the advantage on Dimension 1 and disadvantage on Dimension 2 tend to diminish as the proficiency correlation increases. Compared with these methods, Methods 5-8 provide slightly worse estimates on Dimension 1 but much better ones on Dimension 2. Also, Method 9 gives the best results on Dimension 2, which may be due to the fact that eight out of ten common items in this method are from Cluster 2. Therefore, one conclusion is that the effect size recovery on each dimension largely depends on the number of common items measuring that dimension.

For the same content coverage, the method with medium difficulty items yields the best results, followed by the method with items from all difficulty levels. However, as the correlation between proficiencies increases, the difference between the estimated effect sizes from all selection methods decreases. Generally speaking, in consideration of a good recovery on both dimensions, Methods 6 and 9 perform the best among all methods.

The standard deviations are quite small, especially for Methods 1-4 and when the proficiency correlation is very low. This indicates that the estimates of effect sizes are fairly stable across replications and the difference across methods is substantial in consideration of random errors.

Figures 4.14 and 4.15 show the recovery of effect sizes for diferent methods. From both figures, when the proficiency correlation is 0.8, the points representing the estimates from different methods are all clustered together. One explanation is that the proficiency estimation on one dimension can 'borrow' information from other dimensions in the MIRT calibration and more information can be 'borrowed' when the correlation between them is high. However, from Figure 4.15 for the effect size recovery on Dimension 2, when the proficiency correlation is low, the points representing methods with partial content coverage are much lower than those for other methods.

Table 4.14. Recovery of Effect Sizes for Proficiencies in Part I

| | Dimension 1 | | | | Dimension 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | p0 | p0.4 | p0.6 | p0.8 | p0 | p0.4 | p0.6 | p0.8 |
| TRUE | 0.2315 | 0.1474 | 0.1788 | 0.1936 | 0.1891 | 0.1616 | 0.1841 | 0.1954 |
| (1) 1D, Low | 0.1940 | 0.1226 | 0.1384 | 0.1679 | -0.0022 | 0.0294 | 0.0649 | 0.1428 |
| Std | *0.0131* | *0.0092* | *0.0104* | *0.0120* | *0.0044* | *0.0028* | *0.0083* | *0.0101* |
| (2) 1D, Medium | 0.2165 | 0.1264 | 0.1586 | 0.1781 | -0.0030 | 0.0306 | 0.0801 | 0.1516 |
| Std | *0.0109* | *0.0097* | *0.0098* | *0.0112* | *0.0041* | *0.0047* | *0.0074* | *0.0098* |
| (3) 1D, High | 0.2017 | 0.1155 | 0.1439 | 0.1684 | 0.0027 | 0.0305 | 0.0736 | 0.1441 |
| Std | *0.0110* | *0.0085* | *0.0090* | *0.0124* | *0.0044* | *0.0039* | *0.0070* | *0.0103* |
| (4) 1D, All | 0.2087 | 0.1236 | 0.1487 | 0.1753 | -0.0112 | 0.0243 | 0.0692 | 0.1477 |
| Std | *0.0136* | *0.0069* | *0.0088* | *0.0106* | *0.0029* | *0.0031* | *0.0056* | *0.0091* |
| (5) 2D, Low | 0.1501 | 0.1097 | 0.1305 | 0.1619 | 0.1348 | 0.1072 | 0.1322 | 0.1632 |
| Std | *0.0114* | *0.0129* | *0.0108* | *0.0102* | *0.0124* | *0.0121* | *0.0118* | *0.0093* |
| (6) 2D, Medium | 0.1757 | 0.1176 | 0.1452 | 0.1758 | 0.1490 | 0.1272 | 0.1490 | 0.1772 |
| Std | *0.0113* | *0.0106* | *0.0106* | *0.0104* | *0.0109* | *0.0104* | *0.0122* | *0.0121* |
| (7) 2D, High | 0.1471 | 0.0870 | 0.1181 | 0.1645 | 0.1159 | 0.0971 | 0.1200 | 0.1634 |
| Std | *0.0129* | *0.0107* | *0.0147* | *0.0095* | *0.0139* | *0.0122* | *0.0136* | *0.0125* |
| (8) 2D, All | 0.1632 | 0.1068 | 0.1380 | 0.1734 | 0.1395 | 0.1175 | 0.1367 | 0.1696 |
| Std | *0.0115* | *0.0126* | *0.0114* | *0.0124* | *0.0161* | *0.0114* | *0.0108* | *0.0101* |
| (9) Classical | 0.1400 | 0.1048 | 0.1374 | 0.1760 | 0.1854 | 0.1478 | 0.1691 | 0.1877 |
| Std | *0.0137* | *0.0084* | *0.0091* | *0.0107* | *0.0100* | *0.0078* | *0.0112* | *0.0094* |

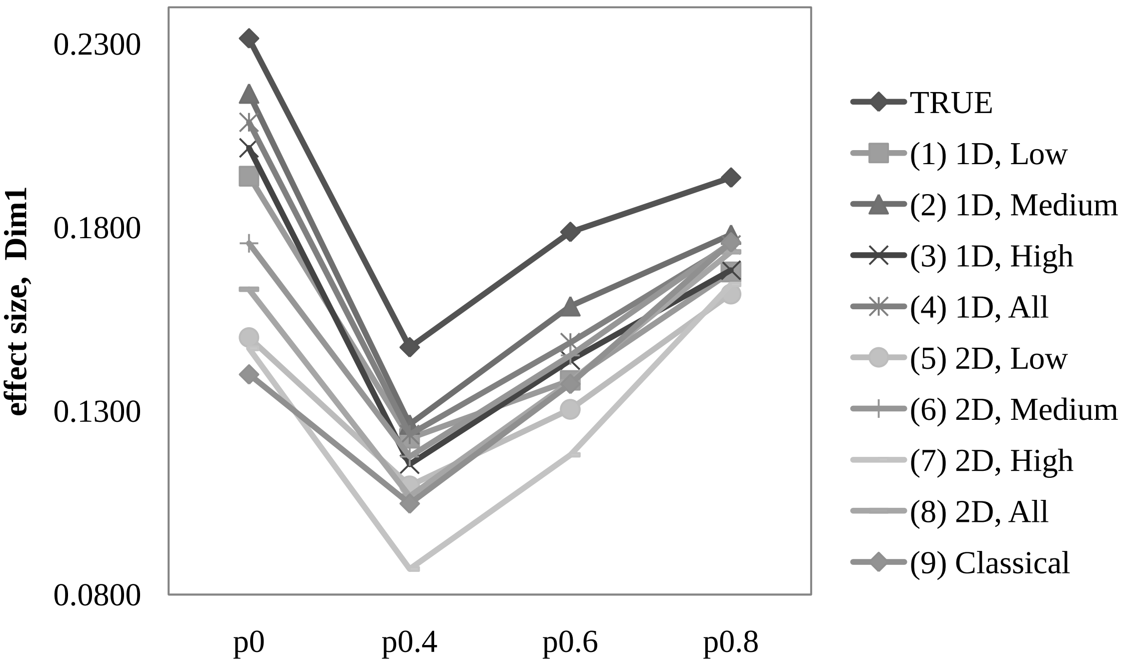

Figure 4.14. Recovery of Effect Size for the Proficiency on Dimension 1 in Part I

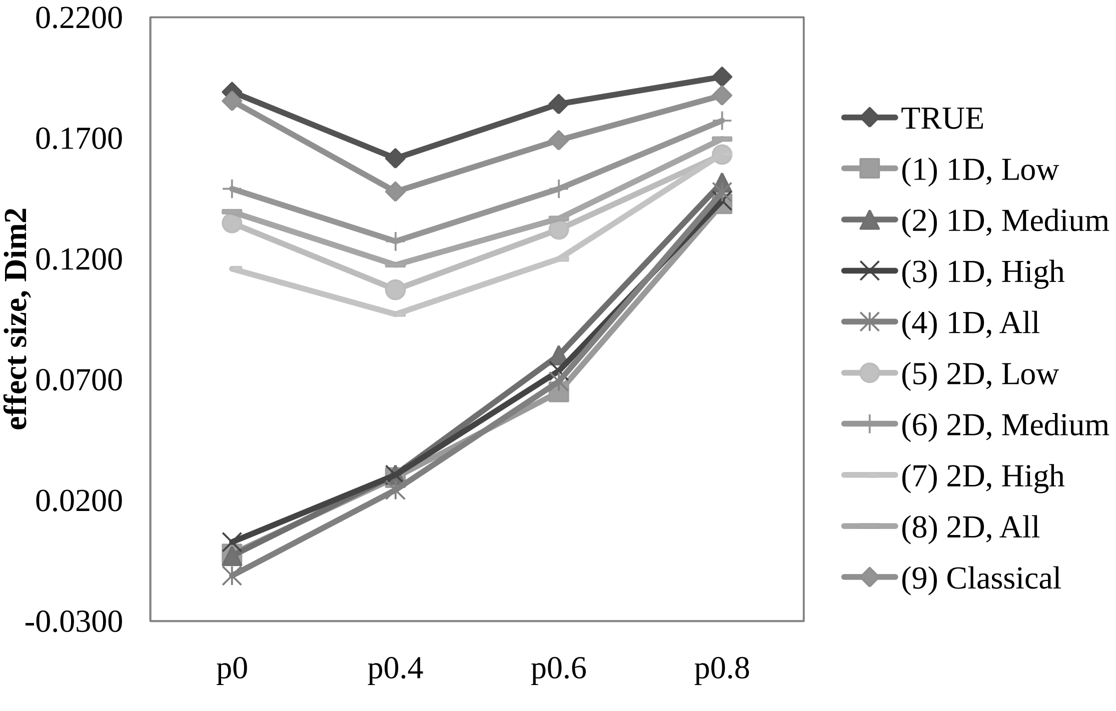

Figure 4.15. Recovery of Effect Size for the Proficiency on Dimension 2 in Part I

# CHAPTER 5

# PART II: DIFFERENT CONSTRUCTS

## 5.1 Parameters and Designs

In this part, two constructs are measured in the lower grade test; besides these two, one more construct is measured in the upper grade test. There are 40 unique items in the test of each grade. Additionally, 10 common items in both tests are selected according to four different methods. For different grades, person proficiencies are simulated from the multivariate normal distributions with different mean vectors and variance-covariance matrices.

### 5.1.1 Unique items

The numbers of unique items in different content domains were chosen to be the same as that in the study by Reckase and Li (2007). With the context from that study, the allocation of unique items in the test of each grade is shown in Table 5.1. Note that there are no algebra items in the lower grade test and the numbers of unique items in different content domains are not balanced.

Table 5.1. Allocation of Unique Items in Different Content Domains and Grades in Part II

| Grade | Arithmetic | Problem Solving | Algebra |
|---|---|---|---|
| Lower Grade | 17 | 23 | 0 |
| Upper Grade | 11 | 18 | 11 |

The cluster number, item parameters, multidimensional difficulty, multidimensional discrimination and direction angles of unique items for tests in both grades are listed in Tables 5.2 and 5.3. The means of *MDISC*s are quite similar for different grade levels; however, the mean and standard deviation of *MDIFFs* are -0.2 and 0.78 for the lower grade, while they are 0 and .67 for the upper grade. As can be observed, the $a_3$-parameters are always of small values

Table 5.2. Unique Item Parameters and Statistics for Lower Grade in Part II

| Cluster | $a_1$ | $a_2$ | $a_3$ | $d$ | *MDIFF* | *MDISC* | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.88 | 0.01 | 0.02 | 0.51 | -0.58 | 0.88 | 2 | 89 | 89 |
| 1 | 0.94 | 0.01 | 0.04 | -0.20 | 0.21 | 0.94 | 3 | 90 | 87 |
| 1 | 0.91 | 0.03 | 0.20 | 0.58 | -0.63 | 0.93 | 12 | 88 | 78 |
| 1 | 0.89 | 0.03 | 0.22 | -0.04 | 0.05 | 0.92 | 14 | 88 | 76 |
| 1 | 0.91 | 0.01 | 0.04 | 0.95 | -1.05 | 0.91 | 2 | 90 | 88 |
| 1 | 0.77 | 0.03 | 0.13 | 0.21 | -0.26 | 0.78 | 10 | 88 | 80 |
| 1 | 1.08 | 0.06 | 0.09 | -0.55 | 0.51 | 1.09 | 6 | 87 | 85 |
| 1 | 1.08 | 0.04 | 0.05 | -1.12 | 1.03 | 1.08 | 3 | 88 | 87 |
| 1 | 1.09 | 0.15 | 0.22 | -0.59 | 0.53 | 1.12 | 14 | 82 | 79 |
| 1 | 0.97 | 0.00 | 0.01 | 0.84 | -0.87 | 0.97 | 1 | 90 | 89 |
| 1 | 1.20 | 0.01 | 0.18 | 0.95 | -0.78 | 1.21 | 8 | 89 | 82 |
| 1 | 0.90 | 0.10 | 0.18 | -0.99 | 1.07 | 0.92 | 13 | 84 | 79 |
| 1 | 1.06 | 0.14 | 0.11 | -0.02 | 0.02 | 1.07 | 10 | 82 | 84 |
| 1 | 1.35 | 0.03 | 0.02 | 1.00 | -0.74 | 1.35 | 1 | 89 | 89 |
| 1 | 0.92 | 0.17 | 0.12 | 0.73 | -0.77 | 0.94 | 13 | 79 | 83 |
| 1 | 1.03 | 0.01 | 0.25 | 0.17 | -0.16 | 1.06 | 14 | 89 | 76 |
| 1 | 1.38 | 0.00 | 0.00 | 1.33 | -0.97 | 1.38 | 0 | 90 | 90 |
| 2 | 0.10 | 0.94 | 0.14 | 0.88 | -0.91 | 0.96 | 84 | 10 | 82 |
| 2 | 0.17 | 0.99 | 0.18 | -1.13 | 1.11 | 1.02 | 81 | 14 | 80 |
| 2 | 0.02 | 1.20 | 0.03 | 0.56 | -0.47 | 1.20 | 89 | 2 | 88 |
| 2 | 0.01 | 0.79 | 0.20 | 1.38 | -1.68 | 0.82 | 90 | 14 | 76 |
| 2 | 0.07 | 1.43 | 0.12 | -1.62 | 1.13 | 1.43 | 87 | 6 | 85 |
| 2 | 0.02 | 0.68 | 0.10 | 0.58 | -0.85 | 0.69 | 88 | 8 | 82 |
| 2 | 0.01 | 0.87 | 0.13 | -0.53 | 0.60 | 0.88 | 89 | 9 | 81 |
| 2 | 0.01 | 0.98 | 0.24 | -0.39 | 0.39 | 1.01 | 90 | 14 | 76 |
| 2 | 0.11 | 0.98 | 0.07 | -0.97 | 0.98 | 0.99 | 83 | 8 | 86 |
| 2 | 0.06 | 1.25 | 0.18 | 0.60 | -0.47 | 1.26 | 87 | 9 | 82 |
| 2 | 0.18 | 1.16 | 0.10 | 0.78 | -0.66 | 1.18 | 81 | 10 | 85 |
| 2 | 0.14 | 1.13 | 0.05 | 1.17 | -1.02 | 1.14 | 83 | 8 | 87 |
| 2 | 0.02 | 0.79 | 0.06 | -0.59 | 0.75 | 0.79 | 89 | 4 | 86 |
| 2 | 0.09 | 0.69 | 0.16 | 0.80 | -1.13 | 0.71 | 83 | 15 | 77 |
| 2 | 0.11 | 0.89 | 0.11 | 0.73 | -0.81 | 0.90 | 83 | 10 | 83 |
| 2 | 0.15 | 0.86 | 0.03 | 0.17 | -0.19 | 0.87 | 80 | 10 | 88 |
| 2 | 0.00 | 0.89 | 0.13 | 0.53 | -0.58 | 0.90 | 90 | 8 | 82 |
| 2 | 0.06 | 0.93 | 0.12 | 0.91 | -0.97 | 0.94 | 86 | 8 | 83 |
| 2 | 0.01 | 1.47 | 0.01 | 0.37 | -0.25 | 1.47 | 90 | 0 | 90 |
| 2 | 0.10 | 1.01 | 0.03 | -0.81 | 0.80 | 1.01 | 84 | 6 | 89 |
| 2 | 0.02 | 1.20 | 0.04 | 1.52 | -1.26 | 1.20 | 89 | 2 | 88 |
| 2 | 0.00 | 0.83 | 0.02 | -0.23 | 0.28 | 0.83 | 90 | 1 | 89 |
| 2 | 0.00 | 1.38 | 0.02 | -0.89 | 0.64 | 1.38 | 90 | 1 | 89 |
| Mean | 0.47 | 0.61 | 0.10 | 0.19 | -0.20 | 1.03 | | | |
| Std | 0.49 | 0.51 | 0.07 | 0.81 | 0.78 | 0.20 | | | |

Table 5.3. Unique Item Parameters and Statistics for Upper Grade in Part II

| Cluster | $a_1$ | $a_2$ | $a_3$ | $d$ | *MDIFF* | *MDISC* | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.08 | 0.04 | 0.16 | -0.23 | 0.21 | 1.09 | 9 | 88 | 81 |
| 1 | 1.00 | 0.19 | 0.19 | -0.14 | 0.14 | 1.03 | 15 | 79 | 80 |
| 1 | 0.67 | 0.01 | 0.04 | -0.69 | 1.02 | 0.67 | 3 | 89 | 87 |
| 1 | 0.81 | 0.10 | 0.18 | 0.10 | -0.12 | 0.84 | 14 | 83 | 78 |
| 1 | 0.94 | 0.06 | 0.21 | 0.86 | -0.89 | 0.96 | 13 | 86 | 78 |
| 1 | 0.97 | 0.07 | 0.04 | 0.88 | -0.90 | 0.98 | 5 | 86 | 88 |
| 1 | 0.67 | 0.05 | 0.04 | -0.15 | 0.22 | 0.67 | 5 | 86 | 87 |
| 1 | 0.97 | 0.03 | 0.09 | -0.50 | 0.52 | 0.97 | 5 | 88 | 85 |
| 1 | 0.87 | 0.04 | 0.02 | 0.31 | -0.36 | 0.87 | 3 | 87 | 89 |
| 1 | 1.04 | 0.08 | 0.08 | -0.09 | 0.09 | 1.04 | 6 | 85 | 86 |
| 1 | 0.93 | 0.11 | 0.03 | -0.07 | 0.07 | 0.94 | 7 | 83 | 88 |
| 2 | 0.02 | 0.77 | 0.15 | -0.48 | 0.61 | 0.79 | 88 | 11 | 79 |
| 2 | 0.07 | 0.97 | 0.14 | 0.09 | -0.09 | 0.98 | 86 | 9 | 81 |
| 2 | 0.05 | 1.16 | 0.18 | -0.48 | 0.41 | 1.18 | 87 | 9 | 81 |
| 2 | 0.01 | 0.89 | 0.15 | -1.09 | 1.21 | 0.90 | 89 | 10 | 80 |
| 2 | 0.00 | 1.18 | 0.01 | -0.27 | 0.23 | 1.18 | 90 | 0 | 90 |
| 2 | 0.08 | 0.83 | 0.04 | -0.26 | 0.31 | 0.83 | 85 | 6 | 87 |
| 2 | 0.07 | 0.78 | 0.05 | 0.45 | -0.57 | 0.78 | 85 | 6 | 86 |
| 2 | 0.00 | 1.18 | 0.17 | 0.73 | -0.61 | 1.19 | 90 | 8 | 82 |
| 2 | 0.03 | 1.11 | 0.15 | -0.59 | 0.53 | 1.13 | 88 | 8 | 82 |
| 2 | 0.10 | 1.19 | 0.12 | 0.68 | -0.57 | 1.20 | 85 | 7 | 84 |
| 2 | 0.05 | 1.09 | 0.28 | 0.78 | -0.70 | 1.12 | 87 | 15 | 75 |
| 2 | 0.00 | 1.14 | 0.14 | -0.99 | 0.86 | 1.15 | 90 | 7 | 83 |
| 2 | 0.01 | 1.16 | 0.03 | -0.75 | 0.65 | 1.16 | 89 | 1 | 89 |
| 2 | 0.11 | 0.87 | 0.16 | -0.32 | 0.36 | 0.89 | 83 | 13 | 79 |
| 2 | 0.00 | 0.95 | 0.17 | 0.62 | -0.64 | 0.97 | 90 | 10 | 80 |
| 2 | 0.13 | 1.19 | 0.07 | 1.18 | -0.99 | 1.20 | 84 | 7 | 87 |
| 2 | 0.03 | 0.76 | 0.05 | 0.64 | -0.84 | 0.76 | 88 | 5 | 86 |
| 2 | 0.05 | 0.94 | 0.13 | 0.16 | -0.17 | 0.95 | 87 | 9 | 82 |
| 3 | 0.05 | 0.04 | 1.19 | -0.62 | 0.52 | 1.19 | 87 | 88 | 3 |
| 3 | 0.06 | 0.22 | 0.91 | -0.65 | 0.69 | 0.94 | 86 | 76 | 14 |
| 3 | 0.27 | 0.09 | 1.09 | 0.03 | -0.03 | 1.12 | 76 | 86 | 15 |
| 3 | 0.08 | 0.06 | 1.52 | 0.70 | -0.46 | 1.53 | 87 | 88 | 3 |
| 3 | 0.12 | 0.02 | 0.95 | 0.02 | -0.02 | 0.96 | 83 | 89 | 7 |
| 3 | 0.03 | 0.14 | 1.02 | -0.99 | 0.96 | 1.03 | 88 | 82 | 8 |
| 3 | 0.15 | 0.04 | 1.02 | -0.08 | 0.08 | 1.03 | 82 | 88 | 9 |
| 3 | 0.02 | 0.01 | 0.71 | 0.72 | -1.01 | 0.71 | 88 | 89 | 2 |
| 3 | 0.06 | 0.03 | 0.79 | 1.31 | -1.65 | 0.79 | 86 | 88 | 5 |
| 3 | 0.09 | 0.16 | 0.92 | 0.21 | -0.23 | 0.94 | 85 | 80 | 11 |
| 3 | 0.08 | 0.10 | 0.76 | -0.89 | 1.15 | 0.77 | 84 | 82 | 9 |
| Mean | 0.29 | 0.50 | 0.35 | 0.00 | 0.00 | 0.99 | | | |
| Std | 0.39 | 0.48 | 0.42 | 0.63 | 0.67 | 0.18 | | | |

for the lower grade, which indicates that this content domain is not designed to be measured by unique items of the test in this grade.

### 5.1.2 Common items

An item pool with 100 items was generated in a similar way as the unique items for both grade levels. Items in this pool were divided into several categories according to three content domains and three difficulty levels as shown in Table 5.4.

Table 5.4. Number of Items for Different Content and Difficulty Categories in Item Pool of Part II

| | Low | Medium | High | All |
|---|---|---|---|---|
| Dimension 1 | 14 | 14 | 14 | 42 |
| Dimension 2 | 14 | 14 | 14 | 42 |
| Dimension 3 | 5 | 6 | 5 | 16 |

Ten common items were selected according to the MIRT methods and the classical correlation method. The details for each method are shown below.

(1) The MIRT methods consist of three methods with different numbers of items selected from the three content domains in order to achieve different degrees of content coverage. For simplicity, common items in these MIRT methods are only selected from medium difficulty items, in view of the results in Part I.

- Method 1: Common items are selected from all content domains. The numbers of items from the three item clusters are four, four and three, respectively. Note that the third content domain is only taught in the upper grade.
- Method 2: Common items are only selected from the first two content domains that are measured by unique items in both tests. Six items are selected from Cluster 1 and four from Cluster 2. Given that the proficiency on Dimension 1 is manipulated to have

a higher correlation with that on Dimension 3, more items are selected from Cluster 1 to replace the items in Cluster 3 that are missing from the common item set.

- Method 3: Common items are only selected from the first two content domains, with four items from Cluster 1 and six from Cluster 2 according to their proportions in the unique items. In order to make results more comparable, Methods 2 and 3 share the same eight common items with four from each cluster.

(2) Method 4, the classical correlation method, chooses common items based on the high item-total-test correlation. The detailed procedure is the same as in Part I.

The statistics for different common item sets are listed in Table 5.5. For each method, the mean of *MDIFF* values is close to zero and the standard deviation is small. From the perspective of MIRT, common items in the classical correlation method appeared to be also selected from the medium difficulty level but with an extremely unbalanced coverage for the three content domains. Almost all the common items in this method were selected from Cluster 2, which is reasonable in that the number of unique items from this cluster is the largest in both tests.

Table 5.5. Statistics of Common Items for Different Selection Methods in Part II

| Selection Method | MDIFF Mean | MDIFF Std | # of items in Dim 1 | # of items in Dim 2 | # of items in Dim 3 |
|---|---|---|---|---|---|
| (1) 3D | -0.11 | 0.25 | 4 | 3 | 3 |
| (2) 2D, Correlation | -0.07 | 0.24 | 6 | 4 | 0 |
| (3) 2D, Proportion | -0.01 | 0.18 | 4 | 6 | 0 |
| (4) Classical | -0.14 | 0.29 | 1 | 9 | 0 |

### 5.1.3 Person parameters

For different grades, the person proficiency parameters were simulated according to the multivariate normal distributions with different mean vectors and variance-covariance matrices.

The mean vectors for person proficiency distribution in each grade are shown in Table 5.6. The variance was set to one for all the proficiencies except for that on Dimension 3 for the lower grade examinees. That proficiency was set to have a much lower mean and smaller variation since examinees in the lower grade were not supposed to have knowledge in this content domain. Note that although the mean differences between examinee proficiencies of the two grades were set to 0.2 in Part I, they were set to 0.7 for the first two dimensions in this part, following the study by Reckase and Li (2007). Also, the correlation matrices for both grades were manipulated to be roughly the same as those in that study.

The variance-covariance matrices for person proficiency distributions in both grades are shown in Tables 5.7 and 5.8. In the lower grade, the correlation between proficiencies on the first two dimensions was 0.7; however, no correlation was assumed between proficiencies on the third and any of the first two dimensions. For the upper grade, the correlations between proficiencies on the second and the other two dimensions were fixed at 0.24 and 0.32 respectively. However, the correlation between proficiencies on the first and third dimensions was manipulated to vary from 0, 0.4, 0.6 to 0.8, in order to check the effect of different correlation levels on the linking results. There was concern that when the correlation was zero, according to the logic of Method 2, more items were supposed to be selected from Cluster 2 instead of Cluster 1. Nevertheless, for simplicity and consistency, this method always selected more items from Cluster 1 for all correlation levels.

Table 5.6. Mean Vectors for Proficiency Distributions of Lower and Upper Grade Examinees in Part II

| Grade | Arithmetic | Problem Solving | Algebra |
|---|---|---|---|
| Lower Grade | -0.5 | -0.7 | -1.5 |
| Upper Grade | 0.2 | 0 | 0 |

Table 5.7. Variance-Covariance Matrix for Proficiency Distribution of Lower Grade Examinees in Part II

| | Arithmetic | Problem Solving | Algebra |
|---|---|---|---|
| Arithmetic | 1 | | |
| Problem solving | 0.7 | 1 | |
| Algebra | 0 | 0 | 0.25 |

Table 5.8. Variance-Covariance Matrix for Proficiency Distribution of Upper Grade Examinees in Part II

| | Arithmetic | Problem Solving | Algebra |
|---|---|---|---|
| Arithmetic | 1 | | |
| Problem solving | 0.24 | 1 | |
| Algebra | **0.6** | 0.32 | 1 |

## 5.2 Estimation

Fifty replications of response matrix were simulated for each proficiency correlation level and the MIRT calibration for each selection method was conducted on the data matrix for the selected common items and all unique items. This resulted in a total of 800 computer runs (4 proficiency correlation levels x 50 replications x 4 item selection methods). The MIRT calibration took more than one hour for each TESTFACT run.

The data layout and TESTFACT syntax in this part are similar to those in Part I, except that a three-dimensional instead of two-dimensional solution was requested for the MIRT calibration. As in Part I, item **a**-parameter estimates from the TESTFACT software were corrected by forcing the mean on each dimension to be positive. However, one problem was found in the sign correction on Dimension 3. It may happen that both positive and negative item discrimination estimates on that dimension have large absolute values, which may be partly due to the low proficiency on that dimension for examinees in the lower grade. Therefore, it was difficult to identify which set of item discrimination estimates, negated or non-negated, was ‘correct’.

Furthermore, because of the weak relationship between the percentage of correct responses and the proficiency estimate on Dimension 3, it was also hard to rely on the TESTFACT software or apply the previous checking method for the sign correction to proficiency estimates on that dimension. The solution to the sign indeterminacies in both item and person estimates was to try all four sign combinations for the proficiencies on Dimension 3, which is to keep it unchanged or changed for either grade, in an attempt to make the item and person estimates a valid pair for the MIRT calibration. Combined with the item estimates, the ‘correct’ signs for person proficiency estimates were selected as the combination that gave the largest correlation value for the recovery of probability matrix and these adjusted estimates were used for further analysis.

## 5.3 Results

### 5.3.1 Recovery of probability matrix

The average correlation values between corresponding elements in the estimated and true probability matrices are listed in Table 5.9. As the proficiency correlation increases, so does the correlation between the true and estimated probabilities for any item selection method. Although Method 3 could give comparatively higher values and Method 4 provides lower values, the differences are very small.

Table 5.9. Correlation for the Recovery of Probability Matrix in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | 0.9641 | 0.9651 | 0.9654 | 0.9664 |
| (2) 2D, Correlation | 0.9639 | 0.9649 | 0.9652 | 0.9664 |
| (3) 2D, Proportion | 0.9641 | 0.9651 | 0.9655 | 0.9669 |
| (4) Classical | 0.9635 | 0.9646 | 0.9651 | 0.9664 |

Tables 5.10 and 5.11 give the bias and RMSE values for the recovery of probability matrix. All methods yield slightly negative biased estimates. With a further examination on the plot for

bias at different parameter values, it was found that, as observed in Part I, the probabilities of large values tend to be underestimated and those of small values tend to be overestimated. However, there are much more points representing the negative bias for the true probabilities with large values than points representing the positive bias for probabilities with small values, and the points representing negative bias are mostly for lower grade examinees. This is reasonable because the means of proficiencies of lower grade examinees are much lower than the mean of multidimensional difficulties for lower grade items and it was already explained in Part I that difficult items are more likely to lead to negative bias for the proficiencies with large values. Also, as the proficiency correlation increases, the RMSE value decreases for all methods. Therefore, it seems to be a general conclusion that as the proficiency correlation increases, the fit between the estimated and true probability matrices becomes better. Furthermore, Methods 3 and 4 give the lower RMSE values for the recovery of probability matrix than the other two methods. These could also be observed in Figure 5.1, which shows the plot between the proficiency correlation level and the RMSE value for each selection method.

Table 5.10. Bias for the Recovery of Probability Matrix in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | -0.0015 | -0.0009 | -0.0011 | -0.0008 |
| (2) 2D, Correlation | -0.0011 | -0.0009 | -0.0013 | -0.0012 |
| (3) 2D, Proportion | -0.0011 | -0.0010 | -0.0011 | -0.0010 |
| (4) Classical | -0.0009 | -0.0007 | -0.0011 | -0.0009 |

Table 5.11. RMSE for the Recovery of Probability Matrix in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | 0.0763 | 0.0755 | 0.0751 | 0.0740 |
| (2) 2D, Correlation | 0.0762 | 0.0754 | 0.0751 | 0.0738 |
| (3) 2D, Proportion | 0.0759 | 0.0751 | 0.0747 | 0.0732 |
| (4) Classical | 0.0760 | 0.0751 | 0.0747 | 0.0733 |

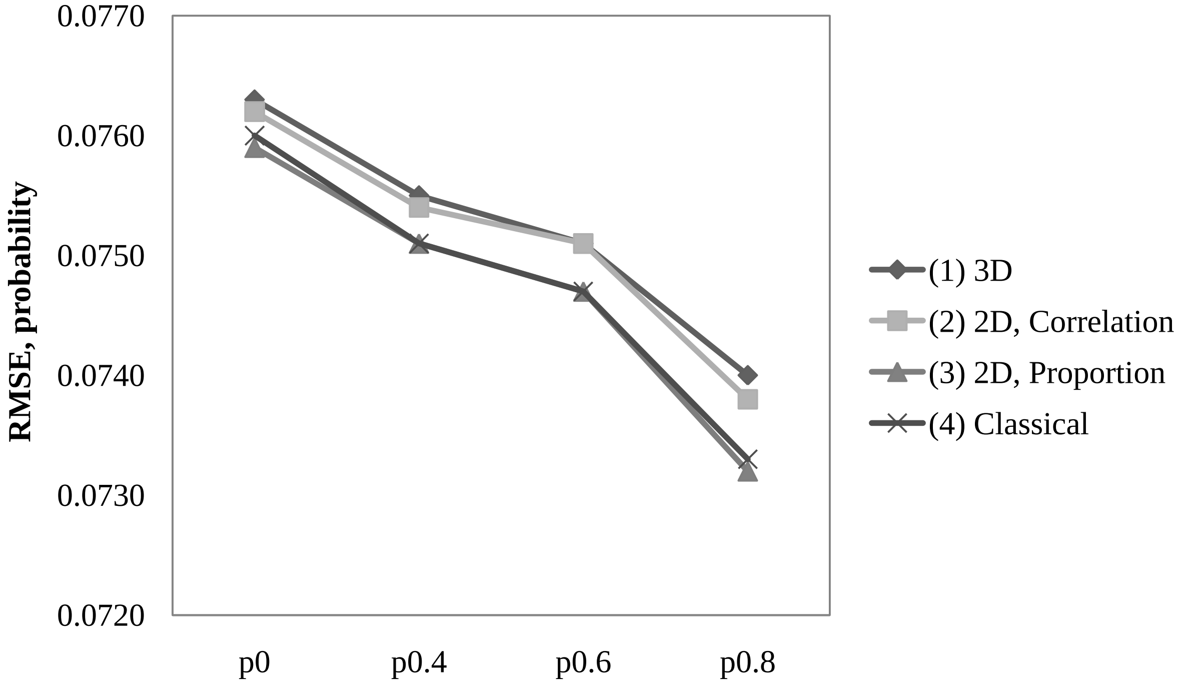

Figure 5.1. RMSE for the Recovery of Probability Matrix in Part II

### 5.3.2 Recovery of a-parameters

The average correlation values between the rotated estimates and true parameters are shown in Table 5.12. Generally speaking, as the correlation between proficiencies increases, the correlation between the rotated estimates and the parameters decreases. This pattern is opposite to that observed in the recovery of the probability matrix. As already explained in Part I, the reason is that as the proficiency correlation increases, it is more difficult to separate its effect from the item estimates obtained with the constraint of identity variance-covariance matrix. Method 1 gives the highest correlation value for the recovery of $a_3$-parameters, which leads to the conclusion that items from all content domains should be included in the common item set for a good recovery of item discrimination parameters on all dimensions. Method 4 does not perform as well as the other three methods under almost all conditions, which may be due to the extremely unbalanced proportion of common items from different content domains.

Table 5.12. Correlation for the Recovery of **a**-parameters in Part II

| | | (1) 3D | (2) 2D, Correlation | (3) 2D, Proportion | (4) Classical |
|---|---|---|---|---|---|
| Dimension 1 | p0 | 0.9759 | 0.9758 | 0.9789 | 0.9732 |
| | p0.4 | 0.9719 | 0.9783 | 0.9815 | 0.9357 |
| | p0.6 | 0.9727 | 0.9851 | 0.9606 | 0.9323 |
| | p0.8 | 0.9547 | 0.9673 | 0.9462 | 0.9185 |
| Dimension 2 | p0 | 0.9856 | 0.9856 | 0.9865 | 0.9734 |
| | p0.4 | 0.9839 | 0.9803 | 0.9781 | 0.9699 |
| | p0.6 | 0.9847 | 0.9702 | 0.9846 | 0.9747 |
| | p0.8 | 0.9813 | 0.9735 | 0.9811 | 0.9676 |
| Dimension 3 | p0 | 0.9695 | 0.9211 | 0.9239 | 0.9222 |
| | p0.4 | 0.9669 | 0.9420 | 0.9386 | 0.8990 |
| | p0.6 | 0.9619 | 0.8873 | 0.8992 | 0.8909 |
| | p0.8 | 0.9442 | 0.8556 | 0.8766 | 0.8712 |

Table 5.13 shows the bias values for the recovery of **a**-parameters. It seems that, in the classical correlation method, the estimates on Dimension 1 are positively biased and those on other dimensions are negatively biased. Also, the estimates on Dimension 3 are negatively biased for all methods. Figures 5.2-5.4 plot the bias values for the recovery of **a**-parameters on each dimension. For all plots, the points representing Method 4 deviate far away from the horizontal zero-line. In addition, the points for Methods 1-3 are very close to each other and to the zero-line for the first two dimensions; however, for the third dimension, it is clear that only the points representing Method 1 are close to the zero-line.

Table 5.13. Bias for the Recovery of **a**-parameters in Part II

| | | (1) 3D | (2) 2D, Correlation | (3) 2D, Proportion | (4) Classical |
|---|---|---|---|---|---|
| Dimension 1 | p0 | 0.0005 | 0.0013 | 0.0030 | 0.0075 |
| | p0.4 | -0.0016 | -0.0010 | 0.0002 | 0.0086 |
| | p0.6 | -0.0019 | 0.0003 | 0.0024 | 0.0056 |
| | p0.8 | -0.0042 | -0.0013 | 0.0002 | 0.0019 |
| Dimension 2 | p0 | -0.0021 | -0.0019 | -0.0028 | -0.0091 |
| | p0.4 | -0.0011 | -0.0004 | -0.0014 | -0.0071 |
| | p0.6 | -0.0005 | 0.0001 | -0.0011 | -0.0047 |
| | p0.8 | -0.0005 | -0.0009 | -0.0017 | -0.0049 |
| Dimension 3 | p0 | -0.0035 | -0.0119 | -0.0111 | -0.0129 |
| | p0.4 | -0.0017 | -0.0046 | -0.0049 | -0.0110 |
| | p0.6 | -0.0019 | -0.0054 | -0.0067 | -0.0078 |
| | p0.8 | -0.0006 | -0.0035 | -0.0039 | -0.0035 |

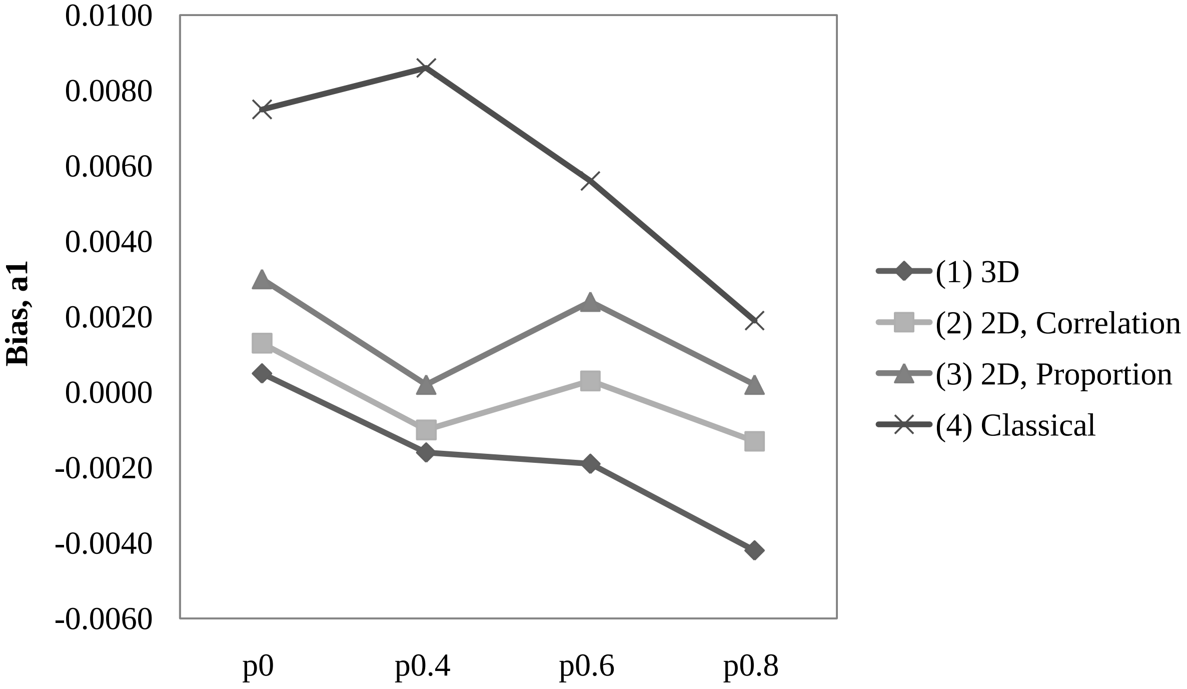

Figure 5.2. Bias for the Recovery of $a_1$-parameters in Part II

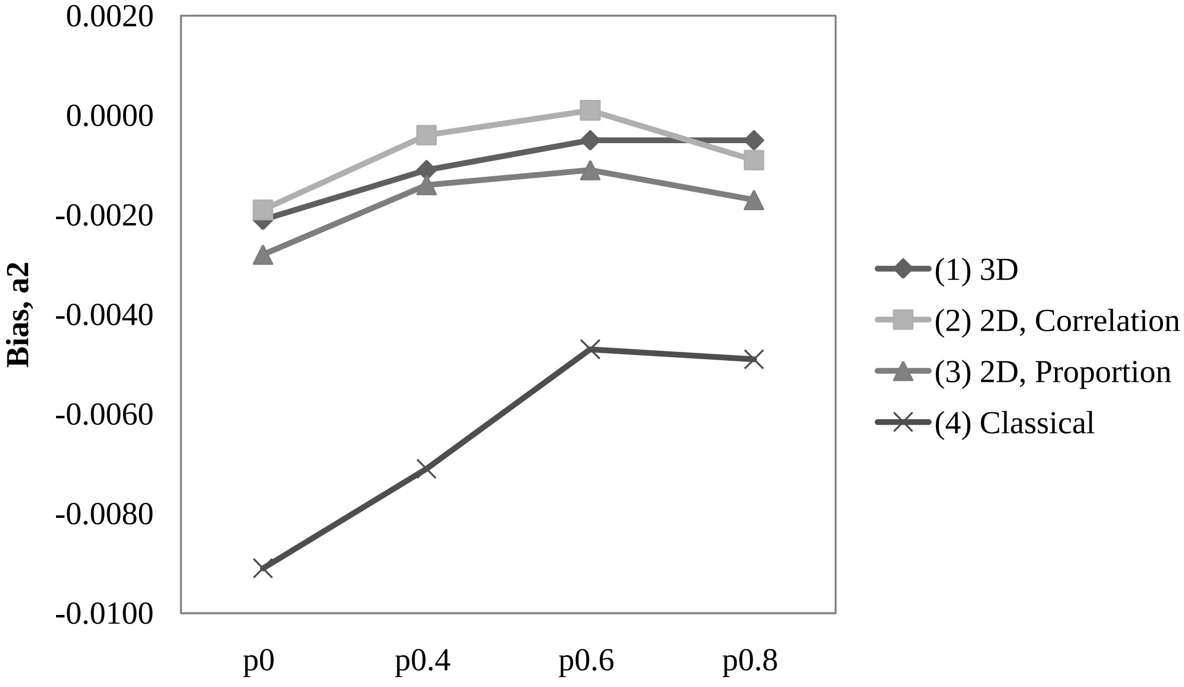

Figure 5.3. Bias for the Recovery of $a_2$-parameters in Part II

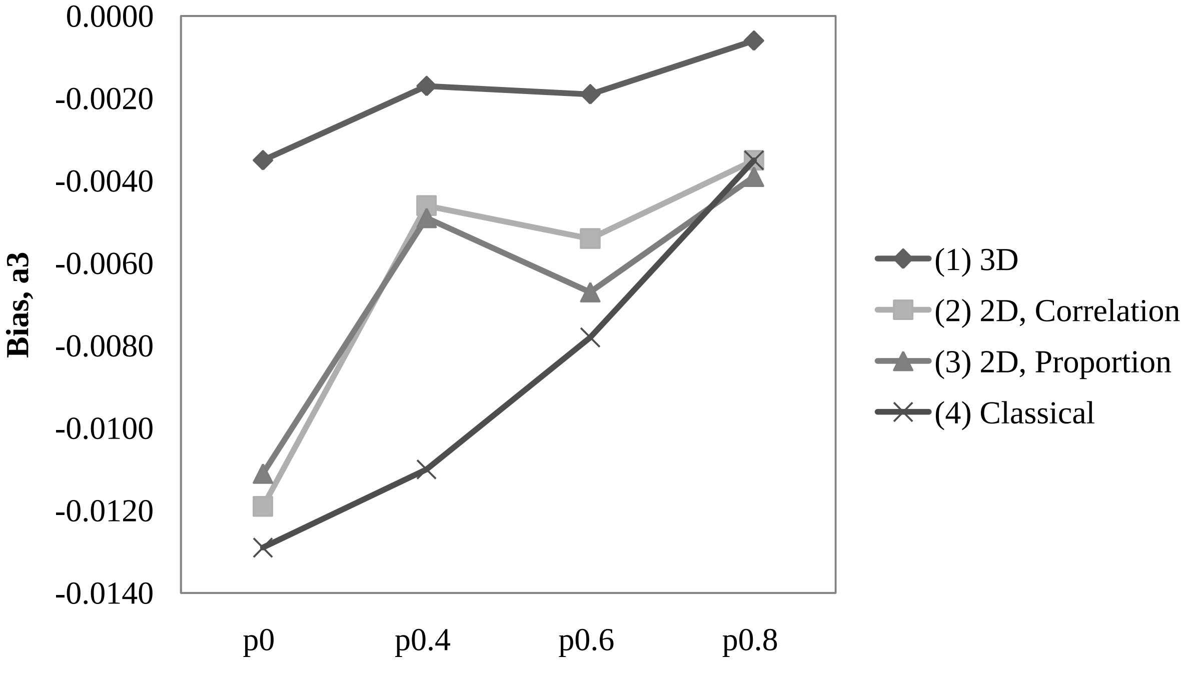

Figure 5.4. Bias for the Recovery of $a_3$-parameters in Part II

In order to evaluate the bias for the parameters of different values, the plot between the bias and **a**-parameters was examined on each dimension and for each method. In the classical method, $a_1$-parameters of small values tend to be slightly overestimated but there is no clear pattern for those of large values. For the remaining *a*-parameters, no clear pattern is found for the bias for those of small values, but *a*-parameters of large values tend to be underestimated. In addition, for $a_3$-parameters, the underestimation tends to become worse but the magnitude of bias is smaller in Method 1 than in other methods. Figure 5.5 shows one example of those plots, which gives the bias of $a_3$-parameters in Method 2 under the zero proficiency correlation condition. Note that the points in the right cluster represent the items dominantly measuring the third dimension.

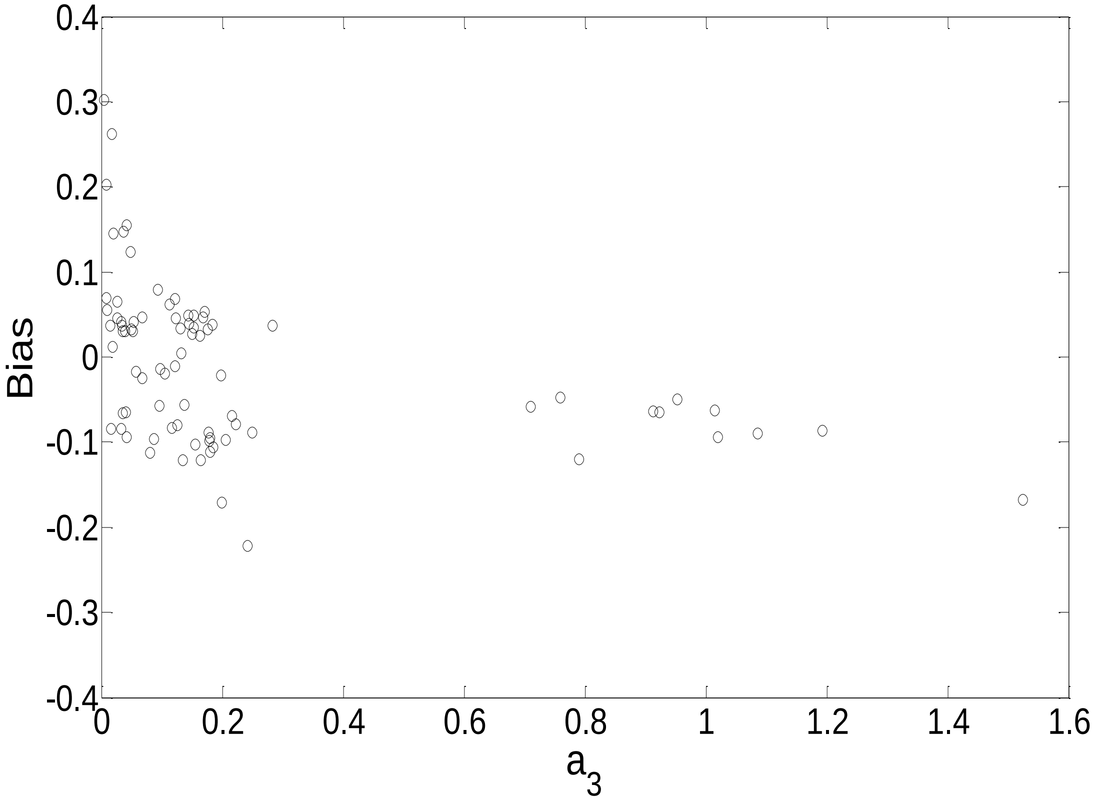

Figure 5.5. Bias for the Recovery of $a_3$-parameters for Method 2 at Zero Proficiency Correlation Level in Part II

Table 5.14 gives the RMSE values for the recovery of **a**-parameters. As the proficiency correlation increases, the estimates tend to deviate further from the true values for all dimensions and for all methods. Method 1, which provides full content coverage, performs a little better on Dimension 2 and gives much lower RMSE values on Dimension 3. Therefore, it confirms again that the *a*-parameter estimates on Dimension 3 are closer to the true values only by including items from that dimension in the common item set. The classical correlation method yields high RMSE values and does not perform as well as the other three methods, which is not very surprising in view of the extremely unbalanced number of common items for each content domain. These results, which are also shown in Figures 5.6-5.8, indicate that the recovery for the *a*-parameters on a certain dimension mostly depends on the number of common items measuring that dimension.

Table 5.14. RMSE for the Recovery of **a**-parameters in Part II

| | | (1) 3D | (2) 2D, Correlation | (3) 2D, Proportion | (4) Classical |
|---|---|---|---|---|---|
| Dimension 1 | p0 | 0.0925 | 0.0935 | 0.0882 | 0.0924 |
| | p0.4 | 0.0980 | 0.0875 | 0.0817 | 0.1386 |
| | p0.6 | 0.0967 | 0.0727 | 0.1140 | 0.1451 |
| | p0.8 | 0.1255 | 0.1073 | 0.1344 | 0.1613 |
| Dimension 2 | p0 | 0.0770 | 0.0769 | 0.0752 | 0.0998 |
| | p0.4 | 0.0798 | 0.0904 | 0.0932 | 0.1052 |
| | p0.6 | 0.0803 | 0.1141 | 0.0808 | 0.0980 |
| | p0.8 | 0.0883 | 0.1064 | 0.0888 | 0.1071 |
| Dimension 3 | p0 | 0.0731 | 0.1134 | 0.1101 | 0.1108 |
| | p0.4 | 0.0785 | 0.0980 | 0.1020 | 0.1279 |
| | p0.6 | 0.0842 | 0.1390 | 0.1323 | 0.1337 |
| | p0.8 | 0.1016 | 0.1573 | 0.1435 | 0.1445 |

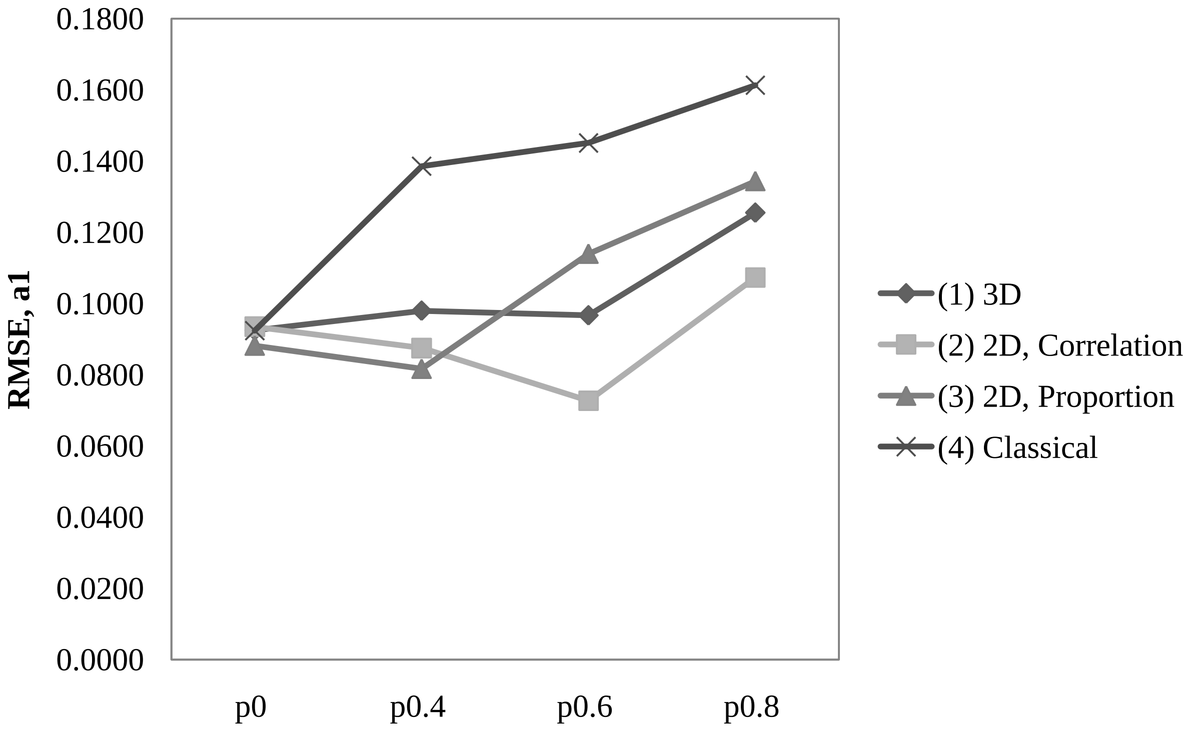

Figure 5.6. RMSE for the Recovery of $a_1$-parameters in Part II

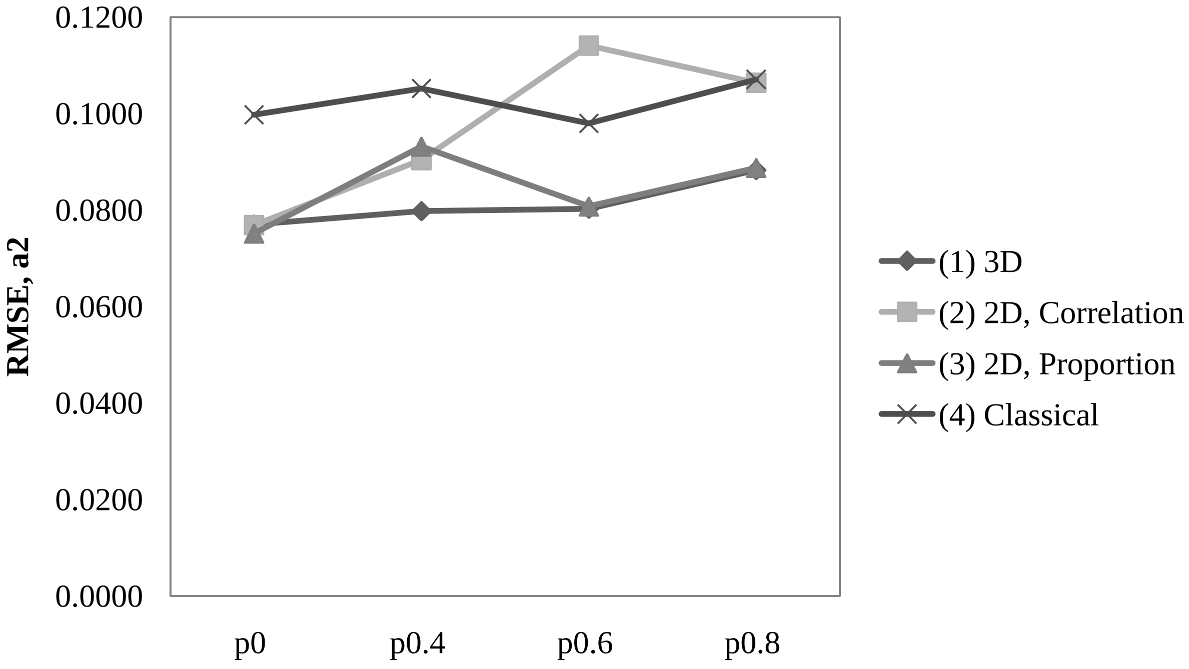

Figure 5.7. RMSE for the Recovery of $a_2$-parameters in Part II

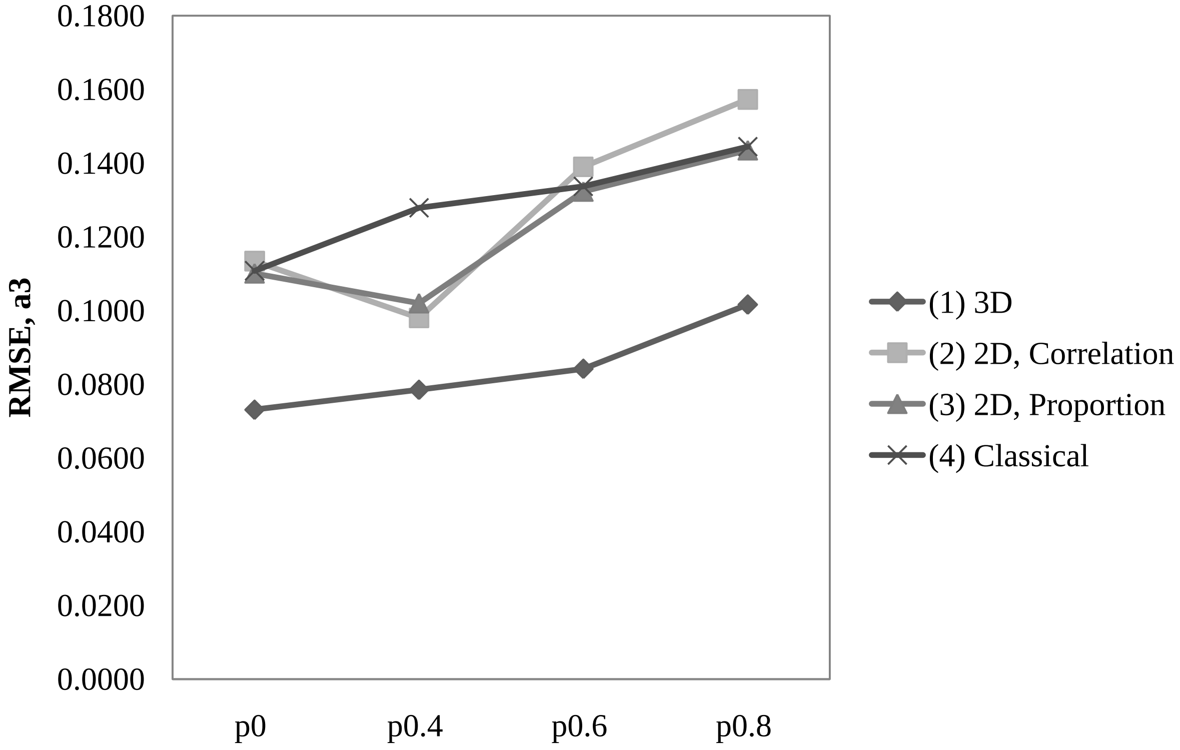

Figure 5.8. RMSE for the Recovery of $a_3$ -parameters in Part II

### 5.3.3 Recovery of *d*-parameters

Tables 5.15-5.17 show the correlation, bias and RMSE values between the adjusted estimates and true values of *d*-parameters for all item selection methods. There is no consistent pattern for the value change of these indices across different proficiency correlation levels. Generally speaking, Method 3 seems to give the highest correlation values for the recovery of *d*-parameters, while Method 2 provides the lowest values. Also, negative bias is found in the *d*-parameter estimates, although the values are quite small. From the RMSE table, Methods 3 and 4 seem to have a little advantage over the other two methods, especially when the proficiency correlation is low. Figure 5.9 gives the plot between the bias values and *d*-parameters for Method 1 at the zero proficiency correlation level, and the plots for all other methods and conditions are similar to this one. Different from that in Part I, there is no clear pattern for the bias at different values of *d*-parameters, but the magnitude of negative bias tends to be slightly larger than that of positive bias. Figures 5.10-5.12 provide the plots of correlation, bias and RMSE for the recovery of *d*-parameters.

Table 5.15. Correlation for the Recovery of *d*-parameters in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | 0.9873 | 0.9874 | 0.9895 | 0.9904 |
| (2) 2D, Correlation | 0.9846 | 0.9871 | 0.9875 | 0.9872 |
| (3) 2D, Proportion | 0.9892 | 0.9901 | 0.9902 | 0.9902 |
| (4) Classical | 0.9891 | 0.9893 | 0.9893 | 0.9885 |

Table 5.16. Bias for the Recovery of *d*-parameters in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | -0.0034 | -0.0015 | -0.0026 | -0.0023 |
| (2) 2D, Correlation | -0.0015 | -0.0022 | -0.0033 | -0.0037 |
| (3) 2D, Proportion | -0.0044 | -0.0029 | -0.0042 | -0.0043 |
| (4) Classical | -0.0007 | -0.0006 | -0.0021 | -0.0023 |

Table 5.17. RMSE for the Recovery of *d*-parameters in Part II

| | p0 | p0.4 | p0.6 | p0.8 |
|---|---|---|---|---|
| (1) 3D | 0.1032 | 0.1046 | 0.0949 | 0.0904 |
| (2) 2D, Correlation | 0.1148 | 0.0999 | 0.0984 | 0.1001 |
| (3) 2D, Proportion | 0.0900 | 0.0868 | 0.0847 | 0.0860 |
| (4) Classical | 0.0904 | 0.0894 | 0.0899 | 0.0945 |

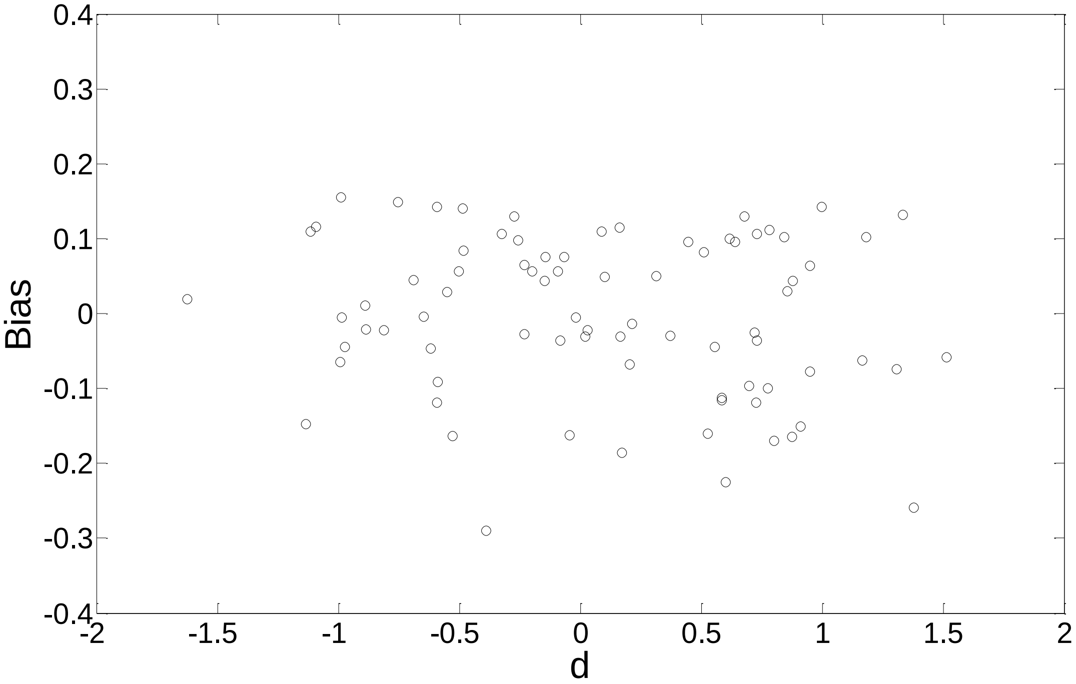

Figure 5.9. Bias of the Recovery of *d*-parameters for Method 1 at Zero Proficiency Correlation Level in Part II

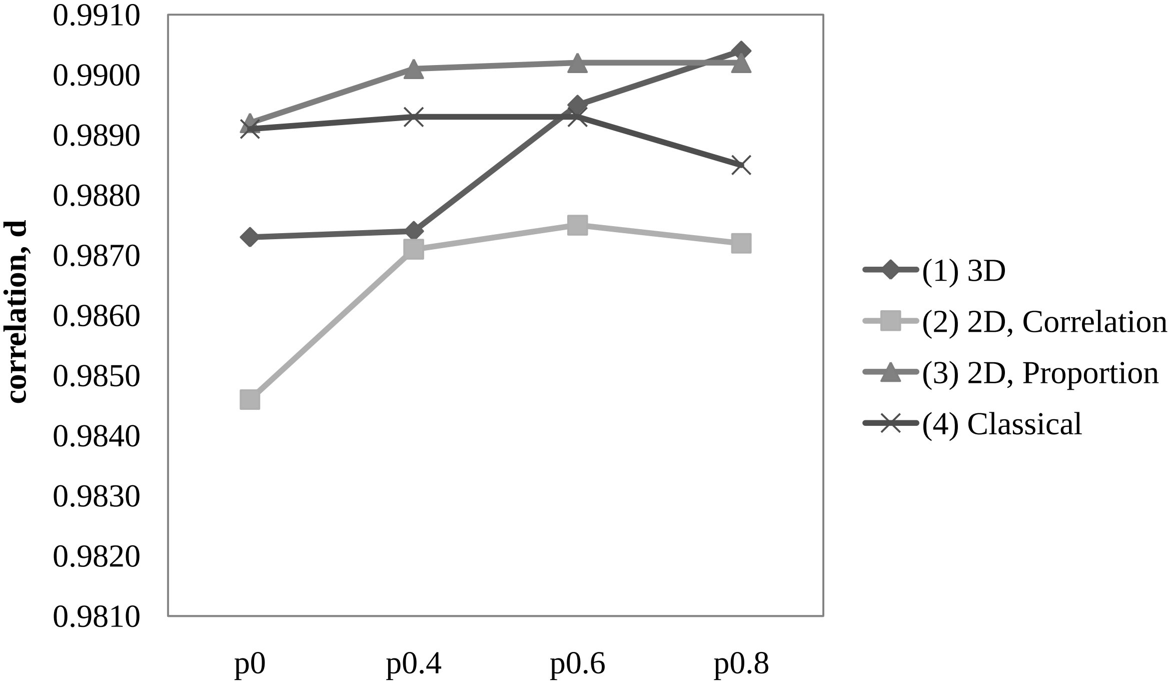

Figure 5.10. Correlation for the Recovery of *d*-parameters in Part II

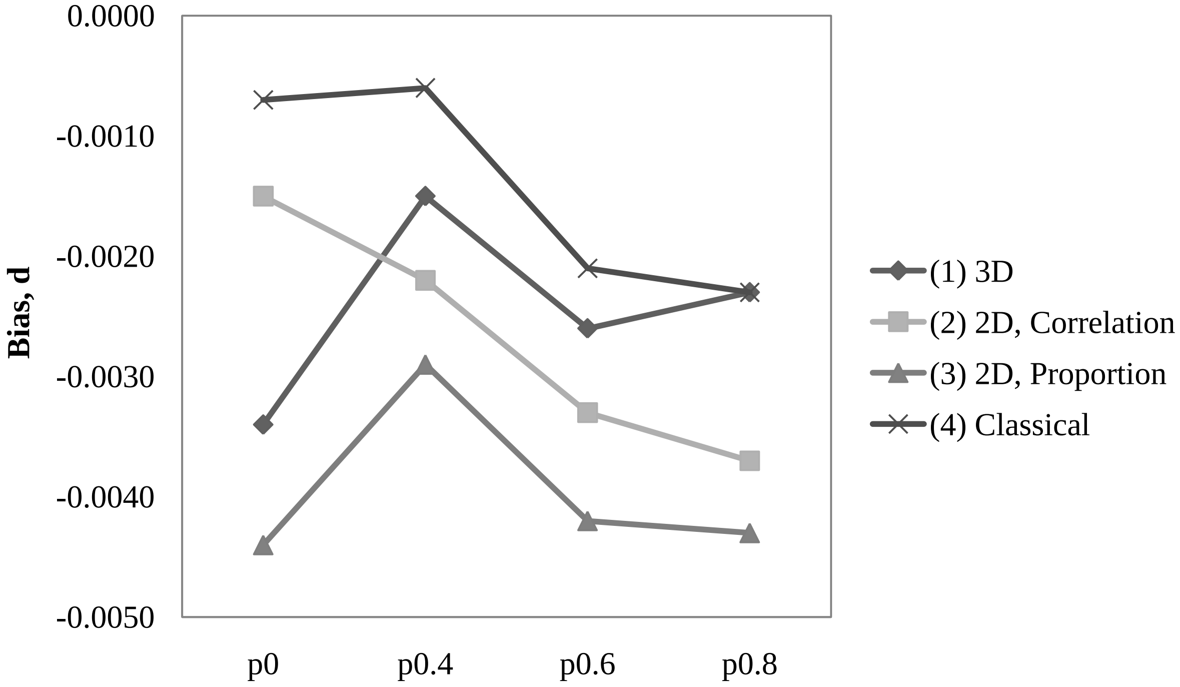

Figure 5.11. Bias for the Recovery of *d*-parameters in Part II

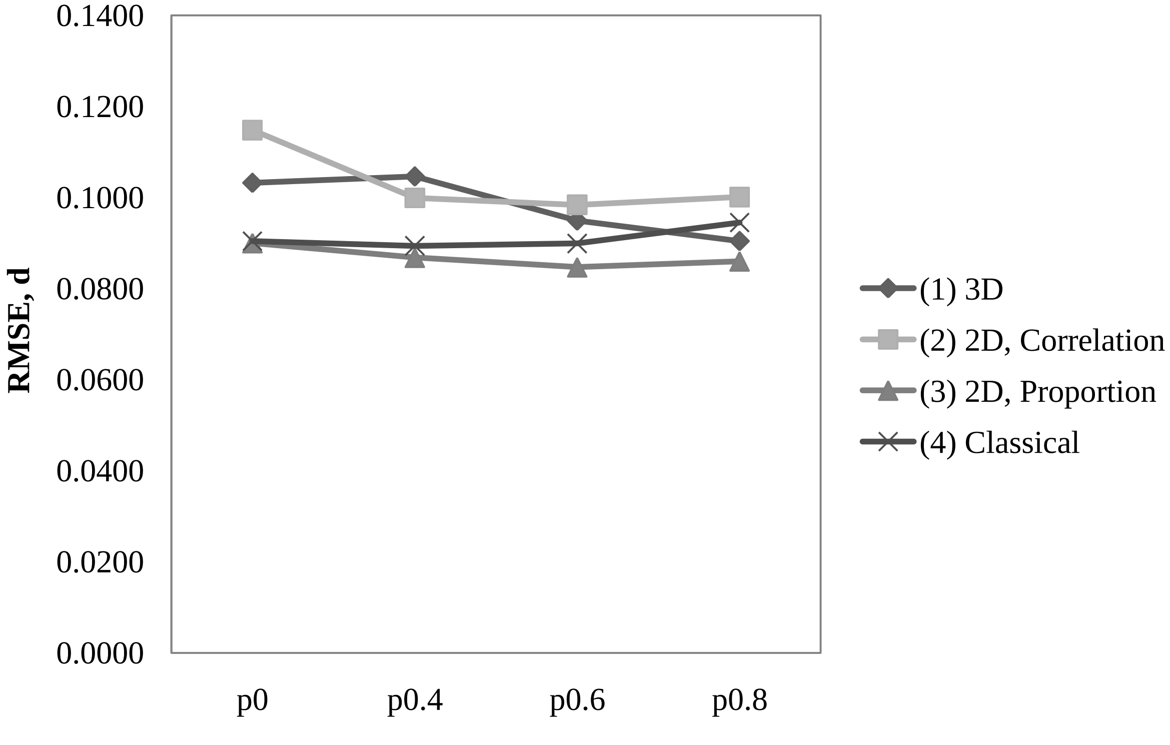

Figure 5.12. RMSE for the Recovery of *d*-parameters in Part II

### 5.3.4 Recovery of effect sizes

The variances of estimated proficiencies on different dimensions were underestimated and the correlations were overestimated, which is similar as in Part I. Table 5.18 shows the true effect sizes as well as the means and standard deviations of estimated effect sizes on the three dimensions for all methods and for all proficiency correlation levels. The first observation is that the estimated effect sizes from Method 4 deviate far away from the true values across all proficiency correlation levels and for all dimensions. The reason may be that almost all common items in this method are from the second content domain.

For Dimension 1, it seems that no method gives a consistent good recovery and the estimates given by Method 1 are closer to the true values only when the proficiency correlation is low. For Dimension 2, Method 3 gives the best recovery among all methods; however, for Dimension 3, Method 1 performs substantially better than the other methods although the estimates still deviate from the true values. Therefore, it can be concluded that without common items dominantly measuring a certain dimension, the effect size on that dimension is highly underestimated. However, as the proficiency correlation increases, the underestimation tends to be less severe, especially for Methods 2 and 3.

The values of standard deviations are quite small, which indicates that estimated effect sizes are fairly stable across replications. In order to show how substantial the difference is between different methods, Figures 5.13-5.15 provide the 95% confidence intervals of effect size for the comparison of Methods 1 and 4 across four proficiency correlation levels and for all three dimensions. From the plots, all the confidence intervals are quite narrow and there is no overlapping of the confidence intervals for the two methods. Figures 5.16-5.18 provide the plots for the recovery of effect sizes for the three dimensions, respectively.

Table 5.18. Recovery of Effect Sizes for Proficiencies in Part II

| | | TRUE | (1) 3D | (2) 2 D, Correlation | (3) 2 D, Proportion | (4) Classical |
|---|---|---|---|---|---|---|
| Dimension 1 | p0 | 0.7122 | 0.7057 | 0.8005 | 0.7974 | 0.6116 |
| | Std | | *0.0187* | *0.0147* | *0.0133* | *0.0130* |
| | p0.4 | 0.7283 | 0.7551 | 0.8028 | 0.7995 | 0.6218 |
| | Std | | *0.0167* | *0.0123* | *0.0135* | *0.0143* |
| | p0.6 | 0.6556 | 0.7309 | 0.7351 | 0.7310 | 0.5619 |
| | Std | | *0.0339* | *0.0117* | *0.0131* | *0.0146* |
| | p0.8 | 0.7147 | 0.8263 | 0.7893 | 0.7878 | 0.6085 |
| | Std | | *0.0327* | *0.0136* | *0.0140* | *0.0226* |
| Dimension 2 | p0 | 0.7198 | 0.6494 | 0.6531 | 0.7184 | 0.7812 |
| | Std | | *0.0190* | *0.0177* | *0.0156* | *0.0130* |
| | p0.4 | 0.7377 | 0.6193 | 0.6493 | 0.7140 | 0.7853 |
| | Std | | *0.0179* | *0.0149* | *0.0135* | *0.0115* |
| | p0.6 | 0.6550 | 0.5996 | 0.5895 | 0.6490 | 0.7177 |
| | Std | | *0.0308* | *0.0142* | *0.0118* | *0.0093* |
| | p0.8 | 0.7058 | 0.6547 | 0.6232 | 0.6853 | 0.7599 |
| | Std | | *0.0204* | *0.0162* | *0.0131* | *0.0089* |
| Dimension 3 | p0 | 1.8427 | 1.1923 | 0.3271 | 0.3956 | 0.4095 |
| | Std | | *0.0240* | *0.0217* | *0.0207* | *0.0157* |
| | p0.4 | 1.9016 | 1.2743 | 0.6102 | 0.6357 | 0.5087 |
| | Std | | *0.0175* | *0.0140* | *0.0171* | *0.0165* |
| | p0.6 | 1.8838 | 1.2346 | 0.6161 | 0.6325 | 0.4939 |
| | Std | | *0.0184* | *0.0156* | *0.0156* | *0.0379* |
| | p0.8 | 1.9483 | 1.2384 | 0.7135 | 0.7262 | 0.5355 |
| | Std | | *0.0182* | *0.0161* | *0.0165* | *0.0922* |

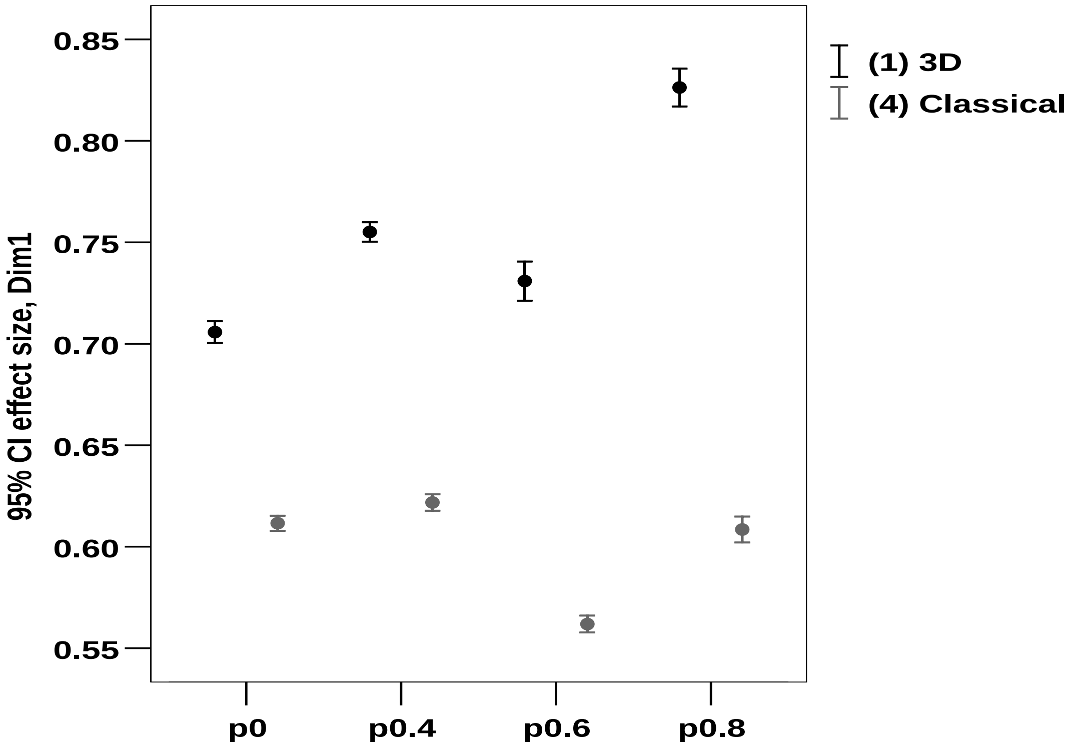

Figure 5.13. Comparison of Effect Size for the Proficiency on Dimension 1 in Part II

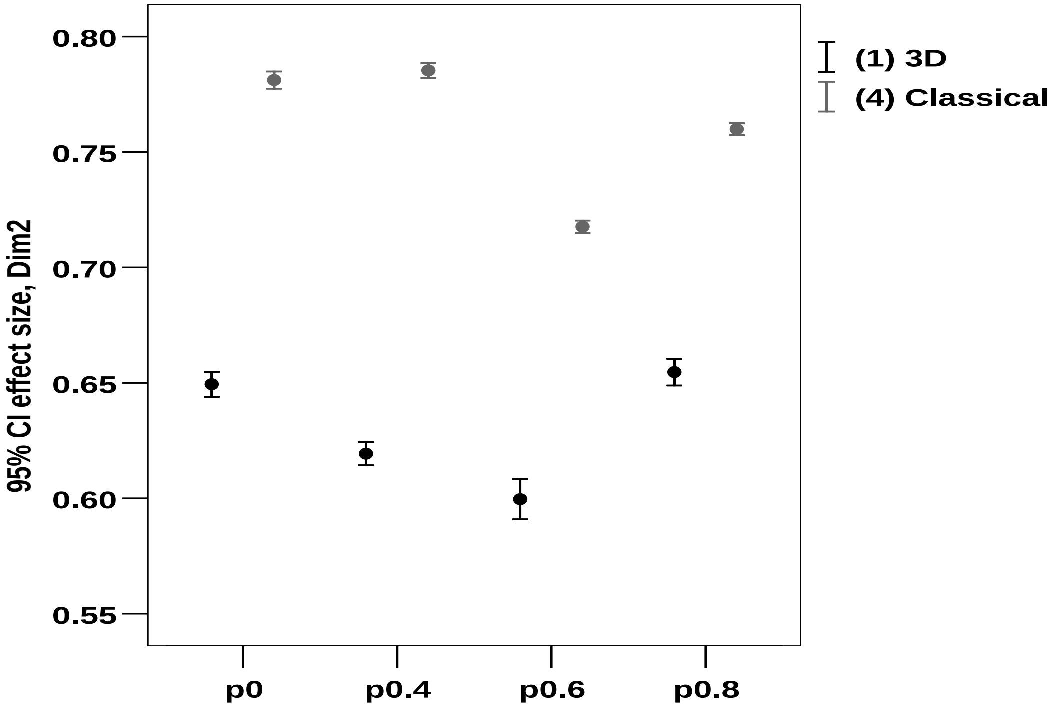

Figure 5.14. Comparison of Effect Size for the Proficiency on Dimension 2 in Part II

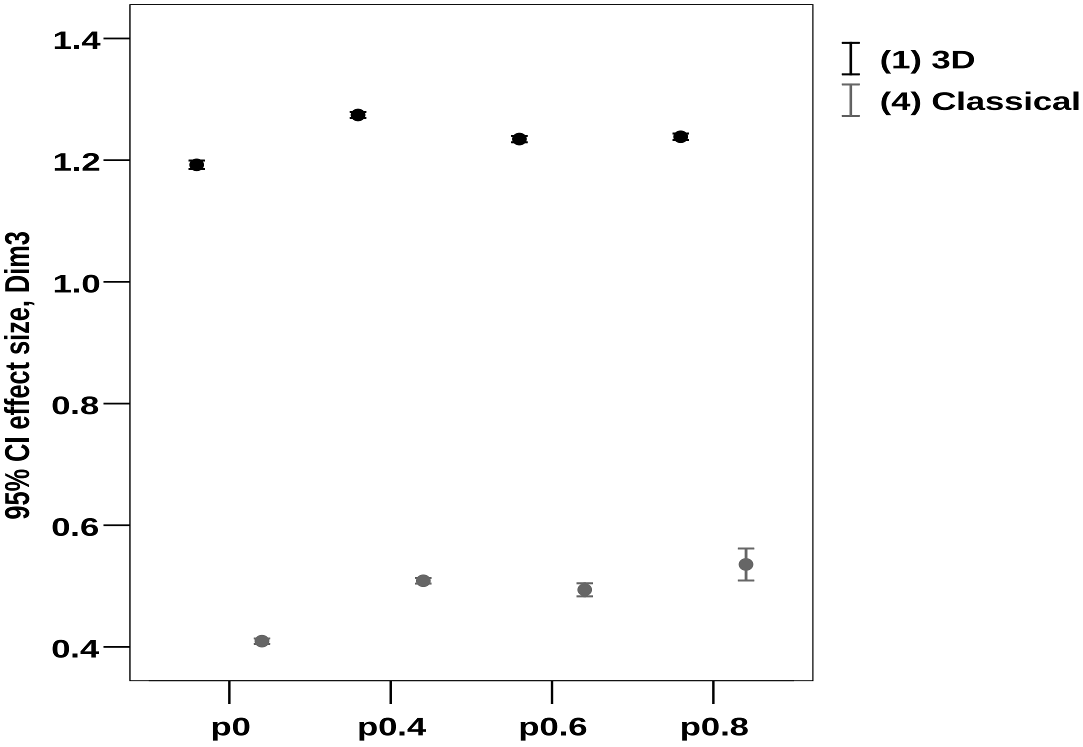

Figure 5.15. Comparison of Effect Size for the Proficiency on Dimension 3 in Part II

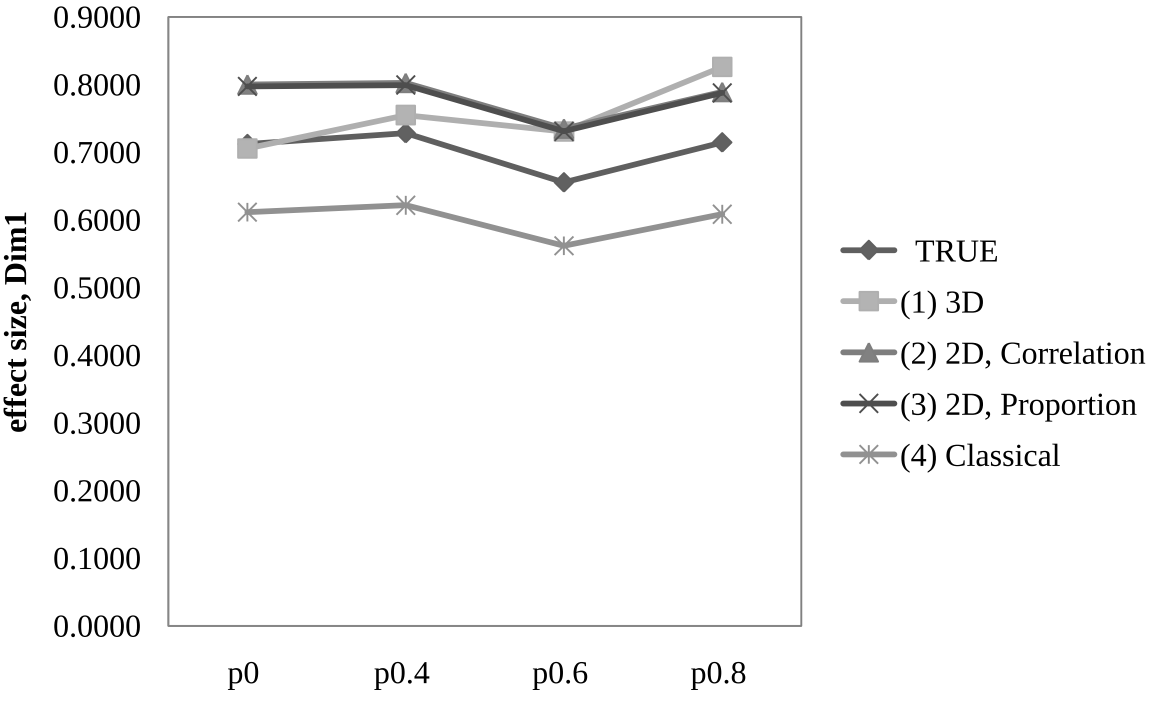

Figure 5.16. Recovery of Effect Size for the Proficiency on Dimension 1 in Part II

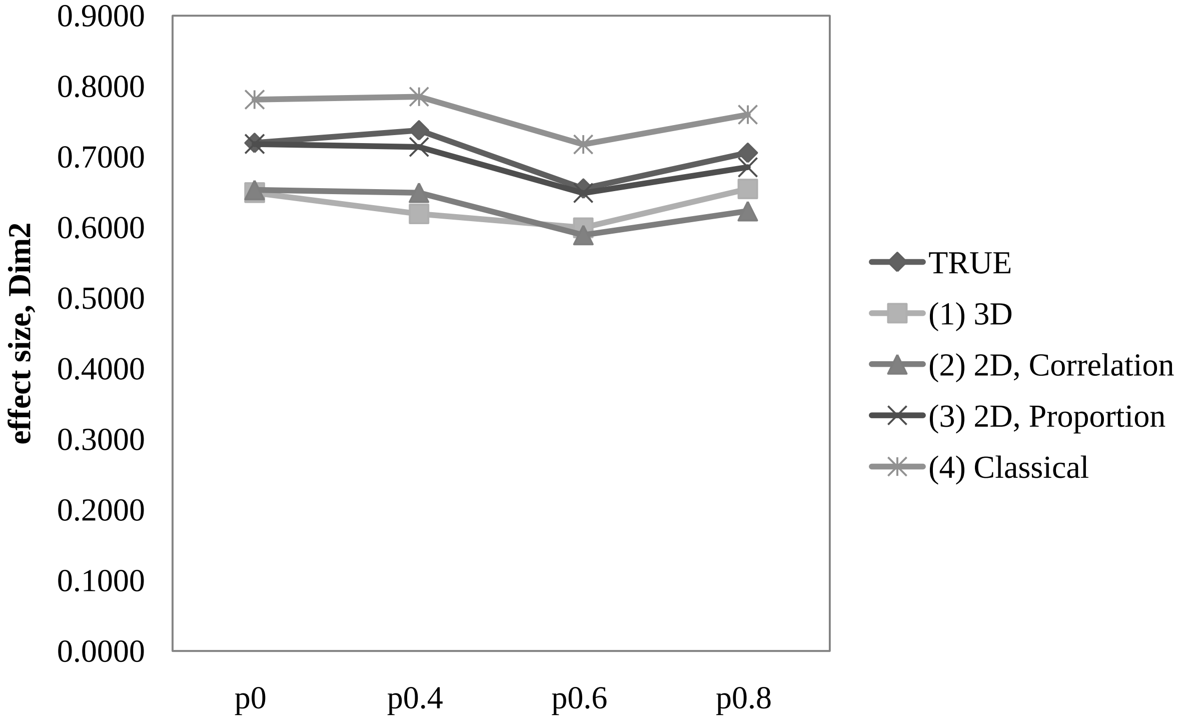

Figure 5.17. Recovery of Effect Size for the Proficiency on Dimension 2 in Part II

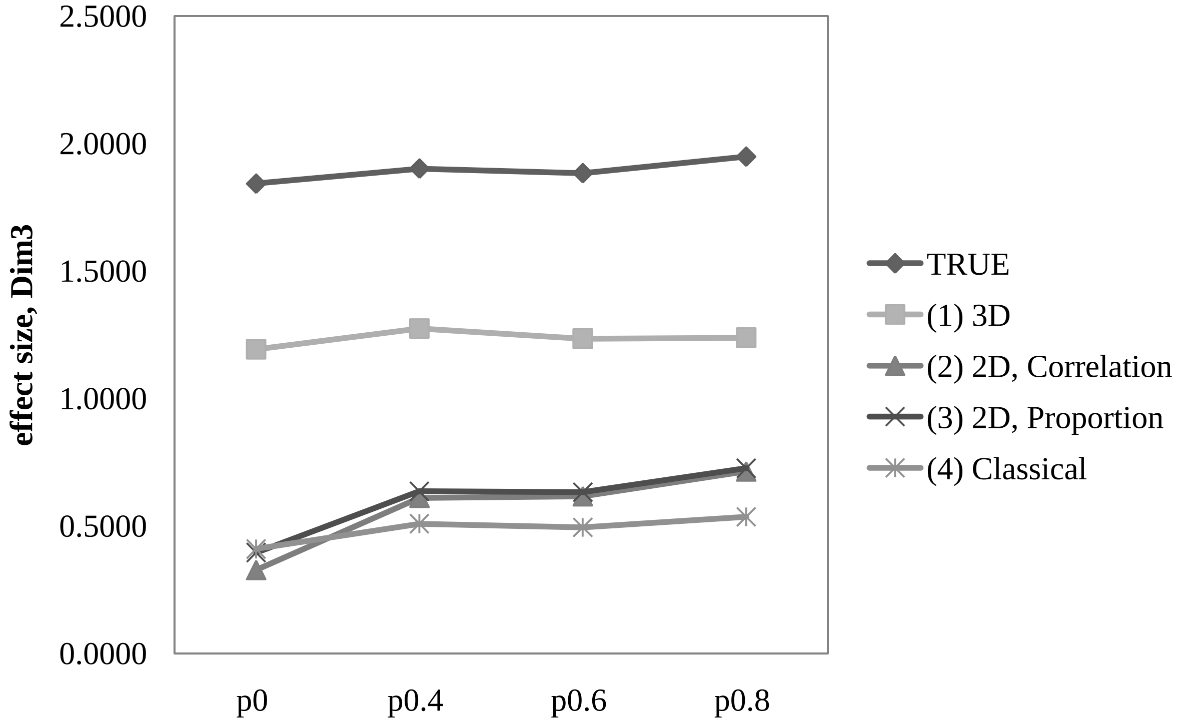

Figure 5.18. Recovery of Effect Size for the Proficiency on Dimension 3 in Part II

# CHAPTER 6

# SUMMARY, LIMITATION AND FUTURE RESEARCH

In this chapter, results and conclusions from the simulation studies are summarized and practical implications are discussed. In addition, limitations and suggestions for future research are provided.

## 6.1 Conclusions and Discussions

Part I focused on the two-dimensional constructs with balanced item design and data were simulated based on proficiencies on the same constructs in both upper and lower grades. In this part, anchor items were selected according to the combination of different content coverage (partial or full content coverage) and difficulty coverage (low difficulty level, medium difficulty level, high difficulty level, or all three difficulty levels). In addition, items with high item-total-test correlations were selected as common items under the classical method. Meanwhile, proficiency correlation level was manipulated to vary from low to high to evaluate its effect on the linking results for each item selection method. The comparisons among different methods were made with respect to the recovery of the probability matrix, item parameters and effect sizes.

The results show that with the increase of the correlation between proficiencies, the probability matrix recovery becomes better. In particular, the correlation between the estimated and true probabilities increases and the RMSE decreases. This was also observed in the study by Fang and Lu (2010). They pointed out that the RMSE for the probability matrix recovery decreases as the proficiency correlation increases when one unidimensional IRT or MIRT calibration is conducted on the data matrix simulated from a two-dimensional MIRT model.

Therefore, for the MIRT calibration on the complete data matrix, when the proficiency correlation is high, it can be assumed that the proficiency estimation on different dimensions can borrow information from each other. Thus, it is reasonable that a better estimation on proficiencies can yield a better recovery of the probability matrix.

However, this is not the case with the **a**-parameter recovery. As the proficiency correlation increases, the correlation between **a**-parameters and estimates decreases and their deviation increases. The reason may be that it becomes more difficult to separate the effect of the proficiency correlation from the **a**-parameter estimates in an attempt to solve the rotational indeterminacy in MIRT.

The recovery of *d*-parameters becomes better as the proficiency correlation increases. This may be due to the fact that the estimates become less affected by differences among dimensions when the data structure approaches unidimensionality.

Generally speaking, the recovery of effect sizes is not much influenced by the magnitude of the proficiency correlation, although it seems to be slightly better as the proficiency correlation increases. One exception is when all common items or most common items dominantly measure one dimension. In this case, the effect size recovery on other dimensions becomes substantially better as the proficiency correlation increases, which is reasonable since the proficiency parameters on different dimensions are more interrelated.

It is obvious that different common item selection methods do give different linking results as expected. Among all methods, three of them are of special interest: Method 6 (full content coverage with items from medium difficulty level), Method 8 (full content coverage with items from all difficulty levels) and Method 9 (items with high item-total-test correlation). The first one is originated from the idea of miditest proposed by Sinharay and Holland (2006b), the second one continues to be the golden rule and favorite of practitioners, and the third one traces back to the in-depth reason for the better equating although it is applied under the framework of multidimensional constructs in this study.

The results show that the classical correlation method gives the best probability recovery. Among all MIRT methods, with the same difficulty coverage, the method achieving full content coverage could give better results than the method achieving partial content coverage. Also, the method selecting medium difficulty items is the best among those selecting items from different difficulty levels under the same content coverage condition. Thus, it is not surprising that Method 6 performs much better than all other MIRT methods.

The **a**-parameter recovery varies in different selection methods. Methods 6 and 8 perform better in the linking as expected. Surprisingly, Method 9 also works better over most methods, although the numbers of common items from different content domains are unbalanced.

For the *d*-parameters, Methods 6 and 9 could give pretty good results for the recovery, followed by Method 8. The methods selecting common items from one content domain do not work as well as those selecting items from both content domains, especially when the proficiency correlation is low. But as the proficiency correlation increases, the difference in the *d*-parameter recovery for these two types of methods becomes smaller.

All effect sizes are underestimated, which may be due to the EAP scoring method for the proficiency estimation. The number of common items from each content domain plays an important role in the effect size recovery. For Methods 1 to 4 where common items come from the first content domain, the effect size on Dimension 2 is highly underestimated. Similar results could also be observed in the classical correlation method, which selects more items from the second content domain than the first one. This method could give a pretty good effect size recovery on Dimension 2 but not on Dimension 1. Therefore, besides the full content coverage, attention should also be paid to the proportion of common items in each content domain.

All in all, this part confirmed the advantage of miditest in the context of vertical scaling, which extended the conclusion by Sinharay and Holland (2006b) that the common item set with medium difficulty items could work better than the minitest in the equating. This is worthy of further attention by practitioners, although the minitest continues to be widely used in practical settings. Furthermore, the linking results also proved the importance of content coverage when multiple proficiencies are measured within one test. Therefore, the conclusion for Part I is that in vertical scaling under the MIRT framework, when the same constructs are measured in both tests, the common item set achieving full content coverage with medium difficulty items perform slightly better than the minitest that covers all content domains with a similar spread of item difficulties as the total test. And these two methods are substantially better than the other item selection methods using MIRT.

All common items selected via the classical correlation method are actually medium difficulty items, which is consistent with the idea that medium difficulty items tend to have higher item-total-test correlations than other items. This method gives good recovery for all parameters except for the effect size on Dimension 1. Since the effect size is much influenced by the proportion of common items in each content domain, given a fixed number of common items that can be used for linking, it is expected that the results would be better if this method selects appropriate number of items from each content domain to achieve proportional representativeness. Therefore, the classical correlation method seems surprisingly promising in the linking of multiple constructs. But there is concern that this method can only be used with careful design, since the correlation should be known in advance and the multidimensionality really limits the rationale and use of the item-total-test correlation.

Part II focused on the three-dimensional constructs with unbalanced item design. The purpose was to evaluate the common item selection methods when the measured constructs are not identical in both grades; in particular, the upper grade test measures more constructs than the lower grade test. In this part, “algebra” was not supposed to be taught in the lower grade; therefore, the proficiency distribution on that construct for lower grade examinees was assumed to have a low mean and small standard deviation.

With the conclusion from Part I that medium difficulty items perform comparatively better than other items in the linking, this part focused more on the content coverage and the proportion of common items from each content domain. As in Part I, comparisons among different item selection methods were made using the criteria on the recovery of probability matrix, item parameters and effect sizes.

The conclusion on the effect of proficiency correlation for the recovery of different parameters is consistent with that in Part I. As the proficiency correlation increases, the recovery of probability matrix and *d*-parameters becomes better, the recovery of **a**-parameters becomes worse, and the effect size recovery does not change too much except on Dimension 3 when no common items are selected to link that dimension.

The performance of different methods varied if different criteria were used for comparison. The content coverage is not as important as expected for the probability matrix recovery; however, it is very crucial for the recovery of **a**-parameters and effect sizes, which are also influenced by the proportions of common items selected from different content domains.

Nevertheless, the disadvantage of not covering all content domains could be partly compensated for by the high correlations between proficiencies.

Method 3, which is to select common items according to proportions of unique items in the two common content domains, seems to yield better results in most recoveries, except for those on Dimension 3. On the other hand, Method 1, which is to select items for full content coverage, is a better choice if item **a**-parameters and effect sizes are expected to be reasonably estimated for all dimensions. The classical correlation method does not work well in this part. This may be due to the unbalanced item design and the complicated content structure, which lead to the extremely unbalanced numbers of common items from different content domains. However, this classical correlation method is worth further analysis if it can be adjusted to achieve the aforementioned proportional representativeness.

The results of this part reconfirm the importance of content coverage, even when the content domain only exists in one grade. Items measuring other highly-correlated proficiencies cannot replace the items from that domain in the common item set. However, it should be noted that including items from all content domains does not ensure a better recovery of probability matrix.

As is well known, vertical scaling with the common item design is currently implemented in many state testing programs, such as California English Language Development Test (CELDT), Colorado Student Assessment Program (CSAP), Connecticut Mastery Test (CMT), Delaware Student Testing Program (DSTP), Mississippi Curriculum Test (MCT), North Carolina End-of-Grade Tests (NCEOG) and Texas English Language Proficiency Assessment System (TELPAS) (Reckase, 2010). Since the scale scores are not comparable across different state testing programs, the common core standards and common assessments are of special interest to practitioners, policy makers and researchers. With the Race to the Top program to motivate reforms in state and local district K-12 education, two consortia, including Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER Balanced Assessment Consortium (SBAC), also show great interest in vertical scales for assessing students' achievement and growth.

However, it is the unidimensional IRT models, either the Rasch model or the 3 PL model, that are commonly adopted in state testing programs. Theoretically, these two unidimensional IRT models can be arguably used and serve as a good approximation to the multidimensional IRT models only when all items in the test measure roughly the same composite of multiple proficiencies (Reckase, Ackerman & Carlson, 1988), or all the proficiencies are highly correlated in consideration of a correlation of 0.7 or more as commonly seen in practice. The high correlation among multiple proficiencies not only explains the reason that mathematical skills can be finely divided into geometry, algebra and etc. or they can be attributed to one general mathematical ability, but also raises heated debates on when to report subscores or one general score from the perspective of psychometrics.

Due to the complex constructs measured in tests and the changes in the curriculum and policy requirements with the change of grade levels, if the constructs measured by different tests in different grade levels are not identical, the interpretation of one single scale score obtained from those tests may not be the same. Therefore, the MIRT model seems to be more appropriate under the situations of multidimensionality and content shift, but there are four main concerns when the MIRT model is practically used in vertical scaling.

First, even with the expert judgment and the dimensionality analysis, it is still difficult to define the constructs measured by different tests across grades. This is even worse in view of the high correlation among proficiencies.

Second, from the results of this study, in order to link the scales for all content areas, off-grade items need to be administered to students even if the content area is not covered by that grade. The importance of the off-grade content in tests is confirmed by Lazer, Mazzeo, Twing, Way, Camara, & Sweeney (2010), who assumed that “this out-of-grade content will mirror the instruction the student has received regardless of his or her grade level or age”. Although students may be reluctant to answer items that are never taught in class, it is difficult to evaluate students’ gains after learning activities if we have no ideas about their pre-knowledge on that content. It is a trade off; unfortunately, the NCLB has prohibited this off-grade testing.

Third, the rotational indeterminacy is a big problem in MIRT vertical scaling. The Varimax and Promax methods are commonly used to constrain items to follow the simple structure by assuming that each item only dominantly measures one proficiency. However, this indeterminacy problem becomes complicated with the existence of mixed structure items and the different correlation structures of proficiencies for students in different grades; therefore, more research studies are needed to better construct the coordinate system for the interpretation of parameter estimates in the MIRT vertical scaling.

Finally, the MIRT calibration is extremely computationally extensive no matter whether the EM algorithm or Markov Chain Monte Carlo (MCMC) algorithm is used, and the computation time increases rapidly with the increase of the number of dimensions for the MIRT model. This may not be acceptable for most testing programs since test results need to be delivered within a short period of time. But as the computer becomes more and more powerful, this may not be a problem.

All in all, the MIRT model is more appropriate to account for the multidimensionality in vertical scales; however, there is still a long way to go to implement this model in practice.

## 6.2 Limitations and Future Research

The above conclusions should be interpreted in light of the limitations inherent in this simulation study. Also, future research needed to make the conclusions more solid and generalizable is discussed below.

First, items in this study were simulated to be approximate simple structure items, which results in a rough alignment of proficiency dimension and content domain. Therefore, sometimes these two terms are used interchangeably in this study. But in practice, multiple proficiencies are often needed to get an item correct (Reckase, 1985). Future studies could be conducted to examine whether mixed structure items can be used in lieu of simple structure items that measure different proficiencies to achieve the full content coverage.

Second, in order to compare the recovery results from different selection methods, item parameter estimates need to be rotated and adjusted to be put on the same coordinate system as the generating parameters. In this study, the target matrix of rotation was defined as the generating parameters, while the study by Reckase and Li (2007) adopted a target matrix with 1s as indicators of measured dimensions and 0s elsewhere. Some trials of matching 0/1 matrices were conducted as well and the results were compared with those from matching with generating parameters. It was found that the difference was quite subtle.

Third, since probability values range from 0 to 1, the nonparametric Spearman’s rank order correlation coefficient may be a better choice than the Pearson's correlation coefficient that requires the assumption of normal distribution for variables. The probability recovery using this rank order correlation coefficient was compared with that using Pearson’s correlation coefficient and the results were quite similar.

Fourth, although multi-group analysis may be deemed as more appropriate for the concurrent calibration in this study, this option is not available in the TESTFACT software. For the multi-group calibration under the unidimensional IRT, most software packages set the default constraints only on the distribution of reference group, and treat the means and variances of distributions for other groups as unknown parameters. However, it can be imagined that in MIRT, if the elements in variance-covariance matrices are regarded as unknown parameters for other groups, the MIRT calibration would become much more complicated and time-consuming. Future research can verify whether the above conclusions still hold when the efficient software package is available for the MIRT multi-group calibration.

Fifth, concurrent calibration is used to align scales across different grade levels in this study. However, vertical scales can also be created by applying an orthogonal or oblique Procrustes rotation method to match the common item parameters estimated from separate calibrations on the two tests. This can be another topic for further studies and some thought should be given to the dimensionality issue when the constructs measured by the two tests are not identical.

Sixth, the performance of the classical correlation method is surprisingly good in Part I. However, the feasibility of this method is a little questionable since the item-total-test correlation values, which depend on the population of examinees, could not be known in advance. Although the results from field tests might be used for reference, they should be used with caution since the sample for the field test may not be representative of the population. Also, the performance of this method becomes worse when more distinctive proficiencies are measured in the test, which makes the unidimensionality assumption more vulnerable.

Finally, because the MIRT calibration in the TESTFACT software is time-consuming, the number of replications in this study is somewhat small compared with other studies. In the future, more replications could be conducted when the computation time shortens. In addition, in order to make the conclusions more generalizable, further studies can focus on the data with more dimensions, since the number of constructs measured by the test is often more than two in practical settings.

# APPENDIX

# APPENDIX

```
% code for the evaluation of Part II results in MATLAB

function evaluation_final
 load item2.dat;
A=item2(:,1:3);
d=item2(:,4);

ns=3000;
np=4;
ndim=3;
ni=40;
nj=4;
nr=50;

result.d=zeros([2*ni,np,nj,nr]);
result.A=zeros([2*ni,ndim,np,nj,nr]);
result.theta=zeros([2*ns,ndim,np,nj,nr]);

result.PBias=zeros([np,nj]);
result.PRmse=zeros([np,nj]);
result.PCorr=zeros([np,nj]);
result.ABias=zeros([ndim,np,nj]);
result.ARmse=zeros([ndim,np,nj]);
result.ACorr=zeros([ndim,np,nj]);
result.dBias=zeros([np,nj]);
result.dRmse=zeros([np,nj]);
result.dCorr=zeros([np,nj]);
result.ES=zeros([ndim,np,nj]);

ES=zeros([ndim,np]);
ESTemp=zeros(ndim,nr);
thetaT=zeros([ns,ndim]);
thetaTemp=zeros([2*ns,ndim]);
numT=zeros([1,ns]);
PBiasTemp=zeros([2*ns,ni,nr]);
PRmseTemp=zeros([2*ns,ni,nr]);
PCorrTemp=zeros([1,nr]);
ABiasTemp=zeros([2*ni,ndim,nr]);
ARmseTemp=zeros([2*ni,ndim,nr]);
ACorrTemp=zeros([ndim,nr]);
dBiasTemp=zeros([2*ni,nr]);
dRmseTemp=zeros([2*ni,nr]);
dCorrTemp=zeros([1,nr]);
```

```
    % person correlation i, method j, replication r
    for i=1:np
        theta=load (['person2', num2str(i),'.dat']);
        P=pfunction(theta,A,d);
        ES(:,i)=ESfunc(theta(3001:end,:), theta(1:3000,:));

        for j=1:nj
            for r=1:nr
                fname=['d2_p',num2str(i),'_r',num2str(r),'_m',num2str(j)];
                fid=fopen([fname,'.PAR'], 'r');
                C=textscan(fid,'%*d %*s %*d %9.6f %9.6f %9.6f %9.6f','headerlines',11);
                fclose(fid);
                result.d(:,i,j,r)=C{1};
                result.A(:,:,i,j,r)=[C{2:4}];

                fid=fopen([fname,'_s1.FSC'],'r');
                    for t=1:ns
                    temp=str2num(fgetl(fid));
                    numT(t)=temp(3);
                    thetaT(t,:)=str2num(fgetl(fid));
                    fgetl(fid);
                    end
                result.theta(1:3000,:,i,j,r)=thetaT;
                fclose(fid);

                fid=fopen([fname,'_s2.FSC'],'r');
                    for t=1:3*ns
                        fgetl(fid);
                    end
                    for t=1:ns
                    temp=str2num(fgetl(fid));
                    numT(t)=temp(3);
                    thetaT(t,:)=str2num(fgetl(fid));
                    fgetl(fid);
                    end
                result.theta(3001:end,:,i,j,r)=thetaT;
                fclose(fid);

    % correction for item discrimination by forcing the mean of item discrimination estimates to
be positive
            if mean(result.A(:,1,i,j,r))<0 result.A(:,1,i,j,r)=(-1)*result.A(:,1,i,j,r); end
            if mean(result.A(:,2,i,j,r))<0 result.A(:,2,i,j,r)=(-1)*result.A(:,2,i,j,r); end
            if mean(result.A(:,3,i,j,r))<0 result.A(:,3,i,j,r)=(-1)*result.A(:,3,i,j,r); end

    % correction for person proficiency estimates by choosing the pair which gives the highest
correlation for the recovery of probability matrix
```

```
       PTemp1=pfunction(result.theta(:,:,i,j,r), result.A(:,:,i,j,r),result.d(:,i,j,r));
       PTemp2=pfunction([result.theta(:,1:2,i,j,r),[(-1)*result.theta(1:3000,3,i,j,r);
       result.theta(3001:6000,3,i,j,r)]], result.A(:,:,i,j,r),result.d(:,i,j,r));
       PTemp3=pfunction([result.theta(:,1:2,i,j,r),[result.theta(1:3000,3,i,j,r);(-1)*
       result.theta(3001:6000,3,i,j,r)]], result.A(:,:,i,j,r),result.d(:,i,j,r));
       PTemp4=pfunction([result.theta(:,1:2,i,j,r),(-1)*result.theta(:,3,i,j,r)],
       result.A(:,:,i,j,r),result.d(:,i,j,r));
       PCorrTemp1=corr(reshape(PTemp1,[],1),reshape(P,[],1));
       PCorrTemp2=corr(reshape(PTemp2,[],1),reshape(P,[],1));
       PCorrTemp3=corr(reshape(PTemp3,[],1),reshape(P,[],1));
       PCorrTemp4=corr(reshape(PTemp4,[],1),reshape(P,[],1));

      PTemp=PTemp1;
      [PCorrTemp_value,ind]=max([PCorrTemp1,PCorrTemp2,PCorrTemp3,
      PCorrTemp4]);
      if ind==2 result.theta(1:3000,3,i,j,r)=(-1)*result.theta(1:3000,3,i,j,r);
      PTemp=PTemp2;
      elseif ind==3 result.theta(3001:6000,3,i,j,r)=(-1)*result.theta(3001:6000,3,i,j,r);
      PTemp=PTemp3;
      elseif ind==4 result.theta(:,3,i,j,r)=(-1)*result.theta(:,3,i,j,r);PTemp=PTemp4;
      end

% oblique Procrustes rotation to match with generating parameters
      T=inv(result.A(:,:,i,j,r)'*result.A(:,:,i,j,r))*result.A(:,:,i,j,r)'*A;
      ATemp=result.A(:,:,i,j,r)*T;
      m=((d-result.d(:,i,j,r))'*ATemp*inv(ATemp'*ATemp))';
      dTemp=result.d(:,i,j,r)+ATemp*m;
      thetaTemp=(inv(T)*result.theta(:,:,i,j,r)'-m*ones(1,size(result.theta(:,:,i,j,r),1)))';

      [PBiasTemp(:,:,r),PRmseTemp(:,:,r)]=criteria(PTemp, P);
      [ABiasTemp(:,:,r),ARmseTemp(:,:,r)]=criteria(ATemp,A);
      [dBiasTemp(:,r),dRmseTemp(:,r)]=criteria(dTemp, d);

      PCorrTemp(r)=PCorrTemp_value;
      ACorrTemp(:,r)=[corr(ATemp(:,1),A(:,1)),corr(ATemp(:,2),A(:,2)),
      corr(ATemp(:,3),A(:,3))];
      dCorrTemp(r)=corr(dTemp, d);

      ESTemp(:,r)=ESfunc(thetaTemp(3001:end,:),thetaTemp(1:3000,:));
      end

   result.PBias(i,j)=mean(mean(mean(PBiasTemp,3),1),2);
   result.PRmse(i,j)=mean(mean(sqrt(mean(PRmseTemp,3)),1),2);
   result.PCorr(i,j)=mean(PCorrTemp);

   result.ABias(:,i,j)=mean(mean(ABiasTemp,3),1);
```

```
    result.ARmse(:,i,j)=mean(sqrt(mean(ARmseTemp,3)),1);
    result.ACorr(:,i,j)=mean(ACorrTemp,2);

    result.dBias(i,j)=mean(mean(dBiasTemp,2),1);
    result.dRmse(i,j)=mean(sqrt(mean(dRmseTemp,2)),1);
    result.dCorr(i,j)=mean(dCorrTemp);

    result.ES(:,i,j)=mean(ESTemp,2);
    end
end

result.PCorr
result.PBias
result.PRmse
result.ACorr
result.ABias
result.ARmse
result.dCorr
result.dBias
result.dRmse
ES
result.ES

% function for probability calculation under 2PL compensatory MIRT
function P=pfunction(theta, A, d)
ns=0.5*size(theta,1);
ni=0.5*size(A,1);
P1=1./(1+exp(-1.7*(theta(1:ns,:)*A(1:ni,:)'+ones(ns,1)*d(1:ni)')));
P2=1./(1+exp(-1.7*(theta(ns+1:end,:)*A(ni+1:end,:)'+ones(ns,1)*d(ni+1:end)')));
P=[P1;P2];

% function related to bias and RMSE calculation
function [diff,diff2]=criteria(estimate, true)
diff=estimate-true;
diff2=(estimate-true).^2;

% function for effect size calculation
function ESvalue=ESfunc(theta1, theta2)
ESvalue=(mean(theta1)-mean(theta2))./sqrt((var(theta1)+var(theta2))*.5);
```

# REFERENCES

# REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, New Jersey: Educational Testing Service.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test* scores (pp. 397-479). Reading, MA: Addison-Wesley.

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

Bock, D., Gibbons, R., Schilling, S., Muraki, E., Wilson, D., & Wood, R. (2003). *TESTFACT 4.0 [Computer software and manual]: Test scoring, item statistics, and item factor analysis*. Lincolnwood, IL: Scientific Software International.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444.

Braun, H. I. (2005). *Using student progress to evaluate teaching: A primer on value-added models* (Tech. Rep.). Princeton, New Jersey: Educational Testing Service.

Camilli, G., Wang, M.-M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. *Journal of Educational Measurement, 32*(1), 79-96.

Doran, H. C., & Cohen, J. (2005). The confounding effect of linking bias on gains estimated from value-added models. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, *22*(4), 249-262.

Fang, Y. (2008). *Using a projection method to estimate subscores from tests with multidimensional structures.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Fang, Y., & Lu, Y. (2010). *The effect of proficiency correlation on the application of multidimensional IRT model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models in latent trait theory*. The University of New England, Center for Behavioral Studies, Armidale, Australia.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

Haertel, E. H. (2004). *The behavior of linking items in test equating* (CSE Technical Report 630). Los Angeles, CA: CRESST/CSE, University of California, Los Angeles, Graduate School of Education and Information Studies.

Hanson, B. A., & Béguin, A. 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.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York: Springer.

Hirsch, T. M. (1989). Mutidimensional equating. *Journal of Educational Measurement*, *26*(4), 337–349.

Holland P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-29). New York: Springer.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *ITBS forms A & B guide to research and development.* Itasca, IL: Riverside.

Hoskens, M., Lewis, D. M., & Patz, R. J. (2003). *Maintaining vertical scales using a common item design.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Jiao, H., & Wang, S. (2006). *Comparison of vertical linking designs*. Paper presented at the National Conference of the Large-Scale Assessment, San Francisco, CA.

Jiao, H., & Wang, S. (2007). *The effects of the selection of vertical linking items on modeling student growth*. Paper presented at the National Conference of the Large-Scale Assessment, Chicago, IL.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131-143.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger Publishers, jointly.

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

Lazer, S., Mazzeo, J., Twing, J. S.,Way, W. D., Camara, W., & Sweeney, K. (2010). Thoughts on an assessment of common core standards. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/6063DE04-2372-4EC4-9642-7B8A584F942F/0/ThoughtonaCommonCoreAssessmentSystem.pdf.

Li, T. (2006). *The effect of dimensionality on vertical scaling.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*(2), 115–138.

Lohnman, D. F., & Hagen, E. (2002). *Cognitive abilities test (Form 6): Research handbook.* Itasca, IL: Riverside.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139-160.

Martineau, J. A. (2004). *The effects of construct shift on growth and accountability models.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35-62.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating.* Los Angeles, University of California, Center for the Study of Evaluation (CSE).

Michigan Department of Education (2005). *Mathematics Grade Level Expectations*. Lansing, MI: Author.

Miller, T. R., & Hirsch, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item-response theory. *Applied Measurement in Education*, *5*(3), 193-211.

Min, K.-S. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations*. Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Muraki, E., & Engelhard, G. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement*, *9*(4), 417-430.

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement*, *37*(4), 357-373.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 253-272). New York: Springer.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed. pp. 221–262*)*. Washington, DC: American Council on Education.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). *The effect of anchor test size in vertical equating with the Rasch and three-parameter models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.

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

Reckase, M. D. (1997b). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.)*, Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics 26: Psychometrics* (pp. 607-641). Amsterdam: North-Holland.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D. (2010). Study of best practices for vertical scaling and standard setting with recommendations for FCAT 2.0. Retrieved from http://www.fldoe.org/asp/k12memo/pdf/StudyBestPracticesVerticalScalingStandardSetting.pdf.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*(3), 193–203.

Reckase, M. D., & Li, T. (2007). Estimating gain in achievement when content specifications change: A multidimensional item response theory approach. In R.W. Lissitz (Ed.) *Assessing and modeling cognitive development in school* (pp. 189-204). Maple Grove, MN: JAM Press.

Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Paper presented to the Committee on Test Design for K-12 Science Achievement, Washington, DC.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*(1), 1-30.

Schmidt, W. H., Houang, R. T., & McKnight, C. C. (2005). Value-added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-added models in education: Theory and applications* (Chapter 6). Maple Grove, MN: JAM Press.

Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

Sinharay, S., & Holland, P. (2006a). *The correlation between the scores of a test and an anchor test* (ETS Research Rep. No. RR-06-04). Princeton, NJ: ETS.

Sinharay, S., & Holland, P. (2006b). *Choice of anchor test in equating* (ETS Research Rep. No. RR-06-35). Princeton, NJ: ETS.

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (Tech. Rep. No. ONR 90-8). Iowa City, IA: ACT.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 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, Department of Psychology, Psychometric Methods Program.

The MathWorks. (2008). *Matlab 2008: The language of technical computing [computer program]*. Natick, MA.

Turhan, A., Tong, Y., & Um, K. R. (2007). *Effects of anchor item properties and dimensionality of test on vertical scaling.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Wang, S., Jiao, H., Young, M. J., & Jin, Y. (2006). *The effects of linking designs in vertical scaling on the growth patterns of student achievement*. Paper presented at the 13th International Objective Measurement Workshop, San Francisco, CA.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347-364.

Yao, L., & Mao, X. (2004). *Unidimensional and multidimensional estimation of vertical Scaled Tests with Complex Structure*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*(4), 299-325.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and thurstone methods of vertical scaling. *Journal of Educational Measurement*, *34*(4), 293-313.