A MULTIDIMENSIONAL AND MIXTURE RANDOM ITEM MODEL FOR MULTIDIMENSIONAL DATA ANALYSIS

by

Yoonsun Jang

(Under the Direction of Seock-Ho Kim and Allan S. Cohen)

Abstract

Multidimensional item response theory (MIRT) models or mixture IRT (MixIRT) models are available for the multidimensional data analyses. The difference of these two IRT models is a type of characteristics used to capture multidimensionality of data. Multidimensionality is explained based on the characteristics of items in MIRT models, while MixIRT models explain multidimensionality as the characteristics of a group of examinees. Sometimes, however, the results of the MIRT or MixIRT models might not be enough to understand multidimensionality of data because it is the results of interaction between examinees and items. The purpose of this study is to propose a multidimensional and mixture random item model (MMixRIM) as an alternative IRT model for multidimensional data analyses. This proposed model is a combination of MIRT model, MixIRT model, and the random item model, and can provide information about both examinees and items to understand multidimensionality of data. One empirical study and one simulation study were conducted to compare the performances of the multidimensional two-parameter logistic (M2PL) model, two-parameter MixIRT (Mix2PL), and two-dimensional MMixRIM (2DMMixRIM) for the multidimensional data analysis. Results of the empirical study indicated that 2DMMixRIM detected some items that measure different latent trait between latent classes, whereas these

items measure the same latent traits based on the analysis by using the M2PL model. Further, results of the simulation study suggested that the Mix2PL model and 2DMMixRIM showed better performances than the M2PL model for the correct model selections based on AIC, BIC, CAIC, AIC_C, and ABIC. On the other hand, recovery of item parameters and class memberships estimated by the M2PL and Mix2PL models were better than MMixRIM.

INDEX WORDS: Multidimensionality, multidimensional item response theory, mixture item response theory, random item model

A MULTIDIMENSIONAL AND MIXTURE RANDOM ITEM MODEL FOR MULTIDIMENSIONAL DATA ANALYSIS

by

Yoonsun Jang

B.S., Ehwa Womens University, Seoul, Korea, 2004

M.A., Ehwa Womens University, Seoul, Korea, 2008

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2016

A MULTIDIMENSIONAL AND MIXTURE RANDOM ITEM MODEL FOR MULTIDIMENSIONAL DATA ANALYSIS

by

Yoonsun Jang

Approved:

Major Professors:	Seock-Ho Kim Allan S. Cohen
Committee:	Gary J. Lautenschlager Zhenqiu (Laura) Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2016

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

In this chapter, the purpose of this study is outlined with a brief review of early research on estimating the effects of multidimensionality on unidimensional IRT models. In addition, a short discussion is presented about current models applied to analyze multidimensional data.

1.1 Statement of the Problem

Unidimensionality is one of the fundamental assumptions in item response theory (IRT; Lord, 1980; Lord & Novick, 1968). An important advantage of unidimensional IRT models is that these models have relatively simple mathematical functions for describing the relationship between persons' characteristics and features of items. In actual educational or psychological measurement situations, however, unidimensionality (i.e., a single person parameter) is a strong assumption (Reckase, 2009). For example, in reality, students can use more than a single ability, such as geometry and algebra, to solve a particular mathematics item; additionally, some items are specifically developed to measure abilities in both geometry and algebra.

A number of studies have investigated the effects of violations of unidimensionality on parameter estimation and have concluded that parameter estimation was somewhat robust to violation of the unidimensionality assumption. Reckase (1979) applied the one- and three-parameter unidimensional IRT models to analyze five sets of empirical multidimensional data and five sets of simulation data. One simulation data set was simulated to be unidimensional data, one set was simulated as two-factor data with a dominant first factor, and three data

sets were estimated to have nine factors with different degrees of dependence between factors. Reckase found that the ability estimates of one-parameter and three-parameter IRT models were different when factors were independent, whereas both models estimated the first principal component when factors were largely dependent. In particular, when the first factor explained at least 20 percent of the variance, the unidimensional IRT model was acceptable for multidimensional data.

Drasgow and Parsons (1983) conducted a simulation study to examine the effects of multidimensionality on the estimations of item and person parameters in the IRT calibration program LOGIST (Wood, Wingersky, & Lord, 1976). This was accomplished by generating five sets of item responses based on five different degrees of correlations between a general factor and common factors. Drasgow and Parsons used the root mean squared differences (RMSDs) for evaluating estimated item parameters and correlations between estimated ability parameters and factor scores as criteria for evaluation of parameter estimation. The results of that research showed that the RMSDs for item discrimination and item difficulty increased as the strength of the general factor decreased. Further, the correlations between estimated ability parameters and factor scores decreased as the correlation between the general factor and the common factors weakened. As in Reckase (1979), Drasgow and Parsons also concluded that a unidimensional model was inappropriate when there was no sufficiently dominant factor.

Dorans and Kingston (1985) examined the effects of violation of unidimensionality on the parameter estimates of items used for equating GRE verbal data. The GRE verbal test consisted of reading comprehension items and discrete verbal items. These two parts of the test measured two highly correlated but distinct abilities, verbal ability and reading comprehension. Dorans and Kingston compared three sets of estimated ability and item parameters. One set of parameters was estimated using only the reading comprehension items, and the second set of estimated parameters was obtained from the discrete verbal items on the test. The last set of estimated parameters was calibrated using both the reading comprehension items and the discrete verbal items. The equated scales were compared to

detect the effects of the violation of unidimensionality on the equating. The conclusion of this research was that, although the results indicated the effect of multidimensionality on the estimated item discriminations and equating, the effects appeared to be insignificant.

Harrison (1986) extended Drasgow and Parsons' (1983) research to more complicated simulation conditions: variations in test length, number of common factors, the distribution of item loadings on common factors, and correlations between the general factor and common factors. The RMSDs for the item parameter estimates from the computer program LOGIST (Wood, Wingersky, & Lord, 1976), which uses joint maximum likelihood estimation, were used as a criterion for evaluating the effect of multidimensionality. Harrison used RMSDs for ability parameter estimates as well. Harrison concluded that the LOGIST parameter estimates were robust to violations of unidimensionality and suggested increasing the number of test items for each factor, increasing the number of common factors, and having an approximately equal number of items for each common factor in order to improve estimation.

Kirisci, Hsu, and Yu (2001) investigated the sensitivity of computer programs, BILOG (Mislevy & Bock, 1990), MULTILOG (Thissen, 1991), and XCALIBRE (Assessment Systems Corporation, 1996), to the violation of unidimensionality and normality. These programs use marginal maximum likelihood to estimate model parameters. The effects of multidimensionality were found to depend on not only the correlation between dimensions, but also the estimation program. According to the results from this study, the estimated item and ability parameters from BILOG had the smallest root mean squared error (RMSE), but the estimated item difficulties from MULTILOG and the estimated item and ability parameters from XCALIBRE were more consistent between replications.

These studies provided evidence of some degree of robustness to violations of unidimensionality; however, these studies generally were limited in that a dominant factor existed in the simulated data. The latent structure used in Dorans and Kingston (1985), for example, was distinct but had highly correlated dimensions. Harrison (1986) used both moderately and highly correlated latent structures (the range of the correlation between the general

factor and common factors was .65 to .80 for the moderately correlated structure and .80 to .95 for the highly correlated structure). Kirisci et al. (2001) suggested the use of multidimensional IRT models (MIRT) if the dimensions were weakly correlated (less than .4). This leaves open the question of whether or not a dominant common factor may be required for accurate parameter estimation of multidimensional data with unidimensional IRT models. It is possible that the estimated parameters may be biased due to the misspecification of a model when unidimensional IRT models are applied to multidimensional data. In such a case, the estimated parameters from unidimensional IRT models for multidimensional data would not accurately reflect the features of the data (Batley & Boss, 1993).

Reckase (2009) notes two types of multidimensionality: content-based multidimensionality and sample based multidimensionality. In many cases, researchers develop an assessment based on a substantive theory. For example, a science assessment contains items to measure multiple concepts, such as osmosis, filtration, and diffusion. In this case, MIRT models could be used. MIRT models are mathematical functions used for describing a relationship between a person's location in a multidimensional coordinate space and a probability of a correct response. This relationship is mediated by a set of item parameters (Reckase, 2009). The set of item parameters represents the features of the test items and can be interpreted based on the theory of multiple concepts in the test. In this case, MIRT models would provide information about item characteristics related to a multidimensional structure.

The second type of multidimensionality involves sample-specific characteristics rather than a feature of the test (Reckase, 1990, 2009). Reckase (1990) classified the usage of the term dimensionality in psychological and educational tests into two categories: psychological dimensionality and statistical dimensionality. The psychological dimensionality is the number of hypothesized constructs of a test. Thus, this is equivalent to the content-based multidimensionality described above. The statistical dimensionality is the required minimum number of mathematical variables to summarize an item response matrix. That is, the statistical dimensionality is a property of a data matrix, not of the test or the examinee population.

To be specific, it is possible the item response data will not meet the assumption of unidimensionality, even though the test was designed to measure a single ability because of the variability due to the particular sample of examinees. A variation of examinees may be caused, for example, by differences in examinees' backgrounds, differences in use of problem solving strategies, or difference in stages of cognitive development. For this case, mixture IRT (MixIRT; Mislevy & Verhelst, 1990; Rost, 1990) models might be more useful for modeling multidimensionality.

MixIRT models are extensions of traditional IRT models. MixIRT models allow for a heterogeneous population, which consists of unobserved subpopulations (Rost, 1990). Each subpopulation is latent in the data, so it is referred to as a latent class. Each latent class is characterized by a unique set of item parameters. Thus, multidimensionality can be understood based on information about the characteristics of a group of examinees in MixIRT models.

As mentioned above, multidimensionality also may be a sample-specific property. There are various possible factors that cause the multidimensionality. For instance, differences in teaching methods or differences in anxiety level of examinees may cause multidimensionality in the data (Nandakumar, 1993). Item information from MIRT models or examinee information from MixIRT models might not be sufficient to fully explain the multidimensional structure. The combination of a MIRT model and a MixIRT model can provide simultaneous information not only about items but also about persons (e.g., Choi & Wilson, 2015).

Both MIRT models and MixIRT models can be used as exploratory models. For example, McKinley and Way (1992) applied the multidimensional three-parameter logistic model (M3PL) to discover the best fitting model for the four sets of responses data to the 146 TOEFL items. Mislevy and Verhelst (1990) used a MixIRT model to describe differences in solution strategies used by examinees taking a physics test. When MIRT or MixIRT models are used for an exploratory purpose, researchers might face difficulties in defining latent traits from MIRT models or latent classes from MixIRT models.

Several explanatory IRT models have been developed to improve the interpretation of results from exploratory analyses. The linear logistic test model (LLTM) is an explanatory IRT model, which can incorporate item properties to explain the difference of item difficulties across items (Wilson & De Boeck, 2004). For example, Kubinger (2009) provided some applications of LLTM to estimate the item position effect, the speededness effect, and the effect of item response format. Random item IRT models (RIMs; De Boeck, 2008) also have been found to be more useful for explanatory analyses than traditional IRT models. This is because RIMs treat both persons and items as random, while traditional IRT models treat items as fixed and persons as random (De Boeck, 2008). Random item parameters are also sometimes more realistic than fixed item parameters, because the effect of item properties might differ across persons. In addition, these item properties might not perfectly explain the differences of item parameters in LLTM (Cho, Gilbert, & Goodwin, 2015; Choi & Wilson, 2015; Jeon, Draney, & Wilson, 2015).

Recently, a number of studies have focused on incorporating RIMs to MIRT, MixIRT, or multidimensional mixture IRT (MMixIRT) models. Frederickx, Tuerlinckx, De Bock, and Magis (2010) introduced a random item mixture model for use in selection of a set of anchor items for detecting differential item functioning (DIF). In that study, items were treated as random. That is, items were treated as being randomly selected from a universe of items. De Jong and Steenkamp (2010) described a finite mixture multilevel multidimensional ordinal IRT model to assess measurement invariance across nations in a large scale cross-cultural study. Instead of treating items as fixed within each subpopulation, as in traditional mixture models, that model allowed item parameters to be assumed to have random distribution within each subpopulation.

An explanatory multidimensional multilevel random item response model was proposed by Cho, Gilbert, and Goodwin (2015). This model incorporated a multilevel structure for the explanatory MIRT model to allow random item parameters at both item and item group levels. Jeon, Draney, and Wilson (2015) proposed a general model as a combination of a linear

logistic latent test model with an error term, a multidimensional Rasch model, and a saltus model. Choi and Wilson (2015) extended the random weights linear logistic test model to mixture modeling framework to identify latent classes that have different multidimensional structures.

As described above, a number of studies introduced alternative approaches that extended models to the random item framework. Few studies, however, considered an extension to multidimensional mixture IRT models with random person and item parameters. Such an extension could produce a model that might be more realistic and informative about multidimensionality in item response data. Choi and Wilson's (2015) proposed model, a mixture generalization of the random weights linear logistic model, cannot be strictly considered as a random item model because the random coefficients vary across persons within an item. Thus, a mixture generalization of the random weights linear logistic model can be considered as a within items multidimensional mixture model. The model proposed by Jeon et al. (2015), a general saltus LLTM-R, is a special case of an extension of the multidimensional random-item mixture IRT model. The saltus model assumes that items within the same subgroup of items have the same amount of differences of item difficulties between latent classes.

1.2 Purpose of the Study

In this study, an alternative model that combines the MIRT and the MixIRT model and treats both persons and items as random (MMixRIM) is proposed for multidimensional data analyses. The motivation for developing the MMixRIM is to provide useful information about the structures of the multidimensionality that otherwise might not be evident in the usual MIRT or MixIRT models. As described below, the MMixRIM provides information about both persons and items to explain multidimensionality. The purpose of this study is (1) to explore the performance of the MMixRIM for analysis of multidimensional item responses data, and (2) to investigate the effects of a multidimensional structure on the estimation of

model parameters. This model will be compared with the performance of both MIRT models and MixIRT model under several conditions of dimensional structures.

Chapter 2

THEORETICAL FRAMEWORK

This chapter begins with a discussion about the definition of the dimensionality of data. In the following sections, several relevant models, MIRT models, MixIRT models, and RIMs, are reviewed. A brief history, main features, and methodological issues for each model are described. Additionally, the multidimensional mixture IRT model with random person and random item parameters is described.

2.1 Dimensionality

The dimensionality of data, either unidimensionality or multidimensionality, is an essential and commonly used term in psychological and educational measurement models. Nevertheless, the definition of dimensionality is varied, somewhat abstract and non-operational (Hambleton & Rouvinelli, 1986). For example, one common definition of dimensionality is the grouping of items on a test, such as "dimensions should exist when items on a test can be grouped into homogeneous bundles" (Walker, Azen, & Schmitt, 2006, p. 722). Li, Jiao, and Lissitz (2012) defined dimensionality as "the number of latent traits test developers would like to extract from the test" (p. 3). In these examples, the dimensionality is considered as a characteristic of a test and mainly based on some hypothesized psychological constructs. In addition, Svetina and Levy (2014) defined dimensionality according to the modeling framework. The dimensionality is the number of factors that "account for student performance on a particular measure within a factor analytical framework [and] the number of latent

variables is necessary to achieve local independence and monotonicity within an IRT framework" (Svetina & Levy, 2014, p. 37). This definition of dimensionality is mainly based on statistical characteristics of the data matrix.

As such, there are various definitions of dimensionality, and the dimensionality commonly used is referred to as the characteristics of the test, such as a bundle of items which measures the same latent trait. As pointed out in several studies (Reckase, 1990, 2009; Svetina & Levy, 2014; Walker, Azen, & Schmitt, 2006), however, this is a much simpler way of defining dimensionality, because the dimensionality is caused by the interaction of items and a sample of examinees. Furthermore, the ignored examinees' characteristics may result in fewer dimensions in the data and misinterpretations. Accordingly, a more technical and realistic definition of dimensionality should be considered in multidimensional research.

Lord and Novick (1968) indicate that the complete latent space should have the same conditional distribution of item scores for fixed latent traits for the population. Consequently, the complete latent space includes not only psychological "important" latent traits that affect examinees' responses to the items, but also latent traits referred to as an "error of measurement." The dimensionality is the number of all latent traits of the complete latent space.

The common factor model is traditionally used to define the dimensionality of a test. The set of items is unidimensional if and only if one common factor model fits the data. The common factor model, however, requires a strong assumption, known as local independence, that items are entirely statistically independent, when the common factors are partialled out as in the equation below:

$$P(U_N = 1|\boldsymbol{\Theta} = \theta) = \prod_{i=1}^{N} P(U_i|\boldsymbol{\Theta} = \theta), \tag{2.1}$$

where $P(U_N = 1|\boldsymbol{\Theta} = \theta)$ is the conditional distribution, U_N is random test responses for a randomly selected examinee, $\boldsymbol{\Theta}$ is a vector of the latent trait, θ is a particular value of $\boldsymbol{\Theta}$, U_i is a randomly selected examinee's response to i-th item, and N is the test length.

McDonald (1981) suggested the definition of dimensionality with the weak assumption that the conditional pair-wise covariances are zero, and nothing higher than the second order joint moments is considered. This can be expressed as the following equation:

$$P(U_i, U_j|\mathbf{\Theta} = \theta) = P(U_i|\mathbf{\Theta} = \theta)P(U_j|\mathbf{\Theta} = \theta). \tag{2.2}$$

Unlike traditional dimensionality that counts all latent traits, Stout (1987, 1990) also suggested a more relaxed definition of dimensionality called 'essential dimensionality,' with a weaker condition of local independence, namely 'essential independence.' The 'essential dimensionality' is the number of dominant latent traits required to satisfy the 'essential independence.' The 'essential independence' can be expressed as below:

$$\sum_{1 \leq i \neq j \leq N} cov(U_i, U_j|\mathbf{\Theta} = \theta) \approx 0, \text{when } N \to \infty. \tag{2.3}$$

That is, the 'essential independence' is the condition that the average of conditional covariances given any particular θ is close to zero as the test length N increases. On the contrary, the traditional local dependence requires the condition that covariances for all pairs of items for all θ must be zero.

As described above, dimensionality is the result of interaction between persons and items. There are many possible factors that affect dimensionality. Moreover, the complete latent space indicated by Lord and Novick (1968) may be too strong in real educational or psychological measurement situations. Therefore, the definition of dimensionality based on statistical characteristics of the data matrix, that is statistical dimensionality as defined by Reckase (1990), with a weak assumption of local independence was adopted in this study. In the following, four different measurement models that can be applied to multidimensional data analysis are described in detail.

2.2 Multidimensional item response theory models

Two statistical methodologies can be considered as bases of MIRT. The one methodology is a factor analysis (FA), and the other is IRT (Reckase, 1997a, 2009). MIRT can be considered as a special case of FA in the perspective of the trying to define unobserved or latent traits and scales by using a data matrix of responses to items. Although both FA and MIRT share many similarities, the goal of each methodology differs. FA mainly focuses on the identifying the structure of latent traits (i.e., factors) in a multidimensional space to reflect similarity among the observed responses (Mulaik, 2010). Unlike FA's main goal of a parsimonious explanation of the data matrix, MIRT focuses on characteristics of items, such as item difficulty to model the interactions between examinees and items (Reckase, 1997b, 2009).

McDonald (1967, 1997) demonstrated that the normal ogive model (Lord, 1952) is approximately equivalent to the nonlinear common factor model; additionally he extended the unidimensional normal ogive model to the multidimensional normal ogive model, called the NOHARM (Normal-Ogive Harmonic Analysis Robust Method) model (Normal-Ogive Harmonic Analysis Robust Method), by defining as following:

$$P(U_i = 1|\theta) = N\{\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \cdots + \beta_{ik}\theta_k\}, \tag{2.4}$$

where U_i is a randomly selected examinee's binary response, which is coded as 0 for incorrect response and as 1 for correct response, to i-th item as before, $P(U_i = 1|\theta)$ is a conditional probability of correct response, β_{i0} is the intercept of i-th item, β_{ik} is factor loading of i-th item on k-th factor, θ is latent trait with k components (i.e., the number of dimensions is k), it is assumed that θ is k-variate normally distributed with zero means and unity variances. $N\{t\}$ is the cumulative normal distribution function and is defined as below:

$$N\{t\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty} \exp\left(\frac{-z^2}{2}\right) dz. \tag{2.5}$$

The other foundational methodology of MIRT is IRT. MIRT can be considered as a multidimensional extension of IRT to model the relation between examinees' responses and a set of items of a test with relaxed assumption that allows multidimensional latent traits measured by a test. Rasch (1961) mentioned about the possibility of application of his model to a higher dimension (i.e., multidimensionality), and Samejima (1974) also proposed the two-parameter normal ogive model in the multidimensional latent space for the continuous responses. McKinley and Reckase (1982) investigated Rasch's (1961) generalized Rasch model and concluded that the reduced vector and product term model was the most capable model in realistic multidimensional data. Later, they proposed the multidimensional two-parameter logistic (M2PL) model which is given as

$$\begin{aligned} &P(U_i = 1|\theta_1, \theta_2, \ldots, \theta_k, a_{i1}, a_{i2}, \ldots, a_{ik}, d_i) \\ &= \frac{\exp(a_{i1}\theta_1 + a_{i2}\theta_2 + \cdots + a_{ik}\theta_k + d_i)}{1 + \exp(a_{i1}\theta_1 + a_{i2}\theta_2 + \cdots + a_{ik}\theta_k + d_i)}, \end{aligned} \tag{2.6}$$

where U_i is the same as defined above, $\theta_1, \theta_2, \ldots, \theta_k$ represent k latent traits, $a_{i1}, a_{i2}, \ldots, a_{ik}$ are item discrimination parameters of i-th item for k latent traits, and d_i is a intercept parameter of i-th item. This equation also can be simply expressed by using a $1\times k$ vector of latent traits,$\boldsymbol{\Theta}$, and a $1\times k$ vector of item discrimination parameters, $\boldsymbol{a}_i$, as following:

$$P(U_i = 1|\boldsymbol{\Theta}, \boldsymbol{a}_i, d_i) = \frac{\exp(\boldsymbol{a}_i{}'\boldsymbol{\Theta} + d_i)}{1 + \exp(\boldsymbol{a}_i{}'\boldsymbol{\Theta} + d_i)}, \tag{2.7}$$

In Equation 2.6, $a_{i1}, a_{i2}, \ldots, a_{ik}$ are equal to item discrimination parameters in a two-parameter logistic IRT model; however, d_i is not equivalent to item difficulty parameter in a usual IRT model. Instead, the negative of d_i divided by item slope parameter $(-\frac{d}{a_k})$ for each dimension means the relative item difficulty. For instance, suppose that (1) two-dimensional data is used, (2) estimated item slope parameters of a specific item for two dimensions are $a_1 = 1.2$ and $a_2 = 0.3$, and (3) the estimated item intercept parameter of the item is $d = 1.0$. The probability of a correct item response is .5 when $1.2\theta_1 + 0.3\theta_2 + 1.0 = 0$, and this can be represented by a linear line as shown in Figure 2.1. If θ_1 is zero, the probability of a correct

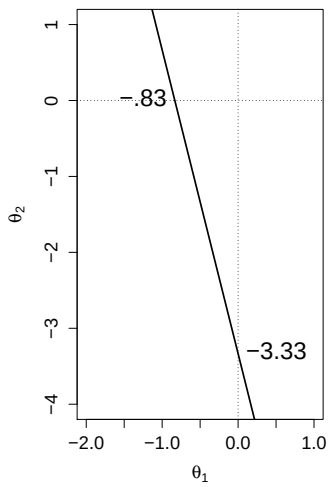

Figure 2.1: Line represented .5 probability of correct response for item with $a_1 = 1.2$, $a_2 = 0.3$, and $d = 1.0$.

response of this item is .5 when $a_2\theta_2 + d$ equals zero. That is, $0.3\theta_2 + 1.0 = 0$. Thus, the probability of a correct item response is .5 when $\theta_1 = 0$ and $\theta_2 = -\frac{d}{a_2} = -\frac{1.0}{0.3} \cong -3.33$, and $\theta_2 = -\frac{d}{a_2}$ is a relative item difficulty related to the second dimension. Similarly, $\theta_1 = -\frac{d}{a_1} = -\frac{1.0}{1.2} \cong -.83$ has a .5 probability of a correct response of this item when θ_2 is zero. These two relative item difficulties of this item are intercepts of two axes in Figure 2.1. For the same interpretation as the item difficulty in a regular IRT model, the distance from the origin to the line that represents a .5 probability of correct responses is used. The distance is called MDIFF, and it can be calculated with the following equation:

$$MDIFF = -\frac{d_i}{\sqrt{\sum_{v=1}^{k} a_{iv}^2}}, \tag{2.8}$$

where the denominator ($\sqrt{\sum_{v=1}^{k} a_{iv}^2}$) is called MDISC.

In the M2PL model, the probability of a correct response is affected by a weighted linear function of latent traits (i.e., $\boldsymbol{a}_i'\Theta$), and this reflects a compensatory relationship. In a

compensatory model, a lack of a specific ability can be compensated for by other abilities. For example, if several cognitive strategies are required to solve a mathematics item, one highly developed strategy (or ability, e.g., making a table) may be able to compensate for a less well-developed strategy ability (e.g., drawing). In contrast, in some situations, a compensatory relationship might not be possible. For example, the lack of mathematics ability typically cannot be compensated for by a high reading ability for solving a mathematics test item. In this case, a non-compensatory model may be more appropriate. Sympson's (1978) model and Whitely's (1981) multicomponent latent trait model (MLTM) can be considered partially compensatory models. The MLTM is given by the following equation:

$$P(U_i = 1|\theta_1, \theta_2, \ldots, \theta_k) = \prod_{v=1}^{k} \frac{\exp(\theta_v - b_{iv})}{1 + \exp(\theta_v - b_{iv})}, \tag{2.9}$$

where U_i is a dichotomous response to i-th item (0 or 1), $\theta_1, \theta_2, \cdots, \theta_k$ are k latent traits, and $b_{i1}, b_{i2}, \ldots, b_{ik}$ are the component-specific difficulty parameters of i-th item. In MLTM, the probability of correct response is a product of probability of being successful on one component (i.e., latent trait) of the item.

The decision of whether to use a compensatory model or a partially compensatory model would depend on a researcher's hypothesis about the test and the item(s) in question. Partially compensatory models may be more appropriate than compensatory models in some cases; however, partially compensatory models require a large enough variability in difficulty to accurately estimate the component-specific difficulty parameters (Bolt & Lall, 2003). For this reason, this study focused on only compensatory models.

2.3 Mixture item response theory models

Basically, the traditional IRT models require three assumptions, which are (1) unidimensionality, (2) local independence, and (3) monotonicity, to guarantee the accurate estimation

of parameters. Realistically, however, these assumptions may not always hold. For instance, the relationship between the ability and the item differs for different groups of examinees for several reasons, such as differential use of cognitive strategies, cultural background, or educational experiences. Thus, the item responses may depend not only on ability but also on other latent traits. In these cases, the unidimensional assumption might be violated, and alternative IRT models may be required.

MixIRT models are combination models of the latent class analysis (LCA) and IRT models. As the name implies, LCA considers latent categorical variables. In LCA, the population consists of unobserved subpopulations, also called latent classes, that is, a heterogeneous population. Additionally, the probability of individual responses is a function of the probability of membership in each latent class, and the probabilities of each observed response is conditional on latent class membership (Collins & Lanza, 2010).

Although the LCA is flexible with respect to a homogeneous population assumption, there is another strong requirement that all examinees in the same latent class must have the same response probabilities. In other words, all examinees' abilities are equal within a latent class. Rost (1990) described a mixture Rasch model which combines the LCA and Rasch models.

Based on features of the LCA and IRT models, the main features of the MixIRT model are that the population is heterogeneous and that it allows different sets of item parameters for each subpopulation or latent class, while the IRT model holds only one set of item parameters for an entire population. Thus, examinees can be characterized by not only a latent ability but also different item parameters between latent classes (Cohen & Bolt, 2005). The mixture Rasch model (MRM; Rost, 1990) is defined as follows:

$$P(u_{ij}=1)=\sum_{g=1}^{G}\pi_g \times P(u_{ij}=1|g,\theta_{jg})=\sum_{g=1}^{G}\frac{\pi_g}{1+\exp(-\theta_{jg}+b_{ig})}, \tag{2.10}$$

where u_{ij} is the response to the i-th item of the j-th examinee who belongs to latent class g; θ_{jg} is the latent ability of the j-th examinee within in latent class g; π_g is the probability of

membership in latent class g with the constraint that $\sum_{g=1}^{G} \pi_g = 1$, b_{ig} is the item difficulty of i-th item in latent class g, and $P(u_{ij} = 1|g, \theta_j)$ indicates the conditional probability of a correct response to the i-th item given an ability θ_{jg} and latent class membership g.

In MixIRT modeling, interpreting and defining the qualitative differences between latent classes may be difficult because of their exploratory nature. Furthermore, a large sample size is required to get accurate estimate results (Smit, Kelderman, & van der Flier, 1999). Because latent classes are permitted to hold different sets of item parameters in MixIRT models, larger numbers of parameters than those present in a traditional IRT model will need to be estimated. For example, 10 item difficulties are estimated when the Rasch model is applied to the analysis of a test with 10 dichotomous items, while the number of item parameters is doubled, when the two-class MRM is used for the same test data (i.e., 10 item parameters for each latent class). In addition, the number of estimated item parameters increases as the number of latent classes increases. For these issues in MixIRT models, some additional information would be helpful to improve the interpretation of the results and the accuracy of parameter estimations.

Mislevy and Verhelst (1990) extended the linear logistic test model (LLTM; Fischer, 1973) to the mixture model. In the LLTM, information about the characteristics of items based on the substantive theory is used to model the relationship between an examinee's ability and items as the following equation:

$$P(u_i = 1|\theta_j, b'_i) = \frac{\exp(\theta_j - b'_i)}{1 + \exp(\theta_j - b'_i)}, \tag{2.11}$$

Equation 2.11 is exactly equal to the Rasch model, but differs in that item difficulty, b'_i, is a linear combination of item characteristics and is defined as follows:

$$b'_i = \sum_{q=1}^{Q} w_{iq} b_q, \tag{2.12}$$

where $w_{iq}(q = 1, \ldots, Q)$ indicates the exhibition of the q-th item characteristics of the i-th item characteristics. That is, this can be considered as an element of a Q-matrix, which is a

matrix that indicates which attributes are measured by each item by coding 0 or 1, and b_q indicates the contribution of the q-th item characteristics to item difficulty. Thus, LLTM is an explanatory IRT model. Similar to MRM, the marginal probability of the correct response of the j-th examinee to the i-th item in the mixture LLTM model is given by

$$P(u_{ij}=1)=\sum_{g=1}^{G}\pi_g \times P(u_{ij}=1|g,\theta_{jg})=\sum_{g=1}^{G}\frac{\pi_g}{1+\exp(-\theta_{jg}+\sum_{q=1}^{Q}w_{iq}b_{qg})}, \tag{2.13}$$

where b_{qg} are class-specific item characteristics coefficients. By adding auxiliary variables about item characteristics, the features of the latent classes would be more easily interpreted. In addition to this, the estimation of parameters can improve by using a more parsimonious model that includes certain constraints based on pre-hypotheses concerning item characteristics.

Smit et al. (1999, 2000) incorporated collateral information in the MRM and the mixed Birnbaum model to reduce the standard error of parameter estimation and improve latent class classification. Item responses, u_j, were assumed to be associated with the latent traits, $theta_j$ and the latent class, g_j, and the dichotomous collateral variable to be associated with the latent class variable. From these assumptions, the joint probability can be expressed as follows:

$$P_{\boldsymbol{u}_j,g_j,y_j,\theta_j}=P_{\boldsymbol{u}_j|g_j,y_j,\theta_j}P_{\theta_j|g_j,y_j}P_{g_j,y_j}=P_{\boldsymbol{u}_j|g_j,\theta_j}P_{\theta_j}P_{g_j,y_j}, \tag{2.14}$$

where $\boldsymbol{u}_j$ is a vector of item responses for the j-th examinee, g_j is a latent class membership for the j-th, y_j is a binary group indicator for the j-th examinee, and θ_j is a latent trait for the the j-th examinee.

Dai (2013) included the manifest group indicator, such as a gender, in MRM as dichotomous predictor. In his model, the probability of membership in latent class g, π_g, is modeled as a logistic regression model with the a group indicator in the following equation:

$$Logit(\pi_{jg})=\beta_{0g}+\beta_{1g}y_j, \tag{2.15}$$

where π_{jg} is a j-th examinee's probability of belonging to latent class g, β_{0g} and β_{1g} are regression coefficients, and y_j is a binary group indicator for the j-th examinee.

Based on the previous studies, incorporating auxiliary variables, whether they are associated with persons or items is helpful for improving parameter estimation, latent class classification, and interpretation of qualitative differences between latent classes in MixIRT models. One issue related to auxiliary variables might arise in that auxiliary variables may have either fixed effects or random effects. For example, one can assume the effect of persons is the same for all persons. Similarly, one can assume that the effect of items is the same for all items. In such a case, the effect of persons or items can be treated as a fixed effect. On the other hand, all persons (or items) within a group might not have the same differences to persons (or items) in other groups. In this case, the effect of an auxiliary variable can be treated as a random effect. Moreover, the effect of the person property may differ over items or the effect of item property may differ over persons. This would be considered a person-by-item property. This type of property is considered in DIF analyses. The following section focuses on random item models, and the issue of incorporating auxiliary variables.

2.4 Random item models

In general, estimation of IRT models has treated items as fixed and persons as random samples from the population. This type of model can be called a 'random person and fixed item model'. In this model, person parameters (i.e., ability) are commonly integrated out by being considered a random component, and then, the item parameters are estimated independently. After the estimation of the item parameters, the ability parameters are estimated based on the estimated item parameters in the previous step. In many educational and psychological measurement situations, however, the main interest is the measurement of persons' latent traits and items are selected from item banks to measure specific latent

traits. Thus, it is more appropriate to consider items as random, such as a person treated as a random component.

De Boeck (2008) investigated, "why items may be considered random some of the time?" The first reason is that, as was described above, the items are from some pre-existing item pool. When this is considered an item population, then items are considered as randomly sampled from this item population. The second reason concerns the measurement uncertainty about the parameters. In Bayesian estimation, measurement uncertainty is quantified by using a probability distribution of unknown parameters. A prior distribution with an unknown variance can reduce this uncertainty.

In addition to these two theoretical reasons, a more practical reason is that IRT models commonly estimate item parameters and an ability distribution using the marginal maximum likelihood method, although the educational or psychological measurement has been interested in the estimation of a person's latent trait. Moreover, the generalization over persons is available based on the estimated ability distribution, but the generalization over items is unavailable with the estimated fixed item parameters. Thus, treating items as random components provides a possibility for the generalization over items.

The second reason is about explanation of item parameters. When item parameters are estimated as fixed components, it is assumed that item parameters are constants with no variation. However, an item parameter with a possible variation is both theoretically and logically preferable to a fixed item parameter. This can be accomplished by treating an item parameter as random across persons. Janssen, Schepers, and Peres' (2004) random effect LLTM (LLTM-R) and Rijmen and De Boeck's (2002) random weights LLTM (RWLLTM) are applications of random item models for this issue. Janssen et al. (2004) added random item variation to the LLTM. The item parameters in that model are random across items within the item group. Rijmen and De Boeck (2002) allowed interactions between item parameters and persons, so that item parameters are random across persons.

The third issue is with respect to DIF analysis. A group of items assumed to be non-DIF items is needed to serve as an anchor for detection of DIF, and the results of DIF analysis depend on the quality of anchor items. In this case, the random item approach may be an alternative to the fixed item approach. Frederickx et al. (2010) applied random item modeling to the mixture model for detecting DIF. Most studies used MixIRT models to classify examinees into latent classes, whereas Frederickx et al. (2010) used a mixture model to classify items, and the groups of examinees are manifest. To be specific, Frederickx et al. (2010) assumed that the item difficulties were random effects with a two-class normal mixture distribution. One class referred to the DIF class and the other class to the non-DIF class. In this way, the items of the non-DIF class can be used as anchor items.

Fox (2010) added more practical advantages of random item modeling. The number of parameters in a random item IRT (RIM) model is smaller than the number of parameters in a regular IRT model, because the RIM will estimate only two distributions, for the ability and the item, respectively. Additionally, it is possible to handle a hierarchical item structure, such as testlet items or item clusters, in a RIM.

Van den Noortgate, De Boeck, and Meulders (2003) treated both persons and items as random in an IRT model and called it a cross-classification multilevel logistic model. They reformulated the Rasch model to a hierarchical two-level logistic model. The hierarchical two-level logistic model treats item responses as repeated measurements nested within persons as shown in Figure 2.2. This model is equivalent to the general IRT model with a fixed item and random person.

Unlike a general IRT model, however, a cross-classification two-level logistic model treats item responses as repeated measurements nested within both persons and items as shown in Figure 2.3. This model is an IRT model with random items and persons. The first level of the cross-classification two-level logistic model is given by

$$Logit(p_{ij}) = \beta_{0j} + \beta_{1j}, \tag{2.16}$$

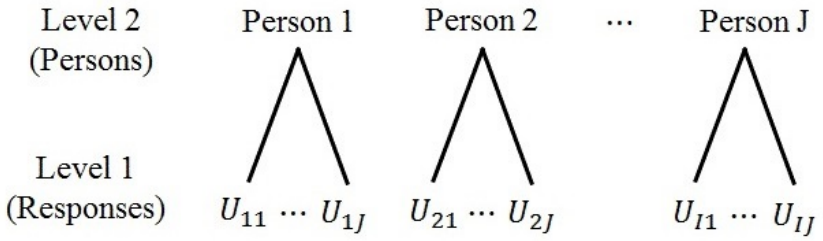

Figure 2.2: Data structure for hierarchical two-level logistic model with random person effects. Adapted from "Cross-classification multilevel logistic models in psychometrics," by W. Van den Noortgate, P. De Boeck, and M. Meulders, 2003, *Journal of Educational and Behavioral Statistics*, 28, p. 370. Copyright 2003 by the American Educational Research Association.

where p_{ij} is a probability of the correct response of the j-th examinee to the i-th item, β_{0j} is equivalent to the j-th examinee's ability, and β_{1j} is the item easiness of the i-th item. The second level of this model can be expressed with the following equations:

$$\beta_{0j} = u_{0j}, \tag{2.17}$$

$$\beta_{1j} = \gamma_0 + u_{1i}, \tag{2.18}$$

where u_{0j} is the random effect of person j, and u_{1j} is the random effect of item i. These two random components are assumed to follow normal distributions with zero means. That is, $u_{0j} \sim N(0, \sigma^2_{u_0})$ with $u_{1i} \sim N(0, \sigma^2_{u_1})$. γ_0 is the mean logit and can be regarded as the mean of item easiness of the item pool. Some auxiliary variables associated with either items or persons can be added as covariates at the second level to improve the explanation of variations in item or ability parameters.

Similar to the cross-classification two-level logistic model in Van den Noortgate et al. (2003), the random item Rasch model can be written as follows:

$$Logit(p_{ij}) = \theta_j - \beta_i, \tag{2.19}$$

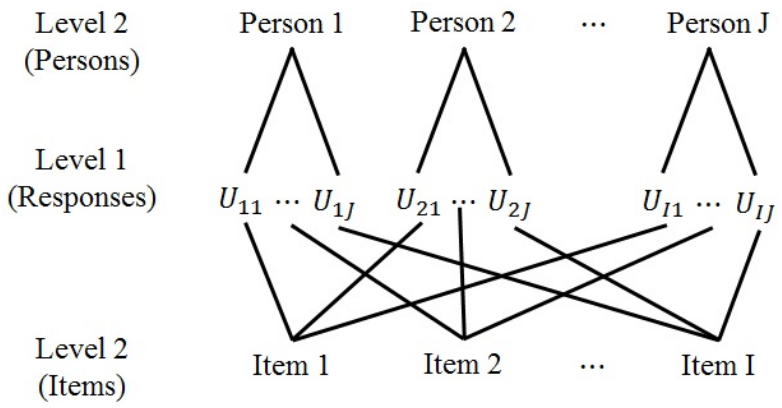

Figure 2.3: Data structure for cross-classified two-level model. Adapted from "Cross-classification multilevel logistic models in psychometrics", by W. Van den Noortgate, P. De Boeck, and M. Meulders, 2003, *Journal of Educational and Behavioral Statistics*, 28, p. 370. Copyright 2003 by the American Educational Research Association.

where θ_j is the j-th person's ability with $\theta_j \sim N(0, \sigma_\theta^2)$ and β_i is the item difficulty of the i-th with $\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$. The difference between the random item Rasch model and a general Rasch model is that the item difficulty has a subscript i to characterize the random sample from the item pool.

Wang (2011) used an extension of MixIRT combined with random item modeling, called the mixture cross-classified item response model (cc-MMixIRTM) to study test speediness. Persons and items were treated as random components, so that covariates related to persons and items could be incorporated in the model. The cc-MMixIRTM with covariates can be expressed in the multilevel framework. The level-1 model is given as

$$Logit(p_{ij}) = \theta_{jg} + \beta_{ig}. \tag{2.20}$$

Equation 2.20 is equivalent to the conditional probability of a correct response to the i-th item given an ability θ_j and latent class membership g in Equation 2.10. The random components related to persons and items are specified in the level-2 model as follows:

$$\theta_{jg} \sim N(\mu_{\theta_{jg}}, \sigma_1^2), \tag{2.21}$$

$$\mu_{\theta_{jg}} = \gamma_{0g} + \sum_{x=1}^{X} \gamma_x X_j, \tag{2.22}$$

$$\beta_{ig} \sim N(\mu_{\beta_{ig}}, \sigma_2^2), \tag{2.23}$$

$$\mu_{\beta_{ig}} = \lambda_{0g} + \sum_{y=1}^{Y} \lambda_y Y_i. \tag{2.24}$$

The person and item covariates of the j-th person and the i-th item are X_j and Y_i, respectively, γ_{0g} is the mean of ability of persons within the latent class g when the effect of person covariate is zero, and $\lambda_{0g}$is the mean of item parameters for the latent class g when the effect of item covariate is zero. Wang (2011) showed that cc-MMixIRT with person and item covariates had smaller bias and RMSE of the estimated item parameters for the non-speed group compared to the unconditional model.

A multidimensional extension of the RIM was applied to analyze multilevel structure data by Cho et al. (2013). Similar to a cross-classification two-level logistic model, the first level was an item response level. The second level was the persons level. Additionally, item-groups were added as the third level in the multidimensional multilevel random item response model (MMRIRM) to handle the dependency of items within the same item group (as illustrated in Figure 2.4). The explanatory MMRIRM with person covariates, item covariates, as well as person by item covariates was applied to analyze three dimensional reading data. Cho et al. (2013) concluded that the additional covariates did not improve the estimation of item and person parameters, because of the multidimensional residual for the persons and items. Instead, the effect of covariates depended on the set of covariates.

As described in the previous sections, some studies have incorporated several approaches in IRT models, such as application of random item modeling to the MixIRT model, MIRT models, or multidimensional extensions of MixIRT models. Although these applications have shown improvement in estimating parameters and understanding results and data, there

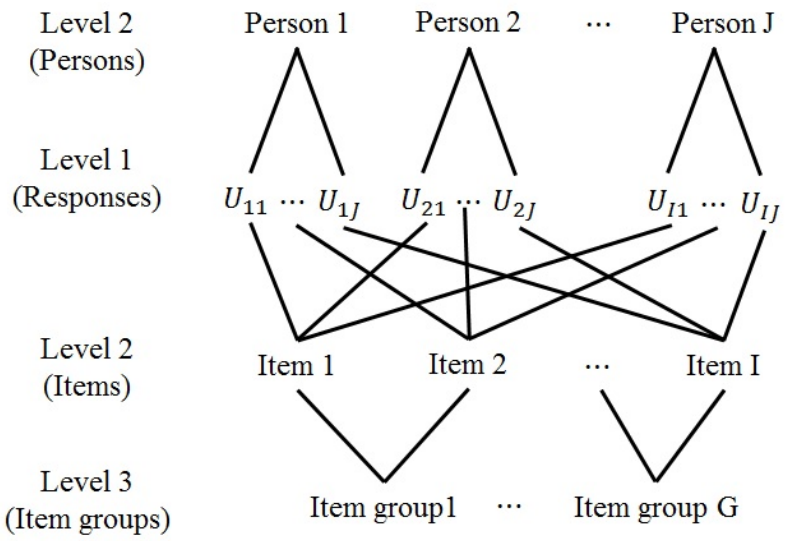

Figure 2.4: Data structure for multidimensional three-level random item response model.

has been little application for understanding of the multidimensionality of the data. Each framework provides specific information based on the statistical features of the approach for multidimensional analysis. In the following section, a multidimensional mixture random item model (MMixRIM), which is combined the MIRT, the MixIRT, and the RIM, is described in detail.

2.5 Multidimensional mixture random item models

The motivation behind proposing a multidimensional mixture random item model (MMixRIM) for multidimensional data analysis is to provide information associated with persons and items and to improve understanding of multidimensionality of the data. A little study proposed multidimensional mixture random item model. Jeon et al. (2015) developed a general Saltus LLTM-R model and provide an application of this model to a cognitive assessment

of deductive reasoning in children. A general Saltus LLTM-R model is a combination of a random effect LLTM, a between items MIRT, and a Saltus model, and is defined as follows:

$$\begin{aligned} P(u_{i(k)j} = 1|\theta_{jk}) &= \sum_{g_k=1}^{G} \pi_{g_k} \times P(u_{i(k)j} = 1|g_k, \theta_{jkg}) \\ &= \sum_{g_k=1}^{G} \frac{\pi_{g_k}}{1 + \exp(-\theta_{jkg} + \sum_{q=0}^{Q} x_{iq} b_q - \varepsilon_i - \sum_{h=1}^{H} \tau_{kgh} w_{ih})}, \end{aligned} \tag{2.25}$$

where π_{g_k} is the probability of membership in latent class g_k in dimension k; θ_{jkg} is the latent ability in dimension k of the j-th examinee in latent class g. x_{iq} indicates the q-th item property of the i-th item; b_q is the effect of the q-th item property; ε_i is a random error and follows $N(0, \sigma^2)$. These three terms came from a random effect LLTM. The last two terms, τ_{kgh} and w_{ih}, reflect a Saltus model, which is a special MRM model to define qualitative differences between latent classes by an effect of item groups called a shift or saltus parameter. τ_{kgh} is the shift parameter for the effect of item group h for latent class g in dimension k; and w_{ih} indicates item group of the i-th item. Jeon et al. (2015) compared the results of the mixture LLTM, unidimensional salute LLTM-R, and two-dimension saltus LLTM-R using the data from the test of deductive reasoning, and concluded that their proposed model can be useful to analyze when an educational or psychological test is well designed based on substantive theory.

In this dissertation, multidimensionality refers to statistical dimensionality, which is a characteristic of the data matrix; it is not theoretical dimensionality, which is a characteristic of the test based on theoretical assumptions. Because of the sample-specific characteristics of dimensionality, several factors should be considered for multidimensional data analysis. The main features of the MMixRIM proposed in this section are (1) compensatory nature, (2) between items and within items multidimensionality, (3) mixture distribution of population, and (4) crossed random variation.

The first feature is a compensatory nature. As described previously, there are two possible multidimensional data situations based on how latent traits interact to response to test items

(De Ayala, 2009). One possible situation is a compensatory situation, the other is a non-compensatory situation. In a compensatory situation, higher latent traits are assumed to compensate for relatively lower latent traits. On the other hand, some higher latent traits cannot compensate for other lower latent traits. In the MMixRIM, multidimensional latent traits are defined as a weighted linear function, reflecting compensatory multidimensionality.

The second feature is related to the structure of latent dimensions. There are two types of multidimensional data: one is between-items multidimensionality, the other is within-items multidimensionality (Adams, Wilson, & Wang, 1997). The between-items multidimensionality means that a test consists of subsets of items that measure one latent trait; within-items multidimensionality means that each item in a test measures several latent traits. These two types of multidimensionality are shown in Figure 2.5. For instance, Item 2 in the case of a between-items multidimensionality measures only Dimension 1, while the Item 2 in the case of a within items multidimensionality measures both Dimension 1 and Dimension 2. The between items multidimensionality has a simple structure, and two types of simple structures are considered. One type is exact simple structure and the other is approximately simple structure. For exact simple structure, only one item discrimination for each item is nonzero and the rest of the item discriminations should be zero. On the other hand, for approximately simple structure, all item discriminations for items can be nonzero, and one item discriminations is dominantly large number, and the rest of item discriminations are relatively small numbers. For example, the item discriminations of Item 1 in Figure 2.5 (a) should be (a_{11}, a_{12}, a_{13})=(1, 0, 0) in the case of exact simple structure. The item discriminations of the same item might be (a_{11}, a_{12}, a_{13})=(1.8, .3, .1) in the case of approximately simple structure. Although exact simple structure is very clear to illustrate between-items multidimensionality, it is unusual in practical situation. This study focuses on both the between-items and within-items multidimensionality with approximately simple structure to reflect a more realistic structure of multidimensionality.

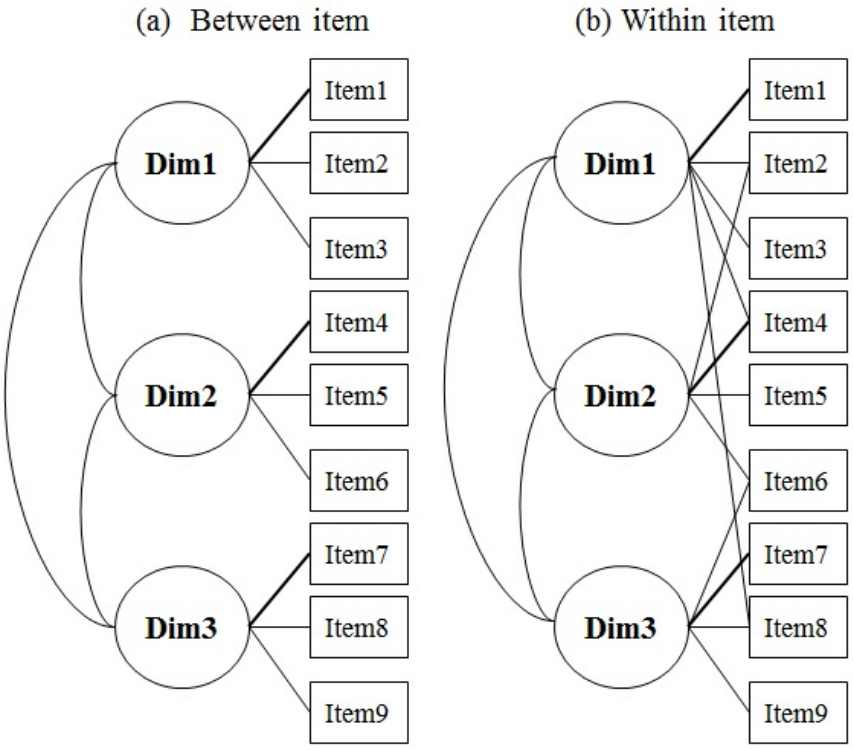

Figure 2.5: Two types of multidimensionality. Adapted from "The multidimensional random coefficients multinomial logit model", by R. J. Adams, M. Wilson, and W. -C, Wang, 1997, *Applied Psychological Measurement*, 21, p. 9. Copyright 1997 by the Sage Publications.

The third feature of the MMixRIM is that it assumes a mixture distribution of the population as multidimensionality could be caused by latent subpopulations. For example, students in different achievement levels would use different abilities to solve the same item, even though the item was developed to measure one ability. Therefore, the responses of students in different achievement levels would be shown to better fit in the multidimensional model than in the unidimensional model. By assuming the mixture distribution of population, the information associated with a person's latent traits can be provided to understand multidimensional data.

The crossed random variation is the fourth feature of the MMixRIM. As described in the previous section, there are several benefits of random item modeling, although treating items

as random in IRT is uncommon. In this study, the person as well as the item is treated as random, that is, the variation of the data is combined with the random item variation and random person variation. The several models explained in this chapter and the MMixRIM are compared based on the features in Table 2.1.

Table 2.1: Overview of Model Characteristics

Models	Population	Item	Ability	Dimension
LCA	Mixture	F	F	U
IRT	Homo	F	R	U
M2PL (McKenley & Reckase, 1983)	Homo	F	R	M
Random item IRT (De Boeck, 2008)	Homo	R	R	U
LLTM (Fischer, 1973)	Homo	F	R	U
LLTM-R (Janssen et al., 2004)	Homo	R	R	U
RWLLTM (Rijmen & De Boeck, 2002)	Homo	F	R	M
MRM (Rost, 1990)	Mixture	F	R	U
MixLLTM (Mislevy & Verhelst,1990)	Mixture	F	R	U
MixRWLLTM (Choi & Wilson, 2015)	Mixture	F	R	M
cc-MMixIRTM (Wang, 2011)	Mixture	R	R	U
general saltus LLTM-R (Jeon et al., 2015)	Mixture	R	R	M
MMixRIM	Mixture	R	R	M

Note. Homo = Homogeneous; F = Fixed; R = Random; U = Unidimensional; M = Multidimensional.

To specify a MMixRIM, first, the M2PL is extended into the MixIRT context as follows:

$$P(u_{ij}=1)=\sum_{g=1}^{G}\pi_g \times P(u_{ij}=1|\boldsymbol{\theta}_{jg},\boldsymbol{a}_{ig},d_{ig},g), \tag{2.26}$$

where $P(u_{ij}=1|\boldsymbol{\theta}_{jg},\boldsymbol{a}_{ig},d_{ig},g)$ is a conditional probability of the j-th examinee within latent class g giving a correct response to the the i-th item, and the π_g is the probability of

membership in latent class g. $P(u_{ij}{=}1|\boldsymbol{\theta}_{jg}, \boldsymbol{a}_{ig}, d_{ig}, g)$ is equivalent to M2PL, described in Equation 2.7 in the previous section, and the only difference is that item parameters and person parameters are subscripted to indicate the latent class as shown in the following equation:

$$P(u_{ij} = 1|\boldsymbol{\theta}_{jg}, \boldsymbol{a}_{ig}, d_{ig}, g) = \frac{\exp(\boldsymbol{a}_{ig}{}'\boldsymbol{\theta}_{jg} + d_{ig})}{1 + \exp(\boldsymbol{a}_{ig}{}'\boldsymbol{\theta}_{jg} + d_{ig})}, \tag{2.27}$$

where $\boldsymbol{\theta}_{jg}$ is a vector of ability with K dimensions, $\boldsymbol{\theta}_{jg} = \boldsymbol{\theta}_{j1g}, \boldsymbol{\theta}_{j2g}, \ldots, \boldsymbol{\theta}_{jKg})'$, $\boldsymbol{a}_{ig}$ is a vector of item discriminations for latent class $g, \boldsymbol{a}_{ig} = (\boldsymbol{a}_{i1g}, \ldots, \mathbf{a}_{iKg})'$, and d_{ig} is an intercept parameter for latent class g.

Next, Equation 2.26 is extended into the random item context by regarding both items and persons treated as random. To be specific, a vector of ability follows a multivariate normal distribution; a vector of item discriminations also follows a multivariate normal distribution; and an intercept parameter follows a normal distribution as follows:

$$\theta_{jg} \sim MVN\left(\begin{bmatrix} \mu_{\theta_{1g}} \\ \vdots \\ \mu_{\theta_{ka}} \end{bmatrix}, \Sigma_{\theta_g}\right), \tag{2.28}$$

$$a_{ikg} \sim N\left(\mu_{a_{ikg}}, \sigma^2_{a_{ikg}}\right), \tag{2.29}$$

$$d_{ig} \sim N\left(\mu_{d_g}, \sigma^2_{d_g}\right). \tag{2.30}$$

To overcome the difficulties in defining the qualitative differences between latent classes and in interpreting multidimensional latent traits, the external factors are incorporated. As explained in the previous section, one motivation of random item modeling is that this framework more easily and realistically incorporates covariates than does fixed item modeling. External item factors, Y_{iy}, is added to explain the intercept parameter, d_{ig}, the external person factors, X_{jx}, is incorporated to explain the probability of latent class membership, π_g, and these external factors are treated as fixed effects as follows:

$$\mu_{d_g} = d_{0g} + \sum_{y=1}^{Y} d_{yg} Y_{iy}, \tag{2.31}$$

$$logit(\pi_{jg}) = \lambda_{0g} + \sum_{x=1}^{X} \lambda_{xg} X_{ix}. \tag{2.32}$$

Thus, d_{0g} is the mean of class specific intercept parameters for the latent class g when the values of all external item factors are zero, and d_{yg} are the class-specific effects of item factor. Similarly, λ_{0g} is the class-specific intercept when the values of all external person factors are zero, and λ_{xg} is the effect of the external person factors on the probability of the j-th examinee's membership in the latent class g.

Chapter 3

METHODS

The purposes of the study were (1) to investigate the utility of MMixRIM, for providing information related to persons and items in the context of multidimensionality in item response data; and (2) to compare the performance of MMixRIM with the performance of MIRT and MixIRT models in the context of different multidimensional structures. In this chapter, details related to the model specification and estimation of MixRIM are described.

3.1 Model specification and estimation

MMixRIM is a multidimensional extension of the MixIRT model with crossed random variation (i.e., random persons and random items). There are several computer programs for estimating MIRT models or MixIRT models, including TESTFACT (Bock, Schilling, Muraki, Wilson, & Wood, 2003) for MIRT and WinMIRA (von Davier, 2001) for MixIRT. There are limits, however, to estimating IRT models with crossed random based on maximum likelihood estimation (MLE) because of the difficulty of MLE in integrals of a highly dimensional distribution. Moreover, the measurement uncertainty that derives from the nature of random items can easily handle prior distribution with unknown parameters (i.e., mean and variance of a prior distribution) in the Bayesian approach.

In this study, Bayesian estimation using a Markov chain Monte Carlo (MCMC) algorithm is implemented in the OpenBUGS computer software (Spiegelhalter, Thomas, Best, & Lunn, 2007). Drawing statistically consistent samples from a distribution is necessary to effectively create a simulation. Even such a simple function cannot be integrated without numerical

methods, while integrals or summations are vital for calculating the expectation or expected values of distributions. An MCMC algorithm provides an answer to the difficult problem of simulation from a highly dimensional distribution (Gamerman & Lopes, 2006). The basic concept of an MCMC algorithm is that a large number of samples are obtained from a stationary distribution, which yields the desired posterior distribution of interest.

In Bayesian estimation, prior distributions for parameters need to be specified. Although there is flexibility in prior distribution of parameters, this study follows prior distributions that are commonly used. For a MMixRIM, the prior distributions of parameters include the latent class membership, g; the probability of latent class membership, π_g; the multidimensional abilities in a latent class g, θ_{jg}; the slope parameter of an item i on a dimension k in a latent class g, a_{ikg}; the intercept parameter of item of an item i in a latent class g, d_{ig}, and latent class-specific means and variances for these parameters as follows:

$$g \sim Multinomial\left(1, (\pi_1, \ldots, \pi_G)\right), (g = 1, \ldots, G), \tag{3.1}$$

$$(\pi_1, \ldots, \pi_G) \sim Dirichlet\left(.5, \ldots, .5\right), \tag{3.2}$$

$$\boldsymbol{\theta}_{jg} \sim MVN\left(\begin{bmatrix} \mu_{\theta_{1g}} \\ \vdots \\ \mu_{\theta_{ka}} \end{bmatrix}, \boldsymbol{\Sigma}_{\theta_g}\right), (j = 1, \ldots, N, k = 1, \ldots, K, g = 1, \ldots, G), \tag{3.3}$$

$$a_{ijk} \sim N\left(\mu_{a_{kg}}, \sigma^2_{a_{kg}}\right) \text{and } a_{ijk} > 0, (i = 1, \cdots, n, k = 1, \cdots, K, g = 1, \cdots, G), \tag{3.4}$$

$$a_{ijk} = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{if} \quad i \neq k \end{cases}, \ (i = 1, \ldots, K, k = 1, \ldots, K, g = 1, \ldots, G), \tag{3.5}$$

$$\mu_{a_{kg}} \sim N\left(0, 1\right) \text{and } \mu_{a_{kg}} > 0, (k = 1, \ldots, K, g = i, \ldots, G), \tag{3.6}$$

$$\sigma^2_{a_{kg}} \sim Inverse - \ Gamma\left(1, 1\right), (g = i, \ldots, G), \tag{3.7}$$

$$d_{ig} \sim N\left(\mu_{d_g}, \sigma^2_{d_g}\right), (i = 1, \ldots, N, g = i, \ldots, G), \tag{3.8}$$

$$d_{ig} = 0, (i = 1, \ldots, K, g = 1, \ldots, G), \tag{3.9}$$

$$\mu_{d_g} \sim N\left(0, 1\right), (g = i, \ldots, G), \tag{3.10}$$

$$\sigma^2_{d_g} \sim Gamma\,(1,1)\,,(g=i,\ldots,G)\,. \tag{3.11}$$

A noninformative prior, Equation 3.2, is specified for the probability of latent class membership (e.g., Choi & Wilson, 2015; Li, Cohen, Kim, & Cho, 2009). Because of the metric indeterminacy problem in IRT, it is necessary to fix a specific metric in order to be able to compare latent classes (Choi, 2014). Choi (2014) found item anchoring to be more effective than either person centering or item centering in MixIRT models. In this study, for the metric indeterminacy, the mean of ability parameters in Equation 3.3 were fixed at zero for each dimension and latent class, and the variance and covariance matrix sets as identify matrix. As shown in Equation 3.4, the item discriminations are generally expected to be positive values. According to Béguin and Glas (2001), a normal distribution with zero mean for item discriminations led to better performance than when the mean of the prior distribution was larger than zero. For the simulation study, normal hyper priors for the item discriminations were assumed with means of zero as shown in Equation 3.6.

Equations 3.5 and 3.9 are the constraints used to solve the rotational indeterminacy problem. Two types of approaches are used for the rotational indeterminacy problem. One approach is that that a_{ik} is zero when $i = 1,\ldots,K-1$, and $k = i+1,\ldots,K$. This approach is used in NOHARM (Fraser, 1988) and BMIRT (Yao, 2003). The other approach suggested by Béguin and Glas (2001) is that a_{ik} is fixed to one if $i = k$. In addition to this constraint, a_{ik} is fixed to zero if $i \ldots k$, and d_i is also fixed to zero when $i = 1,\ldots,K$. In this study, the later approach was applied to solve the rotational indeterminacy problem as in Béguin and Glas (2001) and Kang (2006), because this approach results in the positivity of the item slope parameters.

3.2 Monitoring convergence

An MCMC algorithm is convenient and flexible for fitting complicated statistical models, however, it is difficult to answer the question, at "what point is it reasonable to believe that

the samples are truly representative of the underlying stationary distribution of the Markov chain?" (Cowles & Carlin, 1996, p. 883). Sahlin (2011) has noted that an MCMC algorithm is commonly initiated at a random point, when uninformative prior distributions of parameters are applied. This starting point might or might not be located far from the high density region of the posterior distribution. Consequently, the sample values at the beginning of the simulation might not be close to the true distribution. These invalid samples at the beginning of the simulation are not a big problem, because these initial parts of the chain, which are known as the burn-in period, can be discarded. Determining the numbers of iterations for the burn-in and post-burn-in period, however, is critical so as to minimize the effects of the burn-in period on the samples from the converged part of the chain.

A number of methods for checking convergence have been proposed, such as Brooks, Gelman, and Rubin's (1992) method, Heidelberger and Welch's (1983) method, Raftery-Lewis' (1992) method, and Geweke's (1992) method. Brooks, Gelman, and Rubin's (1992) method is for monitoring convergence of multiple chains, and other three methods are appropriate to monitor convergence of a single chain. These methods are available in the R package CODA (Plummer, Best, Cowles, & Vines, 2006). Heidelberger and Welch's (1983) convergence diagnostic is one popular method. This method assesses the stationary of a single chain. The first step is to assess the stationarity of the initial check-point, the first 10% of a chain. If it passes the stationarity test, the whole chain is applied for the second step. If it fails the stationarity test, this initial check-point is discarded and an additional 10% of the chain is tested until it passes. The procedure is continued until the 10% of chain passes the stationarity test or more than 50% of the chain has been discarded. In the second step, the confidence interval of for the mean of each estimated parameter is generated and tested to determine whether this confidence interval meets a specific accuracy criterion. This step is called the halfwidth test.

In addition to these methods, graphical methods are also a simple way to check convergence of a chain. OpenBUGS provides several kinds of plots for diagnostic convergence,

such as trace plots and autocorrelations. A trace plot is a time series plot that shows the horizontal pattern of sampled parameters across all iterations to be monitored. A stable pattern in the trace plot implies convergence of a chain. A plot of autocorrelations between sampled parameters is another graphical method for checking convergence. High autocorrelations imply a slow convergence, and vice versa. Therefore, if autocorrelations approach zero, the chain can be considered as converged.

A third approach is to examine the Monte Carlo (MC) error for each parameter. When the MC error is less than about 5% of the standard deviation for parameters drawn during the post-burn-in period, the chain is considered to have converged for the parameter. In this study, the percentage of the items that pass Heidelberger and Welch's stationarity test for each replication was used to monitor convergence of the chain. Additionally, the MC error-standard deviation ratio was monitored as a check on convergence.

3.3 Model comparison

Determining the number of dimensions in MIRT models or determining the number of latent classes in MixIRT models is an important step to develop valid and accurate interpretation. For the model comparison of nested models, such as MIRT models (e.g., 1-dimensional M2PL model and 2-dimensional M2PL model), the likelihood ratio test is commonly used. For the model comparison of non-nested models, such as MixIRT models, (e.g., 2-class Mix2PL model and 3-class Mix2PL model), information criteria can be used. In this study, information criteria are used to determine the number of dimensions or latent classes.

Akaike's (1974) information criterion (AIC) and Bayesian information criterion (BIC; Schwarz, 1978) are the most popular information criteria for model selection with IRT models. As summarized by Cohen and Cho (2015), these information criteria were commonly used in several studies for IRT model selection. Both AIC and BIC ban be calculated based on the penalized log-likelihood value by a function of the sample size and the number

of parameters. The model with the lowest value of AIC or BIC is considered the appropriate model. AIC and BIC are computed with the following equations:

$$AIC = -2\log L + 2d, \tag{3.12}$$

$$BIC = -2\log L + d \times \ln(N), \tag{3.13}$$

where L is the maximum value of the likelihood function, d is the number of estimated parameters, and N is the sample size.

As noted previously, this simulation study implemented MCMC algorithm instead of MME. The deviance based on MME, $-2 \times logL$, in equations for AIC and BIC is replaced by the posterior mean of the deviance, $\overline{d(\xi)}$, (Congdon, 2003) where the ξ are estimated parameters. Therefore, AIC and BIC in MCMC algorithm are computed as $\overline{d(\xi)} + 2d$, and $\overline{d(\xi)} + d \times ln(N)$, respectively (Li et al., 2009).

Although AIC and BIC are generally used as criteria for model selection, sometimes, the results based on these information criteria do not agree with each other. Several studies (e.g., Bolt & Johnson, 2009; Dziak, Coffman, Lanza, & Li, 2012; Li et al., 2009) have shown than AIC tends to select a more complex model, whereas BIC tends to select a simpler model. When the sample size is small, there is a tendency to reduce standard error caused by a relatively large number of estimates compared to a sample size. Consequently, under-fitting would be a more common error for the small sample size. On the other hand, as the model with enough parameters to explain the relationship between variables is preferred, over-fitting is a more likely error, when the sample size is large enough. Therefore, AIC seems better with a small sample size, while BIC seems better with large sample size. That is, the performance of information criteria depends on the sample size and the nature of a model (Dziak et al., 2012).

There are some extended versions of information criterion by different penalty function. Bozdogan (1987) modified AIC by adding the sample size to the penalty function for more consistent performance of AIC, and called it consistency AIC (CAIC). For a small sample size

related to the number of model parameters, a small sample AIC (AIC_c; Sugiura, 1978) and an adjusted BIC (ABIC; Sclove, 1987) are suggested. CAIC, AIC_c, and ABIC are computed with the equations below, respectively:

$$CAIC = -2logL + d \times (ln(N) + 1), \tag{3.14}$$

$$\begin{aligned} AIC_c &= -2\log L + 2d + \frac{2d(d+1)}{N-d-1} \\ &= AIC + \frac{2d(d+1)}{N-d-1}, \end{aligned} \tag{3.15}$$

$$ABIC = -2\log L + d \times \ln\left(\frac{N+2}{24}\right). \tag{3.16}$$

CAIC, AIC_c, and ABIC are also computed based on the maximum value of the likelihood function, the number of estimated parameters, and the sample size as AIC and BIC. Therefore, computing these indices and applying these for the model comparisons would be easier than using other indices, such as the deviance information criterions (DIC; Spiegelhalter, Best, Carlin, and van der Linde, 2002). Moreover, several types of IRT models are compared in the simulation study, and some models have a relatively larger number of model parameters compared to the sample size than other models. Consequently, in addition to AIC and BIC, additional three information criteria adjusted to the small sample size (i.e., CAIC, AIC_c, and ABIC) are used for the model comparisons to explore the effect of the relative number of model parameters to the sample size in the simulation study.

Chapter 4

EMPIRICAL STUDY

One empirical data set was analyzed to illustrate the issue with the analyses of multidimensional item response data. Test data from a fractions computation test designed to assess middle school mathematics teachers' understandings of rational numbers were used (Bradshaw, Izsák, Templin & Jacobson, 2014). The data set contained 982 middle school mathematics teachers' responses to 27 items. The test measured four attributes: Referent Units, Partitioning and Iterating, Appropriateness, and Multiplicative Comparison. The test contained a total of 27 items consisting of two multiple choice items with three options, 11 multiple choice items with four options, three multiple choice items with five options, two multiple choice items with six options, and nine short answer items. All items were scored dichotomously, that is, zero for an incorrect answer and one for a correct answer.

4.1 Dimensionality assessment

The first step of the empirical study was a dimensionality assessment. In this study, exploratory approaches were applied. First, a principal component analysis (PCA) was conducted as implemented in the SAS computer software, version 9.4. The plots of eigenvalues and the proportions of explained variances are presented in Figure 4.1. The right panel in Figure 4.1 is the plot of eigenvalues against the factor number. It shows that the eigenvalues of the first eight factors were all greater than one. The left panel in Figure 4.1 is the proportion of explained variance by factors and shows that the explained variance by the first factor was less than 20%. According to Reckase (1979), a data set can be considered

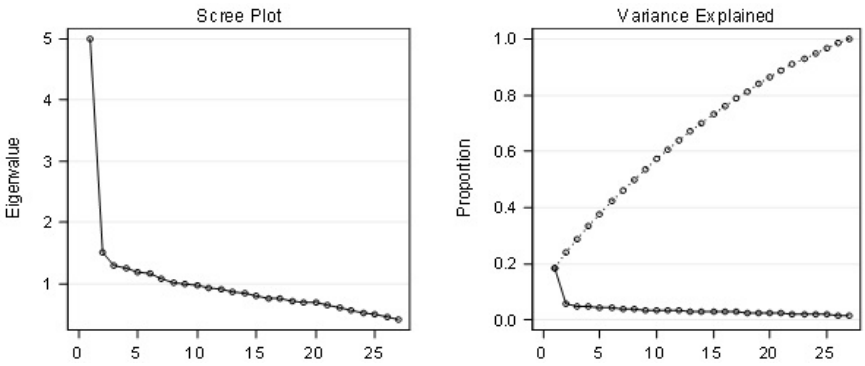

Figure 4.1: Plots for eigenvalues and proportions of explained variances of PCA

unidimensional data, when the first factor explained at least 20% of the variance. Therefore, these plots imply that the data were not unidimensional.

Additionally, exploratory factor analysis (EFA) was conducted using the computer software *Mplus* (Muthén & Muthén, 1998-2012). Also, a dimensionality analysis was conducted using the Dimensionality Evaluation to Enumerate Contributing Traits (DETECT) index implemented in the computer software DIMPACK (Stout, 2006). Table 4.1 shows model goodness of fit statistics of the EFA with one to five factors. In general, the model has a good fit when the Root Mean Square Error of Approximation (RMSEA) is smaller than .05, the Tucker-Lewis Index (TLI) is larger than .90, and the Comparative Fit Index (CFI) is larger than .95.

Table 4.1: Model Goodness of Fit Statistics for the EFA Models

Fit statistics	1-Factor	2-Factor	3-Factor	4-Factor	5-Factor
CFI	.88	.93	.96	.98	.99
TLI	.87	.92	.95	.97	.98
RMSEA	.06	.04	.04	.03	.02
Decreased RMSEA(%)	-	.22	.19	.29	.12

Based on these fit statistics, the EFA models with more than two factors displayed a good fit. The estimation of the EFA model with six factors did not converge, so it is not reported here. Tate (2003) used the proportion of the decreased RMSEA as a criterion for determining the number of factors for an EFA. When the amount of decrease in RMSEA from adding one factor is larger than 10%, the additional factor is considered to provide a better model. Based on Tate's (2003) rule, the EFA model with five factors was most appropriate for the empirical data. The result from DIMPACK (Stout, 2006) indicated the number of dimensions to maximize the DETECT index was five. To sum up, the results from both analyses showed 5-factors for the data, although the factor structures were not exactly the same. The Q-matrix, which is a matrix that indicates which attributes are measured by each item by coding zero or one, and factor structures are presented in Table 4.2.

According to the results from DIMPACK, 12 items were clustered on the first dimension. Most items loading on the first dimension measured Partitioning and Iterating (PI) and Referent Units (RU) attributes. Most items belonged to either the second or the third dimension that measured RU and MC attributes. Items belonging to the fourth dimension were not cleanly classified as belonging to a particular type of attribute, and items clustered into the fifth dimension measured the APP attribute. Similarly, most items mainly measured the first dimension based on the results of the EFA, and these items measure the RU and PI attributes. Most items in the second dimension measured the Appropriateness (APP) attribute, and most items in the third dimension measured the Multiplicative Comparison (MC) attribute. Most items in the fifth dimension measure the PI. A specific pattern of structure with regard to a type of item was not displayed.

4.2 Comparisons of IRT models

Although the empirical data were determined to be multidimensional, the number and structure of the dimensions was not clearly indicated. To figure out the structure of the dimensions

Table 4.2: Factor Structures

Item	Item type	Q-matrix				DIMPACK					EFA in Mplus				
		RU	PI	APP	MC	D1	D2	D3	D4	D5	D1	D2	D3	D4	D5
Q1	MC4	1	0	0	0	1	0	0	0	0	0.54*	0.03	0.03	0.08	0.02
Q2	MC4	0	0	1	0	1	0	0	0	0	0.30*	0.08	0.07	−0.06	0.06
Q3	SA	0	1	0	0	1	0	0	0	0	0.62*	0.00	−0.14	−0.04	0.11
Q4	SA	1	0	0	0	0	1	0	0	0	0.22*	−0.07	0.07	0.17*	−0.07
Q5	MC4	1	0	0	0	0	0	1	0	0	0.40*	0.13	0.01	0.05	−0.08
Q6	SA	0	1	0	0	1	0	0	0	0	0.72*	−0.06	−0.12	0.00	0.01
Q7	MC4	1	0	0	0	0	0	0	1	0	−0.05	0.04	−0.20	0.73*	0.10
Q8	MC4	0	0	1	0	0	0	0	0	1	0.15	0.75*	0.07	0.03	0.02
Q9	MC4	0	0	1	0	0	0	0	0	1	−0.03	0.83*	−0.05	−0.03	−0.02
Q10	MC4	0	0	1	0	1	0	0	0	0	0.12	0.14*	0.05	−0.01	0.09
Q11	MC4	0	0	1	0	0	0	0	0	1	0.03	0.32*	0.05	0.19*	0.20*
Q12	SA	1	0	0	0	0	1	0	0	0	0.29*	−0.21 *	0.05	0.14	−0.04
Q13	SA	0	0	0	1	0	1	0	0	0	−0.08	0.01	1.07*	−0.04	0.02
Q14	SA	1	0	0	1	0	1	0	0	0	0.19	−0.02	0.72*	0.03	0.05
Q15	SA	1	0	0	1	0	1	0	0	0	0.09	0.04	0.73*	0.09	−0.01
Q16	SA	1	0	0	0	0	0	1	0	0	0.23*	0.19*	0.03	0.15*	−0.07
Q17	MC4	1	0	0	0	0	0	0	1	0	0.31*	0.02	0.02	0.40*	0.00
Q18	MC3	0	1	0	1	0	0	0	1	0	0.03	0.07	0.05	0.15*	0.07
Q19	MC4	1	1	0	0	0	0	1	0	0	0.36*	0.11	0.05	0.04	−0.04
Q20	MC5	0	1	0	1	1	0	0	0	0	0.13	0.08	0.15*	−0.01	0.61*
Q21	MC5	0	1	0	0	1	0	0	0	0	0.06	−0.04	−0.03	0.00	0.87*
Q22	MC5	0	1	0	0	1	0	0	0	0	−0.04	0.00	0.01	0.05	0.91*
Q23	MC3	1	0	0	0	0	0	0	1	0	0.17	−0.13	0.07	0.53*	−0.04
Q24	MC6	1	1	0	0	1	0	0	0	0	0.48*	−0.01	0.06	−0.01	0.16*
Q25	MC6	1	1	0	0	1	0	0	0	0	0.45*	0.11	0.07	−0.02	0.08
Q26	MC4	1	0	0	0	1	0	0	0	0	0.42*	0.04	−0.03	0.12	0.07
Q27	SA	1	1	0	0	1	0	0	0	0	0.49*	0.06	0.08	−0.01	0.21*

Note. * $p < .05$; MC3 = Multiple choice item with three options; MC4 = Multiple choice item with four options; MC5 = Multiple choice item with five options; MC6 = Multiple choice item with six options; SA = Short answer item; RU= Referent Units; PI = Partitioning and Iterating; APP = Appropriateness; MC = Multiplicative Comparison.

for the empirical data, an exploratory MixIRT model analysis and MIRT model analysis were conducted. For the exploratory MixIRT model analysis, MRM with one to six classes were compared. Alexeev, Templin, and Cohen (2011) observed that the over-extraction of latent classes could occur when a mixture Rasch model was applied to data generated to fit a two-parameter IRT model. It is possible that a misspecified model might cause over-extracted latent classes in a MixIRT analysis. Mix2PL models with one to five classes were also compared to avoid the incorrect selection of a particular model.

The MixIRT models were estimated using a MCMC algorithm as implemented in the OpenBUGS software. For model selection, three information criteria, AIC, BIC, and CAIC, were used. These are summarized in Table 4.3.

Based on the information criteria, a five-class model was the best fitting model among the six MRM models. Among the Mix2PL models, a four-class model was the best fitting model based on AIC, but a three-class model was the best fitting model based on BIC and CAIC. The number of latent classes for the best fitting model decreased as the number of parameters increased. This result agreed with the finding in Alexeev et al. (2011). The estimated item difficulty parameters of the 5C-MRM are summarized in Table 4.4.

Latent class sizes for the 5C-MRM were 2%, 18%, 11%, 29%, and 40% for Class 1 to Class 5, respectively. Class 1 and Class 2 did not have obvious patterns of estimated item difficulty parameters. For Class 3, three items (i.e., Q8, Q9, and Q11) had relatively high item difficulties. These items measured the APP attribute. For Class 4, item difficulties of Q13, Q14, and Q15 were significantly higher than other classes. These items measured the MC attribute. For Class 5, three items (i.e., Q20, Q21, and Q22) had lower item difficulties; these items measured the PI attribute. Based on these patterns of item difficulties, the characteristics for latent classes are that (1) Class 1 and Class 2 are the average levels, and these classes might be over-extracted; (2) Class 3 lacks in the APP attribute; (3) Class 4 lacks in the MC attribute; and (4) Class 5 is strong in the PI attribute.

Table 4.3: Information Criterion Indices for Exploratory Analysis of Empirical Data

Model	NC	NPAR	log$L$	Deviance	AIC	BIC	CAIC
1CMRM	1	27	$-14,320$	28,640	28,694	28,826.02	28,853.02
2CMRM	2	56	$-13,760$	27,520	27,632	27,905.82	27,961.82
3CMRM	3	85	$-13,360$	26,720	26,890	27,305.62	27,390.62
4CMRM	4	114	$-13,270$	26,540	26,768	27,325.41	27,439.41
5CMRM	5	143	$-13,110$	26,220	**26,506**	**27,205.21**	**27,348.21**
6CMRM	6	172	$-13,350$	26,700	27,044	27,885.01	28,057.01
1CMix2PL	1	54	$-14,010$	28,020	28,128	28,392.04	28,446.04
2CMix2PL	2	110	$-13,460$	26,920	27,140	27,677.86	27,787.86
3CMix2PL	3	166	$-13,160$	26,320	26,652	**27,463.67**	**27,629.67**
4CMix2PL	4	222	$-13,060$	26,120	**26,564**	27,649.49	27,871.49
5CMix2PL	5	278	$-13,070$	26,140	26,696	28,055.31	28,333.31
2DM2PL	1	80	$-13,430$	26,860	27,070	27411.17	27,491.17
2C2DMMixRIM	2	175	$-12,850$	25700	**26,050**	**26,905.68**	**27,080.68**

Note. NC = Number of latent classes; NPAR = Number of estimated parameters; The smallest model information criterion index is bold.

Table 4.4: Estimated Item Difficulty Parameters of the 5C-MRM

Item	Class 1	Class 2	Class 3	Class 4	Class 5
Q1	−0.60	0.12	−0.03	0.09	0.28
Q2	−0.45	−2.26	−2.20	−1.99	−1.47
Q3	0.11	1.18	0.86	0.39	1.01
Q4	−0.10	0.41	−0.28	0.50	1.67
Q5	1.39	0.74	0.49	0.64	1.33
Q6	0.82	2.23	2.16	1.88	2.55
Q7	−0.19	0.29	−1.60	−0.45	0.80
Q8	0.07	−1.90	0.84	−2.17	−1.78
Q9	0.89	−1.96	0.58	−2.59	−1.16
Q10	−0.35	−1.15	−1.50	−1.65	−0.79
Q11	0.26	−0.55	0.61	−0.76	−0.36
Q12	−0.75	0.78	−0.78	0.62	1.44
Q13	−1.78	−3.38	−0.26	0.41	−2.86
Q14	−0.79	−0.61	1.68	3.10	−0.45
Q15	0.01	−0.28	1.84	3.11	0.17
Q16	0.41	−0.12	−0.19	−0.24	0.74
Q17	0.10	0.30	−0.49	0.26	0.85
Q18	0.62	0.01	−1.01	−0.31	0.65
Q19	0.87	1.39	1.43	1.11	1.75
Q20	0.33	1.42	1.01	0.38	−1.01
Q21	−0.47	1.00	−0.76	−1.04	−2.60
Q22	−0.37	0.75	−0.68	−1.10	−2.84
Q23	−0.74	−0.17	−1.40	0.41	0.79
Q24	−0.30	1.13	1.07	0.67	0.69
Q25	−0.64	−0.12	−0.21	−0.34	−0.17
Q26	−0.93	0.92	−0.18	0.63	0.97
Q27	−1.08	0.07	0.20	−0.14	−0.45

Although the attributes measured by the assessment were used to define the characteristics of each latent class, the attributes might not provide enough information to figure out the qualitative differences between latent classes. The items showed significant differences between latent classes that can be more clearly categorized into three item groups. The Q8, Q9, Q10, and Q11 are the first group, Q13, Q14, and Q15 are the second group, and Q20, Q21, and Q22 are the third group. Some of these items measure more than one attribute. For example, Q14 measured both RU and MC attributes. Therefore, the categorization based on the 5C-MRM might not be appropriate for understanding the structure of dimensions of the empirical data.

The estimated item parameters from the 3CMix2PL are summarized in Table 4.5 and Figure 4.2. The class sizes for the three latent classes were 24%, 37%, and 39%, respectively. Overall, the estimated item discriminations are within a general range, except two items (i.e., Q7 and Q23) for Class 2, and there is no specific pattern to characterize the latent classes. As for the estimated item difficulties, only six items (i.e., Q13, Q14, Q15, Q20, Q21, and Q22) have obviously different item difficulties between the latent classes. The item difficulties of Q13, Q14, and Q15 for Class 2 are significantly higher than other classes. Q20, Q21, and Q22 have higher item difficulties for Class 1. Additionally, these items were overlapped with items that showed significant differences between latent classes based on 5C-MRM.

For an exploratory MIRT model analysis, M2PL models with two to six dimensions were compared. These models were estimated using the MCMC algorithm as implemented in the OpenBUGS software. Although AIC, BIC, and CAIC of the M2PL models gradually decreased with each additional dimension, the M2PL models with more than three-dimensions showed problems with convergence. Thus, the estimated item parameters of the 2D-M2PL are reported in Table 4.6. In this table, the angle indicates which dimension is mainly measured by each item. This value is the angle from the axis of the first dimension. Therefore, an item mainly measures the first dimension, when the angle is almost zero, and an item mainly measures the second dimension when the angle is closest to 90°. In

Table 4.5: Estimated Item Parameters of the 3C-Mix2PL

Item	Item discrimination			Item difficulty		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Q1	1.18	0.81	1.19	0.15	−0.11	0.18
Q2	0.78	0.91	0.71	−2.28	−2.61	−2.26
Q3	1.22	1.07	0.81	1.11	0.07	1.01
Q4	0.59	0.52	0.61	1.02	1.08	2.15
Q5	0.66	0.74	0.91	1.50	0.79	1.29
Q6	1.43	1.57	1.03	1.86	0.92	2.47
Q7	1.58	0.16	0.75	0.28	0.19	0.94
Q8	0.88	1.72	1.74	−1.76	−1.83	−1.29
Q9	0.44	1.37	0.84	−2.24	−2.08	−1.73
Q10	0.57	0.62	0.53	−1.32	−2.24	−1.87
Q11	0.70	0.94	0.69	−0.44	−0.77	−0.78
Q12	0.80	0.35	0.61	1.05	1.73	1.89
Q13	2.10	0.49	2.12	2.50	0.78	−1.47
Q14	0.89	1.56	1.58	−0.85	1.95	−0.29
Q15	0.72	2.09	1.23	−0.40	1.98	0.14
Q16	0.56	0.73	0.63	0.32	−0.31	0.82
Q17	1.46	0.56	1.37	0.28	0.36	0.76
Q18	0.42	0.34	0.32	0.76	0.38	0.68
Q19	0.76	0.77	0.97	2.02	1.48	1.64
Q20	2.00	1.55	1.03	0.76	−0.05	−1.22
Q21	1.55	1.25	1.39	0.37	−1.12	−2.46
Q22	1.53	1.29	1.72	0.31	−1.19	−2.34
Q23	1.09	0.28	1.06	−0.08	1.25	0.82
Q24	1.00	1.04	1.04	1.05	0.48	0.52
Q25	1.03	0.87	1.07	−0.06	−0.57	−0.32
Q26	1.41	0.50	1.22	0.93	1.14	0.79
Q27	1.43	1.22	1.17	−0.08	−0.52	−0.48

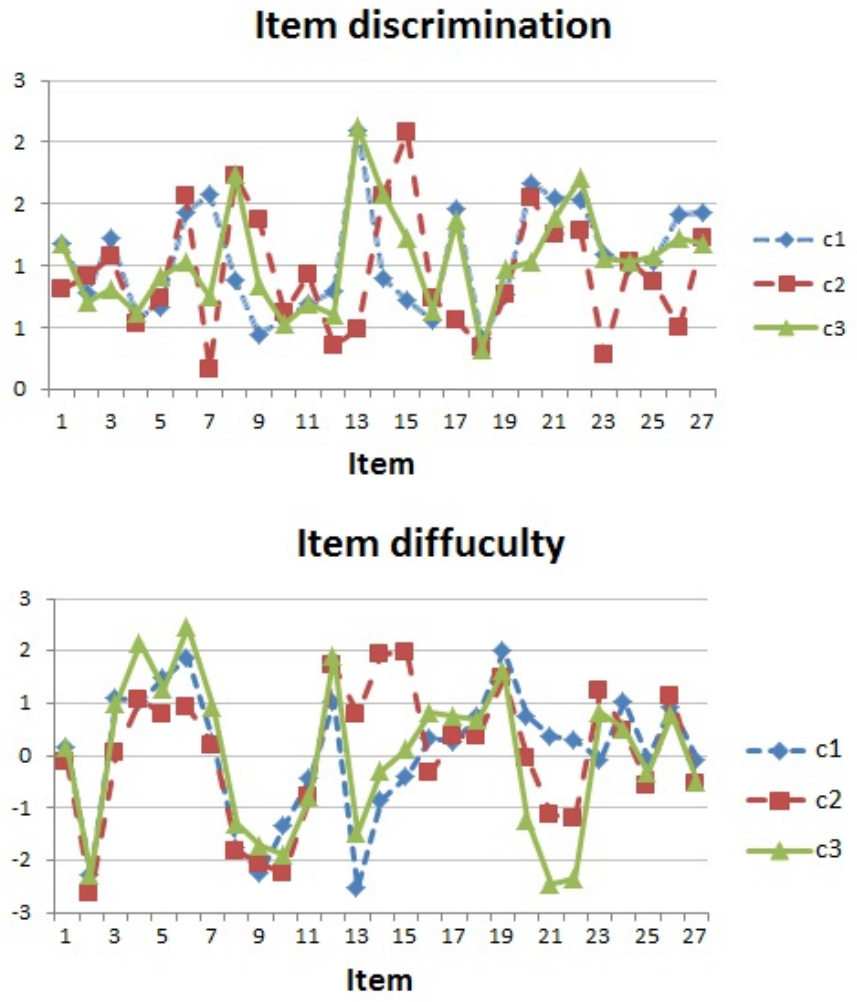

Figure 4.2: Plots for item parameters for each latent class of 3CMix2PL.

Figure 4.3, for example, items in Group A mainly measured Dimension 1, items in Group B mainly measured Dimension 2, and items in Group C measure both Dimensions 1 and 2.

As shown in Table 4.6, angles of most of items were smaller than 30°. Only four items among the 27 items had angles, which were larger than 30°. This means that all items in the test mainly measured the first dimension. The results from the three models (i.e., the 5C-MRM, 3C-Mix2PL, and 2D-M2PL) were inconsistent with each other, and thus were not sufficient to understand the dimensionality of the empirical data. An alternative model, which

can provide information with respect to persons and items, was applied to analyze the data. This was a two-class and two-dimensional MMixRIM which was estimated in OpenBUGS software.

Table 4.6: Estimated Item Parameters of the 2D-M2PL

Item	a_1	a_2	d	MDISC	MDIFF	Angle (°)
Q1	1.03	0.00	−0.49	1.03	−0.48	0.00
Q2	0.72	−0.03	1.41	0.72	1.96	2.45
Q3	0.91	0.29	−1.15	0.95	−1.20	17.85
Q4	0.39	−0.18	−1.08	0.43	−2.53	24.71
Q5	0.69	−0.07	−1.23	0.70	−1.76	5.57
Q6	0.97	0.17	−2.64	0.99	−2.67	9.74
Q7	0.39	0.07	−0.37	0.39	−0.95	10.24
Q8	1.40	−0.04	1.38	1.40	0.99	1.79
Q9	0.72	0.07	1.06	0.72	1.47	5.27
Q10	0.52	0.03	0.75	0.52	1.46	2.90
Q11	0.91	0.08	0.08	0.91	0.09	5.09
Q12	0.34	−0.13	−1.01	0.36	−2.77	20.76
Q13	2.51	−1.55	1.53	2.95	0.52	31.68
Q14	2.45	−1.32	−0.90	2.79	−0.32	28.36
Q15	2.04	−1.24	−1.36	2.38	−0.57	31.40
Q16	0.57	−0.11	−0.50	0.58	−0.87	10.87
Q17	0.86	−0.13	−0.76	0.87	−0.88	8.55
Q18	0.36	−0.02	−0.39	0.36	−1.06	2.85
Q19	0.74	−0.06	−1.71	0.74	−2.30	4.61
Q20	2.00	0.67	−0.48	2.11	−0.23	18.41
Q21	2.17	1.49	0.97	2.63	0.37	34.43
Q22	2.21	1.40	1.09	2.61	0.42	32.45
Q23	0.59	−0.23	−0.54	0.63	−0.85	21.26
Q24	1.14	0.11	−1.13	1.14	−0.98	5.28
Q25	1.07	0.00	−0.12	1.07	−0.11	0.26
Q26	0.86	0.08	−1.08	0.86	−1.25	5.18
Q27	1.46	0.14	−0.13	1.47	−0.09	5.44

Note. MDIFF = Multidimensional item difficulty; MDISC = Multidimensional item discrimination

The estimated item parameters of the 2C2D-MMixRIM are reported in Table 4.7. Because the item slope of the first item on the second dimension was fixed at zero for the

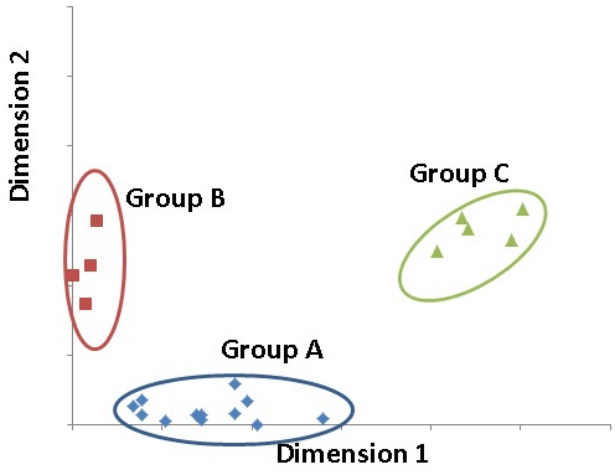

Figure 4.3: Three item groups located in two-dimension space.

rotational indeterminacy problem, the angles of the first item for both classes were zero. Whereas the angles of most items were close to the first dimension, the angles of some items significantly differed between latent classes. Q7, Q13, Q14, Q15, and Q23 were very close to the second dimension, that is, the angles of these items were close to 90°, for Class 1, and these items were close to the first dimension for Class 2. On the other hand, Q21 and Q22 were close to the first dimension for Class 1, but these items were close to the second dimension for Class 2. The class sizes were 50% for the two latent classes, and the AIC and BIC of the 2C2D-MMixRIM was the lowest, as reported in Table 4.3. The locations of items in the two-dimension space based on two different models (i.e., the 2D-M2PL and 2C2D-MMixRIM) are compared in Figure 4.4. As described previously, most of the items are located near the axis of the first dimension, while some of the items are located near the axis of the second dimension.

Table 4.7: Estimated Item Parameters of the 2C2D-MMixRIM

Item	Class 1			Class 2			MDISC		MDIFF		Angle (°)	
	a_1	a_2	d	a_1	a_2	d	Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
Q1	1.01	0.00	−0.88	1.12	0.00	0.42	1.01	1.12	0.87	−0.37	0.00	0.00
Q2	0.50	−0.42	1.19	0.76	0.16	1.79	0.65	0.78	−1.82	−2.30	40.19	12.09
Q3	1.39	−0.01	−1.57	0.73	−0.01	−0.51	1.39	0.73	1.13	0.70	0.56	0.84
Q4	0.18	−0.27	−1.19	0.50	−0.10	−0.88	0.33	0.51	3.61	1.72	56.20	10.94
Q5	0.51	−0.38	−1.47	0.68	0.00	−0.84	0.64	0.68	2.30	1.23	36.58	0.14
Q6	1.70	−0.02	−3.30	0.97	−0.35	−1.82	1.70	1.04	1.94	1.76	0.56	19.80
Q7	0.05	0.57	−0.53	0.94	0.46	−0.24	0.58	1.05	0.92	0.22	84.54	25.92
Q8	0.83	−1.8	1.16	0.75	0.93	1.35	1.98	1.20	−0.59	−1.13	65.27	51.06
Q9	0.36	−1.13	1.15	0.11	0.67	0.55	1.19	0.68	−0.97	−0.81	72.25	80.92
Q10	0.45	−0.39	0.57	0.30	0.16	0.97	0.60	0.34	−0.96	−2.83	41.20	27.29
Q11	0.39	−0.73	−0.22	0.70	0.77	0.19	0.83	1.04	0.27	−0.18	62.09	47.87
Q12	0.27	0.09	−1.15	0.63	−0.17	−0.73	0.28	0.65	4.03	1.12	18.05	15.44
Q13	0.99	−2.76	0.48	2.78	0.58	3.13	2.93	2.84	−0.16	−1.10	70.23	11.76
Q14	1.12	−2.07	−2.22	2.92	0.39	0.63	2.35	2.94	0.94	−0.21	61.73	7.58
Q15	0.77	−1.79	−2.57	2.92	0.11	0.17	1.95	2.93	1.32	−0.06	66.68	2.16
Q16	0.34	−0.35	−0.67	0.58	−0.06	−0.28	0.48	0.58	1.40	0.48	45.96	6.34
Q17	0.42	−0.18	−0.99	1.23	0.09	−0.39	0.46	1.23	2.17	0.31	22.75	4.23
Q18	0.34	−0.14	−0.53	0.37	0.08	−0.18	0.37	0.37	1.44	0.48	22.86	12.27
Q19	0.76	−0.44	−2.04	0.69	−0.07	−1.14	0.88	0.69	2.33	1.65	30.03	6.01
Q20	1.88	−0.80	−1.48	1.42	1.51	0.52	2.04	2.07	0.72	−0.25	23.02	46.74
Q21	3.19	−0.13	−0.07	1.07	2.15	1.69	3.19	2.40	0.02	−0.70	2.40	63.45
Q22	3.18	−0.34	0.06	1.10	2.18	1.76	3.20	2.44	−0.02	−0.72	6.01	63.19
Q23	0.08	0.21	−0.76	1.58	0.16	0.04	0.22	1.59	3.43	−0.02	69.97	5.70
Q24	1.25	−0.46	−1.64	0.93	0.19	−0.22	1.33	0.95	1.23	0.24	20.24	11.29
Q25	1.04	−0.55	−0.54	0.99	0.05	0.65	1.18	0.99	0.46	−0.66	27.88	3.07
Q26	0.76	−0.07	−1.36	1.04	0.11	−0.70	0.76	1.05	1.79	0.67	4.93	6.26
Q27	1.44	−0.66	−0.71	1.27	0.31	0.74	1.58	1.30	0.45	−0.57	24.60	13.56

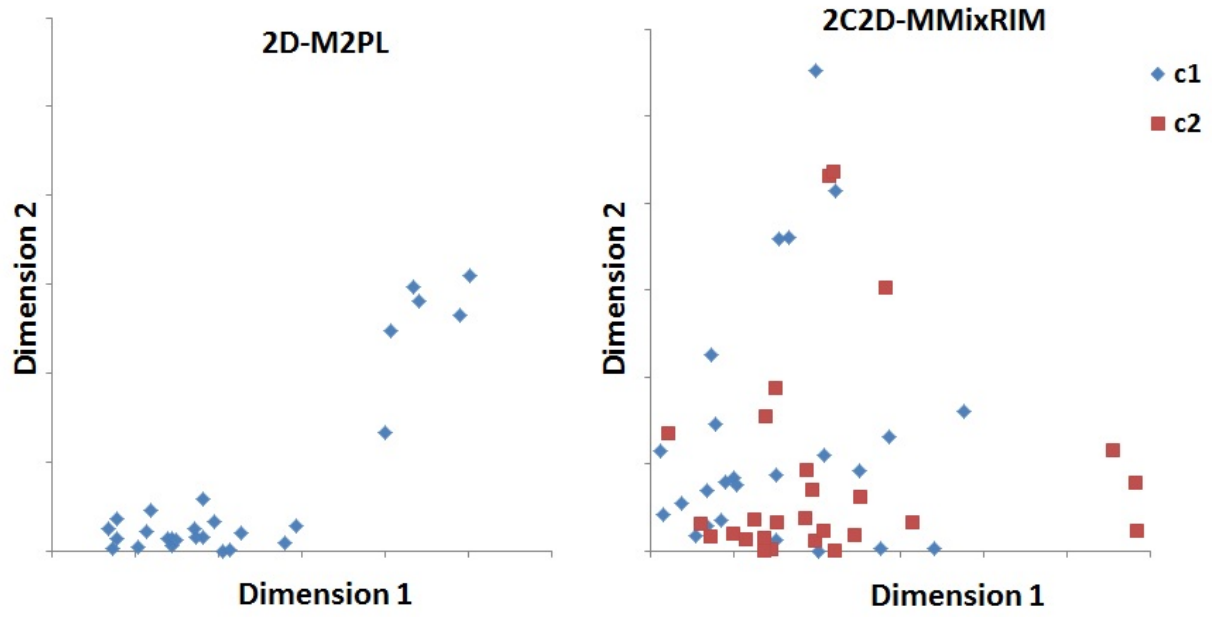

Figure 4.4: Plots of items located in two-dimension space.

4.3 Summary and conclusions

In this chapter, results are reported for the empirical data analyses. In advance of the IRT analyses, the dimensionality of the empirical data was assessed. Three approaches were applied for the dimensionality assessment: the PCA, EFA, and DETECT. The results of these methods indicated the multidimensionality of the empirical data, although the number of dimensions and the structure were not consistent.

First, the MixIRT models were used to explore the structure of the dimensions of the empirical data. Six MRM models with one to six latent classes and five Mix2PL models with one to five latent classes were compared. The information criteria, AIC and BIC, were used to select the best fitting model. AIC, BIC, and CAIC selected the 5C-MRM among the six MRM models as the best model. Among the five Mix2PL models, AIC selected the 4C-Mix2PL, and BIC and CAIC selected the 3C-Mix2PL as the best model. Overall, the 5C-

MRM had the smallest AIC, BIC, and CAIC, however, the first latent class of the 5C-MRM seemed like an over-extracted class, because the class size was very small, about 2%, and the estimated item difficulties did not have a specific feature. The results of the 3C-Mix2PL revealed that the characteristics of latent classes mostly depended on the testlet structure. Three groups of testlet items (Q8 though Q11, Q13 through Q15, and Q20 through A22) caused the main differences between the latent classes.

For further information in terms of items to understand the structure of dimensions, an exploratory MIRT model analysis was conducted. Five M2PL models with two to six dimensions were estimated in OpenBUGS software, but the MIRT models with more than three dimensions had convergence problems. Based on the results of the 2D-M2PL, most items tended to measure the first dimensions. It was difficult to figure out the structure of dimensions by using the results of the 2D-M2PL. Additionally, this result was not consistent with the results of the MixIRT models.

Finally, the MMixRIM with two latent classes and two dimensions was used to attain information in terms of persons as well as items. According to the results, seven items had significantly different angles between latent classes. This result indicates that these seven items measure different latent traits for each latent class, and these items overlapped with items which displayed large differences on estimated item parameters between latent classes. It also reflects the fact that the dimensionality is characteristically associated with both persons and items. In conclusion, the MMixRIM can simultaneously provide information associated with person and item, and this information would be more easily interpreted.

Chapter 5

SIMULATION STUDY

The simulation study was conducted to explore how several types of structures of multidimensionality affect the estimation of model parameter. In this chapter, the design, simulation conditions, and procedures of data generation for the simulation study are described in detail.

5.1 Design of the simulation study

The simulation study focused on how different types of multidimensionality can affect the estimated parameters of different multidimensional extensions of the IRT model. Therefore, without loss of generality, factors that are not associated with the dimension structures were fixed to avoid over complicating the design of the simulation study. There are various suggestions about sample size and test length in IRT models, but a large sample size is generally recommended for the accurate and precise estimation of the parameters. As the complexity of a model increases, the sample size and test length should increase to produce enough information for an accurate estimation.

De la Torre and Hong (2010) compared the effects of test lengths, sample sizes, and correlations between dimensions on the estimation of item and person parameters in the two-dimension higher-order IRT model. Based on De la Torre and Hong's (2010) result, only test length had an obvious effect on the estimation of the item parameters. Both sample size and correlations clearly showed effects on the estimation of the person parameters. In particular, the average root mean square errors (RMSEs) of estimated item parameters were smaller when the sample size was 1,000 and the test length was 20 than when the sample

size was 500 and the test length was 10 or 20. Kose and Demirtasli (2012) reported that increasing sample size and test length had no effect under the unidimensional IRT models, whereas those effects were clear under the MIRT models. Both item and ability estimated parameters under the conditions with a larger sample size and longer test length (the sample size was 1,500 and the test length was 24) had smaller root mean square errors and higher reliabilities for the two-dimension IRT model. In the current study, the sample size was fixed as 1,000 and the test length was fixed as 30 for the stable estimation of parameters.

Two groups of factors were used to manipulate different structures of multidimensionality. One group of factors is associated with the distribution of ability. The other group of factors is associated with the structures of multidimensionality. The maximum number of dimensions and the maximum number of latent classes were fixed as two for the simplest multidimensional and mixture contexts, respectively. Four distributions of ability were considered in the simulation study. The first distribution was a normal distribution, that is, a unidimensional data set. The second distribution followed a bivariate normal distribution in which the population is assumed to be homogeneous (i.e., a one class and two dimensions). The third distribution of ability was for the assumption of the heterogeneous population with two latent classes (i.e., two classes). The distribution of ability for each latent class followed a normal distribution. The last distribution of ability simulated a heterogeneous population with two latent classes and two dimensions. The distribution of ability for each latent class followed a bivariate normal distribution. Additionally, two combinations of class size were used for the third and fourth distributions: one is equal size (i.e., 50% and 50%), and the other is a dominant latent class (i.e., 30% and 70%).

Two more factors were used to manipulate dimensional structures: type of multidimensionality and correlation between dimensions. Two types of multidimensionality were simulated: between-items multidimensionality with an approximately simple structure and within-items multidimensionality. There were two item clusters for the between-items multidimensionality. The first item cluster contained 20 items that predominantly measured

the first dimension. The second item cluster contained 10 items that measured mainly the second dimension. For the within-items multidimensionality, there were three item clusters. The first 10 items had larger item slope parameters for the first dimension than for the second dimension. The next 10 items had the reverse pattern. The last 10 items had roughly equal item slope parameters for both dimensions. As in Bolt and Lall (2003), three levels of correlations between dimensions were applied: (1) independent dimensions (i.e., $\rho_{\theta_1\theta_2}$ =.0); (2) a weak correlation (i.e., $\rho_{\theta_1\theta_2}$ =.3); and (3) a strong correlation (i.e., $\rho_{\theta_1\theta_2}$ = .6).

In summary, four of the simulation factors (i.e., distributions of ability, latent class size, type of dimensionality, and correlations between dimensions) were manipulated to make different multidimensional structures. The other two factors, i.e., sample size and test length, were held constant. A total of 15 combinations were compared in the simulation study, and with 100 replications for each combination. One of the 15 condition was a unidimensional and homogenous population, another six conditions were multidimensional and homogeneous. Two additional conditions were unidimensional and heterogeneous populations, and the last six conditions were multidimensional and heterogeneous populations. These 15 combinations of simulation conditions are summarized in Table 5.1.

5.2 Data simulation procedures

The code for distribution of ability is indicated in the left column of Table 5.1. Four generating models were used according to four different distributions of ability. Specifically, a 2PL IRT model, a two-dimension M2PL (2DM2PL) model, a two-class and 2PL mixture IRT (2CMix2PL) model, and a two-dimension and two-class MMixRIM (2D2C-MMixRIM) were all applied for ability distribution conditions 1C1D, 1C2D, 2C, and 2C2D, respectively. The item parameters were modified from Reckase (2009). Item parameters for ability distribution condition 1C2D were selected from Reckase's (2009, p. 204) Table 7.

Table 5.1: Combinations of Simulation Conditions

	Distribution of ability	Latent class size	Type of multidimensionality	Correlation
1	1C1D	-	-	.0
2	1C2D	-	Between-items	.0
3	1C2D	-	Between-items	.3
4	1C2D	-	Between-items	.6
5	1C2D	-	Within-items	.0
6	1C2D	-	Within-items	.3
7	1C2D	-	Within-items	.6
8	2C1D	50% & 50%	-	.0
9	2C1D	30% & 70%	-	.0
10	2C2D	50% & 50%	Within-items	.0
11	2C2D	50% & 50%	Within-items	.3
12	2C2D	50% & 50%	Within-items	.6
13	2C2D	30% & 70%	Within-items	.0
14	2C2D	30% & 70%	Within-items	.3
15	2C2D	30% & 70%	Within-items	.6

Note. 1C1D = One-class and one-dimension; 1C2D = One-class and two-dimension; 2C = Two-class; 2C2D = Two-class and two-dimension.

For condition 1C1D, MDISC ($\sqrt{a_{i1}^2 + a_{i2}^2}$) in the case of 1C2D with between items multidimensionality were used as the item discriminations. For the item difficulty, the intercept parameter (d_i) was divided by MDISC, because the logistic function of M2PL model is defined as $a_{i1}\theta_1 + a_{i2}\theta_2 + d_i$, while the logistic function of 2PL model is defined as $a_i(\theta + b_i)$. That is, the intercept parameter in M2PL model is equivalent to the multiplication of item discrimination and item difficulty in 2PL model. The item parameters for conditions 1C1C, 1C2D with between-items multidimensionality, and 1C2D with within-items multidimensionality are reported in Table 5.2.

For condition 2C1D, the item parameters for 1C1D were modified. Similar to Choi (2014) and Li et al. (2009), two types of knowledge were assumed, and two latent classes were assumed to perform differently according to the type of knowledge. Suppose that Class 1 performs better at the first type of knowledge, and Class 2 performs better at the other type of knowledge. Additionally, the first 10 items (i.e., Q1 to Q10) have the same item parameters for both latent classes, that is, these items act as anchor items. The next 10 items (i.e., Q11 to Q20) measure the first type of knowledge, and the last 10 items (i.e., Q21 to Q30) measure the second type of knowledge. For a good performance, the item discriminations are generated by adding .5 to the item discriminations for condition 1C1D. Similarly, the item difficulties for a good performance class were generated by subtracting 1.5 from the item difficulties for condition 1C1D, and the item difficulties for a poor performance class were generated by adding 1.5. The item parameters for condition 2C1D are presented in Table 5.3 and Figure 5.1.

The item parameters for 2C2D were modified based on the item parameters for 1C2D with within-items multidimensionality to reflect different multidimensional patterns between classes. To be specific, for Class 1, which is a group of students who are good at the first trait (θ_1), items that measure the first trait (i.e., Q1 through Q20) had large slopes for the first dimension, smaller slopes for the second dimension, and larger intercepts than those parameters for Class 2. Additionally, items that measure the second trait (i.e., Q21

Table 5.2: Generating Item Parameters for 1C1D and 1C2D

Item	1C1D (Between-items)		1C2D (Within-items)			1C2D		
	a	b	a_1	a_2	d	a_1	a_2	d
Q1	0.97	0.94	0.97	0.00	0.91	0.97	0.00	0.91
Q2	1.05	−0.21	1.02	0.25	−0.22	1.02	0.25	−0.22
Q3	0.96	−0.50	0.93	0.24	−0.48	0.93	0.24	−0.48
Q4	0.97	−1.14	0.94	0.21	−1.10	0.94	0.21	−1.10
Q5	0.86	0.48	0.84	0.20	0.42	0.84	0.20	0.42
Q6	0.97	−0.60	0.97	0.05	−0.58	0.97	0.05	−0.58
Q7	1.01	−0.88	1.01	0.06	−0.88	1.01	0.06	−0.88
Q8	1.03	1.12	1.01	0.17	1.15	1.01	0.17	1.15
Q9	1.15	1.01	1.14	0.15	1.16	1.14	0.15	1.16
Q10	0.96	−0.40	0.95	0.14	−0.38	0.95	0.14	−0.38
Q11	1.00	0.13	0.98	0.18	0.13	0.85	0.66	−0.52
Q12	0.87	−0.57	0.84	0.20	−0.49	0.76	0.74	0.30
Q13	1.11	0.41	1.08	0.26	0.46	0.73	0.81	−0.62
Q14	0.83	−1.53	0.81	0.15	−1.26	0.60	0.68	−0.45
Q15	1.04	0.10	1.03	0.10	0.10	0.74	0.70	−0.92
Q16	1.09	−0.08	1.09	0.07	−0.09	0.73	0.72	−0.48
Q17	1.07	0.75	1.05	0.22	0.80	0.71	0.55	−0.75
Q18	1.05	−0.04	1.04	0.19	−0.04	0.62	0.68	0.78
Q19	1.00	0.07	0.97	0.23	0.07	0.77	0.80	0.02
Q20	1.07	0.58	1.05	0.16	0.62	0.70	0.69	0.11
Q21	0.90	0.30	0.14	0.89	0.27	0.14	0.89	0.27
Q22	1.06	1.17	0.04	1.06	1.23	0.04	1.06	1.23
Q23	1.05	−0.09	0.02	1.05	−0.09	0.02	1.05	−0.09
Q24	1.18	−0.20	0.02	1.18	−0.24	0.02	1.18	−0.24
Q25	1.03	0.82	0.03	1.03	0.85	0.03	1.03	0.85
Q26	0.93	−0.83	0.08	0.93	−0.78	0.08	0.93	−0.78
Q27	1.03	−0.83	0.21	1.01	−0.86	0.21	1.01	−0.86
Q28	0.90	0.03	0.22	0.87	0.02	0.22	0.87	0.02
Q29	0.99	−0.23	0.20	0.97	−0.23	0.20	0.97	−0.23
Q30	0.99	−0.12	0.03	0.99	−0.12	0.03	0.99	−0.12

Note. a = Item discrimination; b = Item difficulty; a_1 = Slope parameter on the first dimension; a_2 = Slope parameter on the second dimension; d = Intercept parameter.

Table 5.3: Generating Item Parameters for 2C1D

Item	Class 1		Class 2		Item	Class 1		Class 2	
	a	b	a	b		a	b	a	b
Q1	0.97	0.94	0.97	0.94	Q16	1.59	−1.08	1.09	0.92
Q2	1.05	−0.21	1.05	−0.21	Q17	1.57	−0.25	1.07	1.75
Q3	0.96	−0.50	0.96	−0.50	Q18	1.55	−1.04	1.05	0.96
Q4	0.97	−1.14	0.97	−1.14	Q19	1.50	−0.93	1.00	1.07
Q5	0.86	0.48	0.86	0.48	Q20	1.57	−0.42	1.07	1.58
Q6	0.97	−0.60	0.97	−0.60	Q21	0.90	1.30	1.40	−0.07
Q7	1.01	−0.88	1.01	−0.88	Q22	1.06	2.17	1.56	0.17
Q8	1.03	1.12	1.03	1.12	Q23	1.05	0.91	1.55	−1.09
Q9	1.15	1.01	1.15	1.01	Q24	1.18	0.80	1.68	−1.20
Q10	0.96	−0.40	0.96	−0.40	Q25	1.03	1.82	1.53	−0.18
Q11	1.50	−0.87	1.00	1.13	Q26	0.93	0.17	1.43	−1.83
Q12	1.37	−1.57	0.87	0.44	Q27	1.03	0.17	1.53	−1.83
Q13	1.61	−0.59	1.11	1.41	Q28	0.90	1.03	1.40	−0.97
Q14	1.33	−2.53	0.83	−0.53	Q29	0.99	0.77	1.49	−1.23
Q15	1.54	−0.91	1.04	1.10	Q30	0.99	0.88	1.49	−1.12

Note. a = Item discrimination; b = Item difficulty.

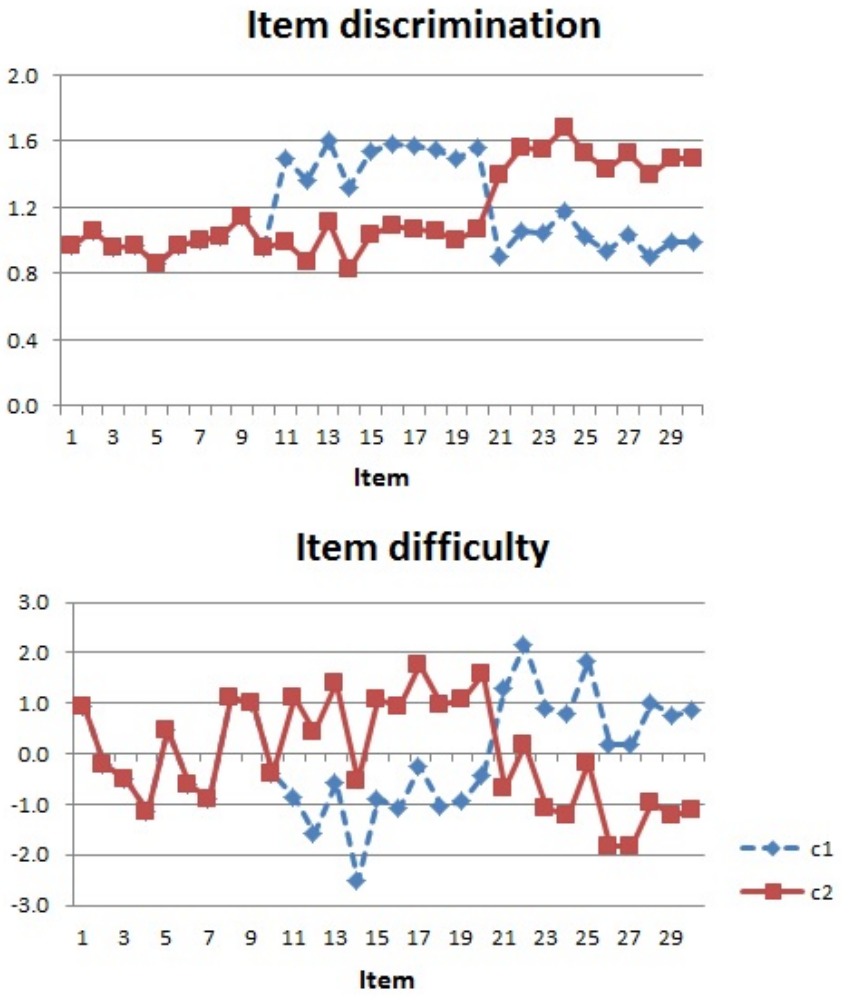

Figure 5.1: Item parameter patterns for two latent classes for 2C1D.

through Q30) had smaller slopes for both dimensions and intercepts than those for Class 2. The opposite pattern was applied to Class 2. That is, Q1 through Q10 had smaller slopes for both dimensions and intercepts than Class 1, and Q11 through Q30 had smaller slope for the first dimension and larger slopes for the second than Class 1. These patterns were formed by adding or subtracting .5 from the item parameters for 1C2D with within-item multidimensionality in the same way used to modify item parameters for 1D2C. The item parameters for 2C2D are presented in Table 5.4 and Figure 5.2.

Table 5.4: Generating Item Parameters for 2C2D

Item	Class 1			Class 2		
	a_1	a_2	d	a_1	a_2	d
Q1	1.47	0.00	1.41	0.47	0.00	0.41
Q2	1.52	0.25	0.28	0.52	0.25	−0.72
Q3	1.43	0.24	0.02	0.43	0.24	−0.98
Q4	1.44	0.21	−0.60	0.44	0.21	−1.60
Q5	1.34	0.20	0.92	0.34	0.20	−0.08
Q6	1.47	0.05	−0.08	0.47	0.05	−1.08
Q7	1.51	0.06	−0.38	0.51	0.06	−1.38
Q8	1.51	0.17	1.65	0.51	0.17	0.65
Q9	1.64	0.15	1.66	0.64	0.15	0.66
Q10	1.45	0.14	0.12	0.45	0.14	−0.88
Q11	1.35	0.16	−0.02	0.35	1.16	−0.02
Q12	1.26	0.24	0.80	0.26	1.24	0.80
Q13	1.23	0.31	−0.12	0.23	1.31	−0.12
Q14	1.10	0.18	0.05	0.10	1.18	0.05
Q15	1.24	0.20	−0.42	0.24	1.20	−0.42
Q16	1.22	0.22	0.02	0.22	1.22	0.02
Q17	1.21	0.05	−0.25	0.21	1.05	−0.25
Q18	1.12	0.18	1.28	0.12	1.18	1.28
Q19	1.27	0.30	0.52	0.27	1.30	0.52
Q20	1.19	0.19	0.61	0.19	1.19	0.61
Q21	0.14	0.39	−0.23	0.14	1.39	0.77
Q22	0.04	0.56	0.73	0.04	1.56	1.73
Q23	0.02	0.55	−0.59	0.02	1.55	0.41
Q24	0.02	0.68	−0.74	0.02	1.68	0.26
Q25	0.02	0.53	0.35	0.02	1.53	1.35
Q26	0.08	0.43	−1.28	0.08	1.43	−0.28
Q27	0.21	0.51	−1.36	0.21	1.51	−0.36
Q28	0.22	0.37	−0.48	0.22	1.37	0.52
Q29	0.20	0.47	−0.73	0.20	1.47	0.27
Q30	0.03	0.49	−0.62	0.03	1.49	0.38

Note. a_1 = Slope parameter on the first dimension; a_2 = Slope parameter on the second dimension; d = Intercept parameter.

The ability parameters for 1C1D were randomly sampled from a standard normal distribution, $\theta \sim N(0,1)$. For 2C1D, the ability parameters for both classes were randomly sampled from a standard normal distribution, $\theta_g \sim N(0,1)$, respectively, similar to the ability parameters for 1C1D. The ability parameters for 1C2D with different levels of correlation between dimensions were simulated by applying the procedure described by Oshima, Raju, and Flowers (1997). For the two independent dimensions condition, the ability parameters (θ_1 and θ_2) were randomly sampled from a bivariate normal distribution with means of 0 and a unit covariance matrix. For the correlated dimensions, the ability parameters were simulated by weighted linear transformation with weights. The weights were the elements of $\boldsymbol{L}^{'}$ when the correlation matrix, $\boldsymbol{R}$, is decomposed as $\boldsymbol{R} = \boldsymbol{L}\boldsymbol{L}^{'}$. For instance, the correlation matrix is $\boldsymbol{R} = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$, and the weights are the elements of $\boldsymbol{L}^{'} = \begin{bmatrix} 1 & .6 \\ 0 & .8 \end{bmatrix}$. Then, new correlated ability parameters were generated as follows:

$$\boldsymbol{L}^{'} \times \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \text{new } \theta_1 \\ \text{new } \theta_2 \end{bmatrix}, \tag{5.1}$$

where θ_1 and θ_2 are independent abilities which are generated for 1C2D.

A total of 15 sets of generated data were analyzed using nine MMixRIM models with combinations of one to four dimensions and one to four latent classes. The maximum numbers of dimensions and latent classes of the generated data were both two. Models with four dimensions or four latent classes were applied to detect the effects of structures of dimensionality on the estimation of model parameters. This was done because more complex models generally tend to fit better than simpler models. A MMixRIM with one class and one dimension is equivalent to the 2PL with a random persons and random items model (Rijmen & De Boeck, 2005). A MMixRIM with one class and two dimensions is equal to the two-dimensional M2PL with a random persons and random items model. Likewise, a MMixRIM with two classes and one dimension corresponds to the two-class Mix2PL with a random persons and random items model. Therefore, the nine IRT models applied

in this simulation study were 2PL model, two- to four-dimensional M2PL models, two-to four-class Mix2PL models, two-dimensional and two-class MMixRIM, and two-dimensional and three-class MMixRIM. All IRT models applied in the simulation study treated both persons and items as random. Thus, 2PL, M2PL, and Mix2PL models mean extensions of a random item model hereafter.

5.3 Recovery analysis

A recovery analysis was conducted to estimate the quality of estimated parameters under different structures of multidimensionality. For the recovery analysis, the estimated item parameters and the generated item parameters were compared using three indices: (1) BIAS, (2) root mean square error (RMSE), and (3) Pearson correlations. These three indices were computed with the following equations:

$$BIAS(\hat{\beta}) = E(\hat{\beta}) - \beta, \tag{5.2}$$

$$RMSE(\hat{\beta}) = \sqrt{E((\hat{\beta} - \beta)^2)}, \tag{5.3}$$

$$Corr(\hat{\beta}, \beta) = \frac{Cov(\hat{\beta}, \beta)}{\sigma_{\hat{\beta}}\sigma_{\beta}}, \tag{5.4}$$

where $\hat{\beta}$ is an estimator for a parameter β and $E(\cdot)$ is the expected value. BIAS and RMSE indicate the accuracy of estimation, and smaller BIAS and RMSE values indicate more accurate estimation.

Unlike BIAS and RMSE, the Pearson correlation can be used to compare parameters in different metrics. In this case, it was used to provide information about the relationship between generated and estimated parameters. The BIAS and RMSE for the estimated item parameters of models without latent class (i.e., a_{ik} and d_i) were computed across replications

and items. The BIAS and RMSE for the estimate item parameters of models with latent classes (i.e., a_{ig}, a_{ikg}, b_{ig}, and d_{ig}) were computed across replication, latent classes, and items. BIAS, RMSE, and the mean of correlations for each over 100 replications are reported for each conditions.

5.4 Linking of scales for recovery analyses

Before computing BIAS and RMSE, estimated and generated parameters should be placed on the same scale. The estimated parameters for each replication, however, were on their own scale. Therefore, the estimated parameters needed to be transformed onto a common scale. In this study, the common scale was the scale of the generating parameters. A classical and simple method is a linear equating. A linear equating assumes a linear relationship between scores on the different scales. This linear relationship is determined by the transformation coefficients (Hambleton, Swaminathan, & Rogers, 1991).

For the unidimensional IRT model, the mean and sigma method was used to determine the transformation coefficients. As the name suggests, the transformation coefficients of the mean and sigma method can be obtained from the means and standard deviations of item parameters on the base scale and the target scale. In this case, the base scale means the scale of the generating item parameters, and the target scale means the scale of the estimated item parameters for each replication. The transformation coefficients (i.e., α and β) are defined as follows:

$$\alpha = \frac{S_{b_B}}{S_{b_T}}, \tag{5.5}$$

$$\beta = \bar{b}_B - \alpha\bar{b}_T, \tag{5.6}$$

where S_{b_T} represents the standard deviation of the item difficulty parameters on the target scale, S_{b_B} represents the standard deviation of the item difficulty parameters on the base

scale, $\bar{b}_T$ represents the mean of the item difficulty parameters on the target scale, and $\bar{b}_B$ is the mean of the item difficulty parameters on the base scale. By using these transformation coefficients, the estimated item parameters are transformed onto the same scale of the generated item parameters as below:

$$\tilde{a}_i = \frac{a_i}{\alpha}, \tag{5.7}$$

$$\tilde{b}_i = \beta + \alpha b_i, \tag{5.8}$$

where a_i and b_i are the estimated item discrimination and difficulty of item i, and $\tilde{a}_i$ and $\tilde{b}_i$ represent transformed parameters of these estimates.

Similar to the unidimensional IRT model, the estimated parameters from the MIRT models are also required to transform onto the same coordinate system of the generated parameters. For the multidimensional IRT model, three types of indeterminacy are in (1) placement of the origin, (2) selection of units of measurement along axes, and (3) orientation of the axes (De Ayala, 2009; Reckase, 2009). That is, the origin, scale, and orientation of each coordinate axis should be adjusted to get comparable item parameters for the multidimensional linking. In this simulation study, the same procedure of multidimensional linking described by Reckase (2009) was conducted as follows:

$$\tilde{\mathbf{a}} = \mathbf{a}\mathbf{M}', \tag{5.9}$$

$$\tilde{\mathbf{d}} = \mathbf{d} - \mathbf{a}\mathbf{M}'\mathbf{s}', \tag{5.10}$$

where $\mathbf{a}$ is a $n \times k$ matrix of the estimated slope parameters, $\mathbf{d}$ is a $n \times 1$ vector of the estimated intercept parameters, and $\tilde{\mathbf{d}}$ and $\tilde{\mathbf{a}}$ are sets of transformed item parameters onto the same coordinate system of the generated parameters. $\mathbf{M}$ is a $k \times k$ matrix to use for the nonorthogonal rotation of the coordinate system, and $\mathbf{s}$ is a $1 \times k$ matrix to use for the shift origin of coordinate axes. These two matrixes, (i.e., $\mathbf{M}$ and $\mathbf{s}$), are calculated as follows:

$$\mathbf{M}^{-1} = \left(\left(\boldsymbol{\theta}_T - \bar{\boldsymbol{\theta}}_T \right)' \left(\boldsymbol{\theta}_T - \bar{\boldsymbol{\theta}}_T \right) \right)^{-1} \left(\boldsymbol{\theta}_T - \bar{\boldsymbol{\theta}}_T \right)' \left(\boldsymbol{\theta}_B - \bar{\boldsymbol{\theta}}_B \right)' \tag{5.11}$$

$$\mathbf{s} = \bar{\boldsymbol{\theta}}_B - \bar{\boldsymbol{\theta}}_T \mathbf{M}^{-1}, \tag{5.12}$$

where $\boldsymbol{\theta}_T$ is a $N \times k$ matrix of the estimated ability parameters of N examinees on the k-dimensional space, $\boldsymbol{\theta}_B$ is a $N \times k$ matrix of the generated ability parameters, and $\bar{\boldsymbol{\theta}}_T$ and $\bar{\boldsymbol{\theta}}_B$ are the mean vectors of the estimated and generated ability parameters, respectively. The results of the simulation study are described in the following chapter.

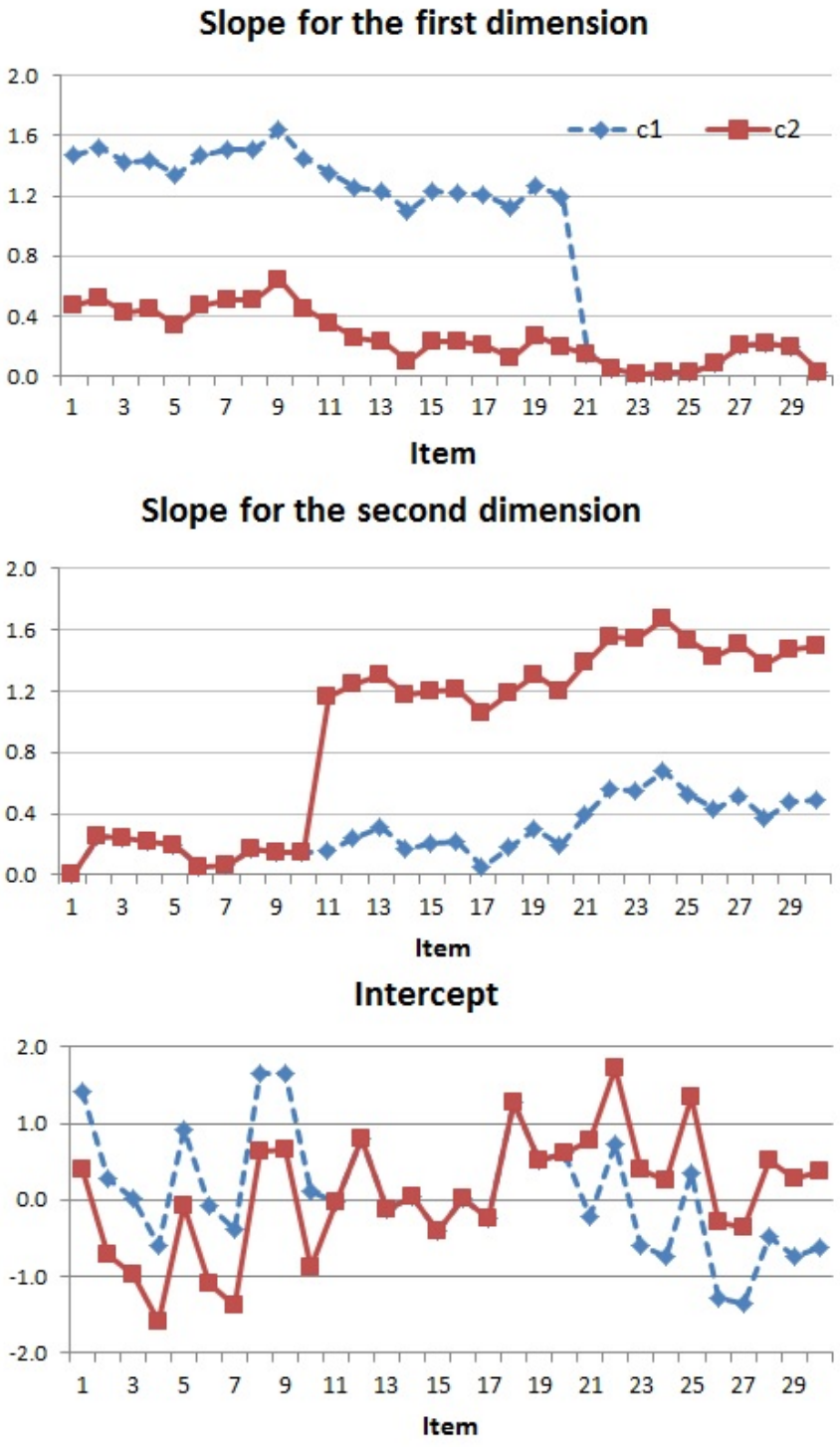

Figure 5.2: Item parameter patterns for two latent classes for 2C2D.

Chapter 6

RESULTS

The results of the simulation study, including monitoring convergence, model comparison, and recovery analyses, are presented in this chapter. The main purpose of this simulation study was to explore the performance of MMixRIM for analysis of the multidimensional data. In addition, the effect of different multidimensional structures on the estimation of model parameters was examined. As described in the previous chapter, the data sets were generated based on the 15 combinations depending on the four different kinds of ability distribution, which are 1) unidimensional distribution, 2) two-dimensional distribution, 3) mixture distribution consisting of two latent classes, and 4) two-dimensional and mixture distribution consisting of two latent classes. For the two-dimensional ability distribution, two types of dimension structure (the between-items multidimensionality and within-items multidimensionality) and three levels of correlations between dimensions (ρ =.0, .3, and .6) were manipulated. For the mixture distribution with two classes, the combinations of latent class sizes (equal size of latent classes, 50% & 50% and a dominant class, 30% & 70%) were manipulated. For convenience, the 15 conditions names represent the combinations of simulation conditions as summarized in Table 6.1.

For example, D1 represents the condition of the unidimensional ability distribution, and D2BR0 represents the combination of the two-dimensional ability distribution (2D), the between-items multidimensionality (B), and two independent dimensions (R0). Similarly, D1C2E represents the condition of the mixture ability distribution with two latent classes (D1C2) and the sizes of latent class are equal (E). D2C2R3D indicates the combinations of the two-dimensional and mixture ability distribution (D2C2), weakly correlated dimensions

Table 6.1: Names of Simulation Conditions

Name	ND	NC	Type of multidimensionality	Correlation	Latent class size
D1	1	1	-	-	-
D2BR0	2	1	Between-items	.0	-
D2BR3	2	1	Between-items	.3	-
D2BR6	2	1	Between-items	.6	-
D2WR0	2	1	Within-items	.0	-
D2WR3	2	1	Within-items	.3	-
D2WR6	2	1	Within-items	.6	-
D1C2E	1	2	-	-	50/50
D1C2D	1	2	-	-	30/70
D2C2R0E	2	2	Within-items	.0	50/50
D2C2R3E	2	2	Within-items	.3	50/50
D2C2R6E	2	2	Within-items	.6	50/50
D2C2R0D	2	2	Within-items	.0	30/70
D2C2R3D	2	2	Within-items	.3	30/70
D2C2R6D	2	2	Within-items	.6	30/70

Note. ND=Number of dimensions; NC=Number of latent classes.

(R3), and a dominant latent class (D).

6.1 Monitoring Convergence

For estimation with the MCMC algorithm, the chain should be converged to get accurate parameter estimates. As described in Chapter 3, several methods were suggested for monitoring convergence of the chain. In this study, a single chain with 10,000 iterations for the burn-in period and 10,000 iterations for the post-burn-in period were used. The convergence of all estimated item parameters was tested by using Heidelberger and Welch (1983) index. This index provides a convergence diagnostic for a single long chain.

For the unidimensional data condition (i.e., D1), the number of estimated item parameters was 60, which includes the 30 item discriminations and 30 item difficulties. For the two-dimensional structure conditions, (i.e., D2BR0, D2BR3, D2BR6, D2WR0, D2WR3, and D2WR6), the number of estimated item parameters was 84; this includes 56 (= 28 items $\times$ 2 dimensions) item slope parameters and 28 item intercept parameters because the six item parameters of the two items were fixed for the model identification. For the two latent classes conditions (i.e., D1C25 and D1C23), 120 item parameters were estimated. That is, 30 item discriminations and 30 item difficulties for each latent class were estimated. For the two-dimensional and two-class conditions, (i.e., D2C2R0E, D2C2R3E, D2C2R6E, D2C2R0D, D2C2R3D, and D2C2R6D), a total of 168 item parameters were estimated. Specifically, 56 item slopes and 28 item intercept parameters were estimated for each latent class. Tables B.1 and B.2 indicate the percentages of the passed item parameters based on Heidelberger and Welch's approach for each condition and replication.

For the unidimensional data condition, the item parameters were estimated by a 2PL model. As summarized in Table B.1, for the unidimensional condition, the percentages of the passed item parameters were higher than .9 in the 90 replications, and the average of percentages of passed item parameters over 100 replications was .96. That is, the chains

of most replications converged based on Heidelberger and Welch's convergence diagnostics. The Mix2PL model was applied to estimate the item parameters for the two-class conditions. The percentages of passed item parameters of D1C2D were relatively smaller than those of D1C2E. The averages of percentages of passed item parameters over all replications for D1C2D and D1C2E were .84 and .94, respectively. For the two-dimensional conditions, an M2PL model was applied to estimate the item parameters. Most replications for the condition of two independent dimensions (i.e., D2WR0 and D2BR0) achieved almost perfect convergence. The average of the percentages of passed item parameters was .99 for both conditions. The conditions of weakly (or strongly) correlated dimensions were also mostly converged. For these conditions, the average percentages of passed parameters were higher than .96. A MMixRIM was applied to estimate item parameters for the two-dimensional and two-class conditions. Similar to the other conditions, most replications showed converged chains, except for three replications. For these three replications, the percentages were smaller than .50. One replication was in condition D2C2R0D, another replication was in condition D2C2R3D, and the third replication was in D2C2R3E. The averages of percentages of passed item parameters for the two-dimensional and two-class conditions were between .95 and .98. To sum up, the convergence for each condition was good enough, although it was not percept based on Heidelberger and Welch's convergence diagnostics.

In addition to Heidelberger and Welch convergence diagnostics, the ratio of MC error to the standard deviation of item parameter estimates from the post-burn-in iterations was used for monitoring convergence of the chain. The percentages of item parameters that the MC error is less then 5% of the standard deviation are summarized in Tables B.3 and B.4.

For the unidimensional condition, all replications except one had perfectly converged chains based on the ratio of MC error to the standard deviation. Unlike the results of Heidelberger and Welch convergence diagnostics in Tables B.1 and B.2, all replications for the two-class conditions with a dominant latent class (i.e., D1C2D) failed to converge. The ratios of MC error to the standard deviations of most items were larger than 5%. For the

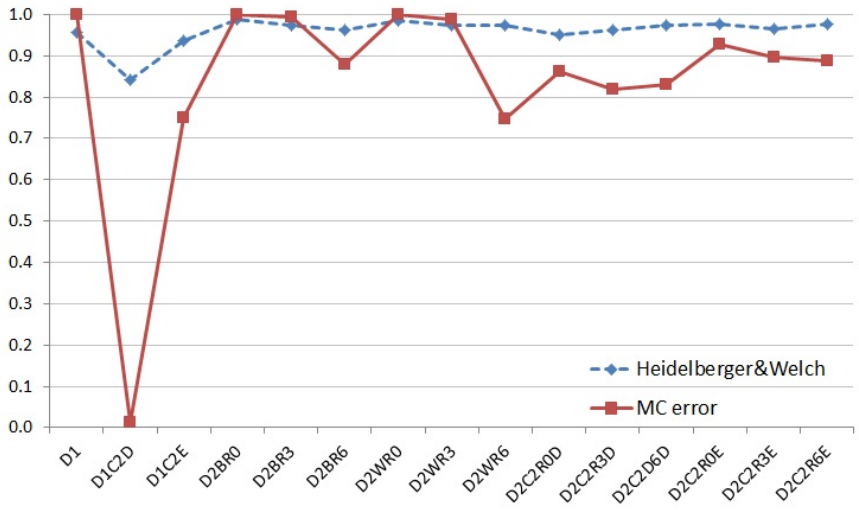

Figure 6.1: Percentages of passed item parameters based on Heidelberger and Welch's convergence diagnostics and the ratio of MC error to standard deviation.

D1C2E, the MC errors of all estimated item difficulties for Class 2 were larger than 5% of the standard deviation in every replication. Thus, the percentages of items for which the ratio of MC error to the standard deviation was less than 5% were 75% (30 item parameters among 120 item parameters) for all conditions. The MC errors of all estimated item parameters for the two-dimensional data with independent dimensions (i.e., D2BR0 and D2WR0) were smaller than 5% of the standard deviation in all replications. For the two-dimensional and two-class conditions, the chains mostly converged based on the ratios of MC error to the standard deviation, except for some replications.

Overall, the replications, except for some conditions, had good convergence. The averages of percentages of passed items based on Heidelberger and Welch's way and the ratio of MC error to standard deviation are compared in Figure 6.1. For the unidimensional conditions, the percentages of passed items parameters based on Heidelberger and Welch convergence diagnostics were slightly smaller than these percentages based on the ratio of MC error to the standard deviation. For the two-class conditions, the item parameters more easily passed

the convergence test when the Heidelberger and Welch index was used than when the ratio of MC error to the standard deviation was used. For the two-dimensional conditions, the results of monitoring convergence were similar for Heidelberger and Welch method and the ratio of MC error to the standard deviation, except when the correlation between dimensions was .6. For the two-dimensional and two-class conditions, the percentages of the passed item parameters based on Heidelberger and Welch's approach were larger than those based on the ratio of MC error to the standard deviation.

6.2 Model Comparison

The data sets for this simulation study were generated based on the 15 different conditions described above to have different dimensional structures. Four types of IRT models, 2PL, M2PL, Mix2PL, and MMixRIM models, were applied to these generated data sets to compare the performance of each model for the multidimensional data sets with a different dimensional structure.

In the first part of this section, the percentages of correct model selections based on the five different information criterion indices (i.e., AIC, BIC, CAIC, AIC_c, and ABIC) are summarized. The correct model selection means that each index suggests the true model, which is a model used to generate data sets for each condition, as the best-fitting model among several candidate models. The true model for the unidimensional condition (i.e., D1) was the 2PL model, and the true model for two-dimensional conditions (i.e., D2BR0, D2BR3, D2BR6, D2WR0, D2WR3, and D2WR6) was the two-dimensional M2PL model. The Mix2PL with two classes model was the true model for the two latent classes conditions (i.e., D1C25 and D1C23), and the two-dimensional and two-class MMixRIM was the true model for the two-dimensional and two latent classes conditions (i.e., D2C2R0E, D2C2R3E, D2C2R6E, D2C2R0D, D2C2R3D, and D2C2R6D). To calculate the percentage of correct model selections, the IRT models within the same type of the true model were compared.

For example, the comparison of one- to four-class Mix2PL models was done to determine the percentage of correct model selections for condition D1C2E. This was not possible for the other five IRT models (i.e., two- to four-dimensional M2P models, two-dimensional and two-class MMixRIM and two-dimensional and 3-class MMixRIM). In the second part of this section, the results of the comparison of IRT models within a different type from the true model are presented. For instance, for D1C2E, the results of comparisons between two- to four-dimensional M2PL models based on information criterion indices are described in the second section of this chapter.

6.2.1 Correct model selections

The percentages of correct model selections based on five different information criterion indices (i.e., AIC, BIC, CAIC, AIC_c, and ABIC) are summarized in Table 6.2 and Figures 6.2 to 6.4. As summarized in Table 6.2 and Figure 6.2, the percentages of correct model selections based on AIC differed depending on a type of the true mode. For the unidimensional condition (i.e., D1), AIC detected the true model, which is 2PL model, as the best-fitting model in 78 replications among 100 replications. For the two-dimensional conditions (i.e., D2B2R0E, D2B2R3E, D2B2R6E, D2B2R0D, D2B2R3D, and D2B2R6D), however, AIC failed to detect the true model. No replication was observed for correct model selection. The percentages of correct model selections based on AIC were very high for the two latent classes conditions (i.e., D1C25 and D1C23). The percentages of correct model selections for these two conditions were 98% and 97%, respectively. Similarly, AIC performed well for the two-dimensional and two-class conditions (i.e., D2C2R0E, D2C2R3E, D2C2R6E, D2C2R0D, D2C2D3D, and D2C2R6D). When the sizes of two classes were equal (i.e., D2C2R0E, D2C2R0E, and D2C2R6E), the percentages of correct model selections based on AIC were 88%, 93%, and 91%, respectively. When the size of one class was larger than another class (i.e., D2C2R0D, D2C2R3D, D2C2R6D), the percentages of correct model selections were 84%, 88% and 87%, respectively.

Table 6.2: Percentage of Correct Model Selection

True model	Structure	Correlation	Latent class size	AIC	BIC	CAIC	AIC_c	ABIC
2PL				.78	.91	.91	.00	.00
2DM2PL	Between-items	.0		.00	.09	.02	.00	.00
		.3		.00	.09	.00	.00	.00
		.6		.00	.21	.00	.00	.00
	Within-items	.0		.00	.18	.04	.00	.00
		.3		.00	.23	.06	.00	.00
		.6		.00	.27	.10	.00	.00
2CMix2PL			50% & 50%	.98	1.00	1.00	.88	.89
			30% & 70%	.97	1.00	.99	.88	.89
2D2CMMixRIM	Within-items	.0	50% & 50%	.88	.01	.04	.93	.96
		.3	50% & 50%	.93	.01	.07	.97	.96
		.6	50% & 50%	.91	.01	.15	.95	.95
		.0	30% & 70%	.84	.00	.01	.88	.89
		.3	30% & 70%	.88	.00	.02	.90	.89
		.6	30% & 70%	.87	.01	.09	.93	.94
2PL				.78	.91	.91	.00	.00
2dM2PL				.00	.18	.05	.00	.00
Mix2PL2c				.98	1.00	1.00	.88	.89
2dMMixRIM2c				.89	.07	.06	.93	.93
	Between-items			.00	.13	.03	.00	.00
	Within-items			.59	.08	.06	.62	.00
		.0		.43	.30	.27	.67	.69
		.3		.45	.08	.04	.47	.69
		.6		.45	.13	.01	.47	.72
			50% & 50%	.93	.26	.32	.93	.94
			30% & 70%	.89	.26	.28	.90	.90

Note. 2DM2PL = Two-dimensional M2PL; 2CMix2PL = Two-class Mix2PL; 2D2CMMixRIM = Two-dimensional and two-class MMixRIM.

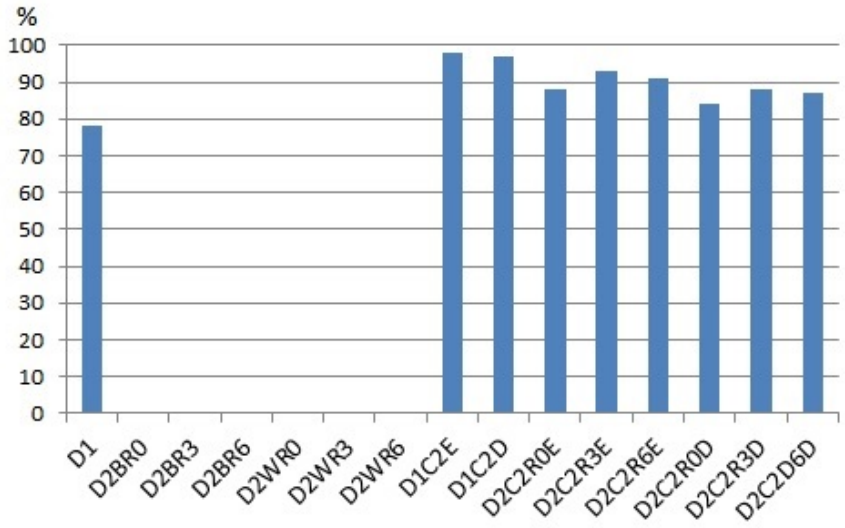

Figure 6.2: Percentages of correct model selection based on AIC.

Table 6.2 and Figure 6.3 show the percentages of correct model selections based in BIC. BIC detected the correct model 91% of the replications for condition D1. For the two-dimensional conditions, however, BIC did not work well to detect the true model among one- to four-dimensional M2PL models regardless of the type of dimensionality structure and the correlation between dimensions. BIC tended to suggest the three-dimensional M2PL model as the best-fitting model, instead of the correct two-dimensional M2PL model. For the two-class conditions, BIC perfectly detected the true model. Unlike AIC, however, BIC could not detect the true model for the two-dimensional and two-class conditions regardless of the correlation between dimensions and the size of latent classes.

Table 6.2 and Figure 6.4 showed the percentages of correct model selection based on CAIC, and the patterns of model selections were similar to those based on BIC. Only the uni-dimensional condition (i.e., D1) and two-class conditions (i.e., D1C2E and D1C2D) observed high percentages of correct model detection based on CAIC. The percentages of correct model selection for D1, D1C2E, and D1C2D were 91%, 100%, and 99%, respectively. The percent-

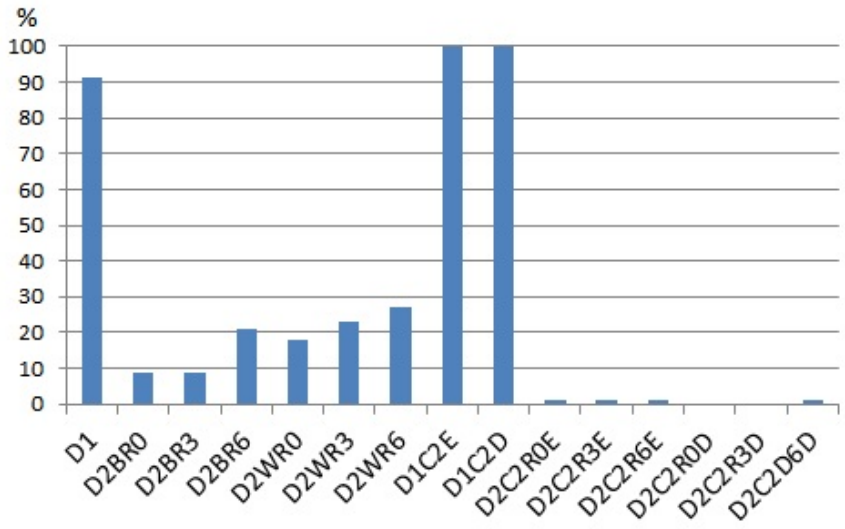

Figure 6.3: Percentages of correct model selection based on BIC.

ages of true model detections were below 10% for the two-dimensional conditions regardless of the type of multidimensional structure and the correlation between dimensions. For the two-dimensional and two-class conditions, although the percentages based on CAIC were slightly higher than the percentages based on BIC, these percentages were still very low, less than 15%.

The results of correct model selections based on AIC_c and ABIC were consistent as can be seen in Figures 6.5 and 6.6. These two information criteria failed to detect the true model for both the unidimensional condition (i.e., D1) and the two-dimensional conditions (i.e., D2BR0, D2BR3, D2BR6, D2WR0, D2WR3, and D2WR6). The percentages of correct detections for these conditions were zero. On the other hand, AIC_c and ABIC performed well for the two-class conditions and the two-dimensional and two-class conditions. For the two-class conditions (i.e., D1C2E and D1C2D), the percentages based on AIC_c and ABIC were 88% and 89%, respectively. For the two-dimensional and two-class conditions, the percentages of correct model selections based on AIC_c and ABIC were above 90% when the sizes of latent

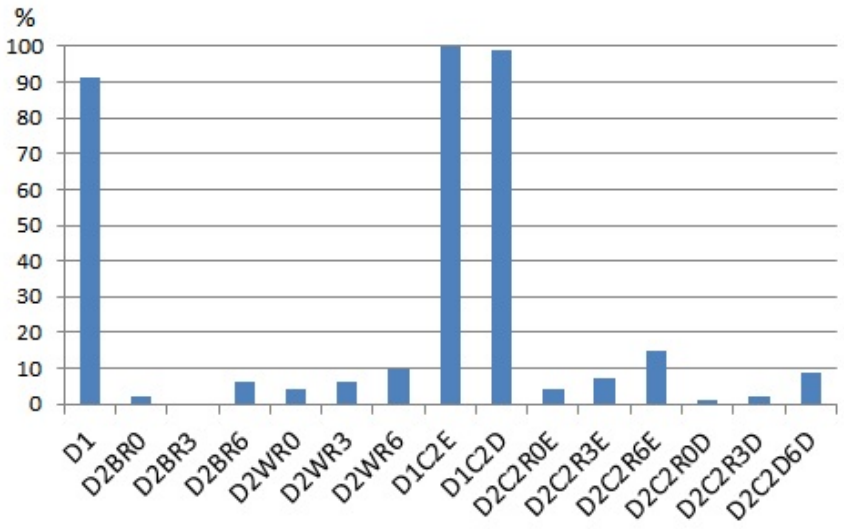

Figure 6.4: Percentages of correct model selection based on CAIC.

classes were equal (i.e., D2C2R0E, D2C2R3E, and D2C2R6E). For the two-dimensional and one dominant latent class conditions (i.e., D2C2R0D, D2C2R3D, and D2C2R6D), percentages of correct model selection based on these two information criterion indices were between 88% and 94%.

In summary, the five information criterion indices performed differently depending on the type of true IRT models. All indices performed very well regardless of the size of latent classes for the Mix2PL model, while their performance for M2PL model was poor. For M2PL model, BIC and CAIC tended to suggest the three-dimensional M2PL model, whereas AIC, AIC_c and ABIC tended to select the four-dimensional M2PL model. For the MMixRIM, AIC, AIC_c and ABIC seems to perform better than BIC and CAIC. For these conditions, BIC and CAIC tended to suggest two-dimensional and one-class MMixRIM rather than two-dimensional and two-class MMixRIM.

As described in the Methods chapter, the performance of information criteria depended on the sample size relative to the number of model parameters. Moreover, AIC tended to

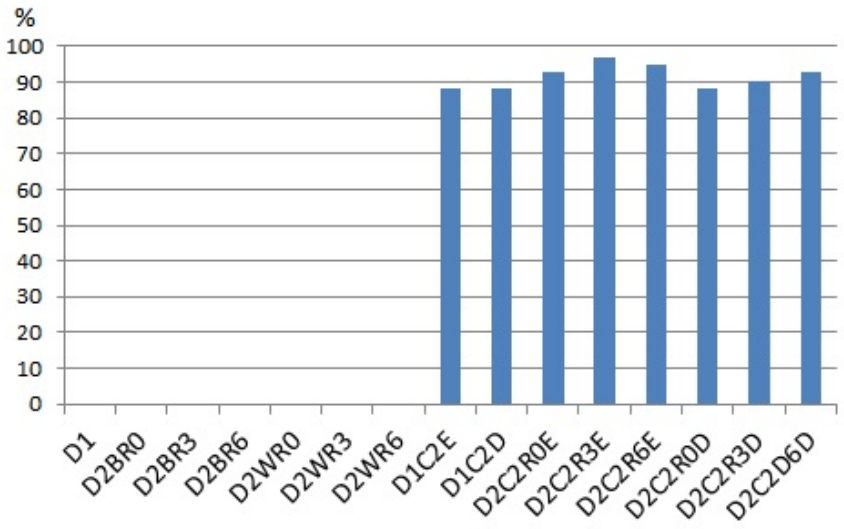

Figure 6.5: Percentages of correct model selection based on AIC_c.

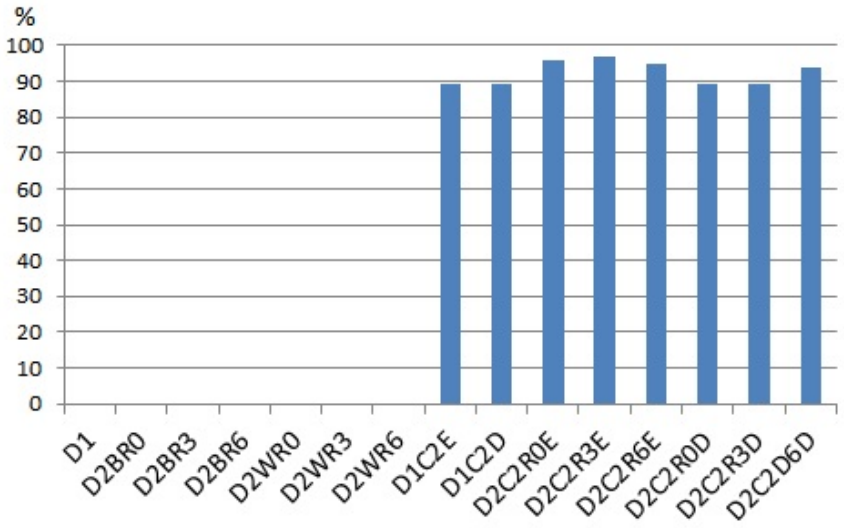

Figure 6.6: Percentages of correct model selection based on ABIC

perform better for a small sample size. BIC, on the other hand, was better for a large sample size (Dziak et al., 2012). In this study, the sample size and the number of items were fixed

at 1,000 and 30, respectively, but the number of model parameters differed based on the type of model. For example, the number of parameter for the 2PL model is 64 (30 item discriminations, 30 item difficulties, one mean and variance of item difficulty, and one mean and variance of item discriminations). For the two-dimensional M2PL model, the number of model parameters is 90, consisting of 28 item slopes for the first dimension, 28 item slopes for the second dimension, 28 intercepts, three means and three variances of item parameter distributions (i.e., $28 \times 3 + 3 + 3 = 90$). The number of model parameters for the two-dimensional and two-class MMixRIM is twice of the number of parameters for the two-dimensional M2PL model, that is 180, which is 90 model parameters per each latent class. For the two-class Mix2PL model, the number of parameters is 130, containing 30 item discriminations, 30 item difficulties, two means and variances of the distribution of item parameters for each class, one mixing proportion, and one mean of ability distribution (i.e., $2 \times (30 + 30 + 4) + 1 + 1 = 130$). Therefore, the relative sample size for each model differed, although the sample size and the number of items were fixed. According to the results of model selections, AIC and sample adjusted information criteria (i.e., AIC_c and ABIC) performed much better than BIC and CAIC to detect the true model for MMixRIM, which is the more complex model than M2PL and Mix2PL models. For the simpler model (i.e., 2PL), BIC and CAIC performed better than AIC, and AIC_c and ABIC showed poor performance to detect the true model. Consequently, the different results of model selections by each information criterion indices might be results from the different ratio of sample size to the number of model parameters.

6.2.2 Model comparisons within the different type of IRT models from the true model

In addition to IRT models within the same type to the true model, other types of IRT models were applied to explore how the dimensional structure appears in different IRT models. The results of comparisons between IRT models within the different type from the true model are

summarized in this section. For the data sets of the two-dimensional conditions (i.e., D2BR0, D2BR3, D2BR6, D2WR0, D2WR3, and D2WR6), the results of the comparison between Mix2PL models and the comparison between MMixRIMs are summarized in Tables 6.3 and 6.4, respectively. The numbers in these tables indicate the number of replications that have been selected as the best-fitting model over 100 replications.

As shown in Table 6.3, all five information criterion indices mainly suggested the two-class Mix2PL as the best-fitting model, when Mix2PL models were applied to the two-dimensional data sets, regardless of the type of multidimensionality and the degree of correlation between dimensions. Table 6.4 presents the results of model comparisons when two-dimensional MMixRIMs were applied to the two-dimensional data sets. All information criterion indices mostly suggested the two-dimensional and one-class MMIxRIM, which is equivalent to the M2PL model. Compared with other indices, AIC suggested M2PL model as the best-fitting model in fewer replications.

In addition to Mix2PL models, M2PL models and MMixRIMs were also applied to the data sets of two-class conditions. The results of the comparison of M2PL models and the comparison of MMixRIMs are summarized in Tables 6.5 and 6.6, respectively. When M2PL models were applied to the data sets of the two-class conditions, the suggestions by each index were not consistent. AIC, AIC_c, and ABIC selected the four-dimensional M2PL model as the best-fitting model in most replications regardless of the size of latent classes. BIC, however, selected the two- or three-dimensional M2PL model as the best-fitting model about half the time, respectively. The two-dimensional M2PL model was selected in 51 and 57 replications for D1C2E and D1C2D, respectively. The next most suggested model based on BIC was the three-dimensional M2PL model. The three-dimensional M2PL model was selected in the 43 and 39 replications for D1C2E and D1C2D, respectively. CAIC selected three-dimensional M2PL model in about 50% of replications (47 replications for D1C2E and 52 replications for D1C2D).

Table 6.3: Results of Comparisons between Mix2PL Models for Two-dimensional Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D2BR0	2PL	0	0	0	0	0
	2CMix2PL	**93**	**100**	**100**	**97**	**98**
	3CMix2PL	3	0	0	3	2
	4CMix2PL	4	0	0	0	0
D2BR3	2PL	0	0	0	0	0
	2CMix2PL	**100**	**100**	**100**	**100**	**100**
	3CMix2PL	0	0	0	0	0
	4CMix2PL	0	0	0	0	0
D2BR6	2PL	0	2	1	0	0
	2CMix2PL	**98**	**98**	**99**	**99**	**99**
	3CMix2PL	2	0	0	1	1
	4CMix2PL	0	0	0	0	0
D2WR0	2PL	0	0	0	0	0
	2CMix2PL	**97**	**99**	**99**	**99**	**99**
	3CMix2PL	2	1	1	1	1
	4CMix2PL	1	0	0	0	0
D2WR3	2PL	0	0	0	0	0
	2CMix2PL	**100**	**100**	**100**	**100**	**100**
	3CMix2PL	0	0	0	0	0
	4CMix2PL	0	0	0	0	0
D2WR3	2PL	1	12	11	2	5
	2CMix2PL	**89**	**88**	**89**	**94**	**93**
	3CMix2PL	9	0	0	4	2
	4CMix2Pl	1	0	0	0	0

Note. 2CMix2PL = Two-class Mix2PL; 3CMix2PL = Three-class Mix2PL; 4CMix2PL = Four-class Mix2PL. The largest number is bold.

Table 6.4: Results of Comparisons between MMixRIMs for Two-dimensional Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D2BR0	M2PL	**85**	**99**	**98**	**88**	**91**
	2D2CMMixRIM	10	1	2	9	7
	2D3CMMixRIM	5	0	0	3	2
D2BR3	M2PL	**86**	**100**	**100**	**89**	**93**
	2D2CMMixRIM	7	0	0	6	5
	2D3CMMixRIM	7	0	0	5	2
D2BR6	M2PL	**81**	**100**	**100**	**88**	**93**
	2D2CMMixRIM	7	0	0	6	4
	2D3CMMixRIM	12	0	0	6	3
D2WR0	M2PL	**85**	**98**	**98**	**94**	**95**
	2D2CMMixRIM	9	2	2	5	4
	2D3CMMixRIM	6	0	0	1	1
D2WR3	M2PL	**79**	**100**	**96**	**88**	**92**
	2D2CMMixRIM	9	0	4	7	5
	2D3CMMixRIM	12	0	0	5	3
D2WR6	M2PL	**74**	**98**	**97**	**84**	**89**
	2D2CMMixRIM	15	2	3	14	10
	2D3CMMixRIM	11	0	0	2	1

Note. 2D2CMMixRIM = Two-dimensional and two-class MMixRIM; 2D3CMMixRIM = Two-dimensional and three-class MMixRIM. The largest number is bold.

Table 6.5: Results of Comparisons between M2PL Models for Two-class Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D1C2E	M2PL	0	0	0	0	0
	2DM2PL	0	**51**	24	0	1
	3DM2PL	4	43	**47**	7	17
	4DM2PL	**96**	6	29	**93**	**82**
D1C2D	M2PL	0	0	0	0	0
	2DM2PL	0	**57**	25	0	0
	3DM2PL	1	39	**52**	4	12
	4DM2PL	**99**	4	23	**96**	**88**

Note. 2DM2PL = Two-dimensional M2PL; 3DM2PL = Three-dimensional M2PL; 4DM2PL = Four-dimensional M2PL. The largest number is bold.

As summarized in Table 6.6, the different patterns of model selections were observed depending on information criterion indices. AIC and AIC_c largely suggested the two-dimensional and two-class MMixRIM as the best-fitting, while BIC and CAIC selected two-dimensional and one-class MMixRIM in most replications. The model mostly suggested based on ABIC was the two-dimensional and one-class MMixRIM (about 55%); the two-dimensional and two-class MMixRIM was second most (about 40%).

For the two-dimensional and two-class conditions, M2PL and Mix2PL models were additionally applied. The results of the comparisons of M2PL and Mix2PL models are presented in Tables 6.7 and 6.8, respectively. As observed in other model comparisons, AIC and AIC_c tended to select the more complex model than other indices. These indices selected the four-dimensional M2PL model in more than half of replications regardless of the type of multidimensionality and the correlation between dimensions. The percentages of the selection of the four-dimensional M2PL model based on AIC were between 74% and 92%. The percentages

Table 6.6: Results of Comparisons between MMixRIMs for Two-class Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D1C2E	M2PL	9	**98**	**97**	19	43
	2D2CMMixRIM	**73**	2	3	**76**	**54**
	2D3CMMixRIM	18	0	0	5	3
D1C2D	M2PL	2	**100**	**99**	8	40
	2D2CMMixRIM	**85**	0	1	**86**	**55**
	2D3CMMixRIM	13	0	0	6	5

Note. 2D2CMMixRIM = Two-dimensional and two-class MMixRIM; 2D3CMMixRIM = Two-dimensional and three-class MMixRIM. The largest number is bold.

of the selection of the four-dimensional M2PL model based on AIC_c were slightly smaller than those based on AIC, ranging between 56% and 83%. On the other hand, BIC tended to suggest the two- or three-dimensional M2PL models as the best-fitting model. For D2C2R0E, D2C2R0D, D2C2R3D, and D2C2R6D, the three-dimensional M2PL was mostly suggested as the best-fitting model based on BIC. For D2C2R3E and D2C2E6E, the two-dimensional M2PL was favored by BIC. CAIC mainly suggested the three-dimensional M2PL model for all conditions. The percentages of the selection of the three-dimensional M2PL model were between 62% and 89%. ABIC preferred the four-dimensional M2PL model as the best-fitting model for D2C2R0E, D2C2R0D, D2C2R3D, and D2C2R6D. For D2C2R3E and D2C2R6E, the three-dimensional M2PL model was mostly selected.

Overall, when the M2PL models were applied to the data sets of the two-dimensional and two-class condition, more than three-dimensional M2PL models were selected as the best-fitting model, except for the D2C2R3E and D2C2R6E conditions. That is, the multidimensionality caused by the two-dimensional traits and the heterogeneous population

observed as additional dimensions when the heterogeneous population were not considered in the model.

Table 6.8 presents the numbers of replications that have been selected as the best-fitting model when Mix2PL models were applied to the data sets of the two-dimensional and two-class conditions. All five information criterion indices dominantly suggested two-class Mix2PL model as the best-fitting model for all conditions. The percentages of the selection of two-class Mix2PL model based on AIC were slightly lower than those based on other indices for all conditions. Consequently, additional classes were not observed for the data sets two-dimensional and two-class conditions, although the applied model (i.e., Mix2PL model) did not consider multidimensional traits within a latent class.

Based on these results, the two-class Mix2PL model might be appropriate to analyze the data sets of the two-dimensional and two-class conditions. For further understanding about the performances of Mix2PL model and MMixRIM, these two models were compared based on the five information criterion indices, in particular for the two-class conditions (i.e., D1C2E and D1C2D) and the two-dimensional and two-class conditions (i.2., D2C2R0E, D2C2R3E, D2C2R6E, D2C2R0D, D2C2R3D, and D2C2R6D). The results of model comparisons are presented in Table 6.9.

According to Table 6.9, the model selections were consistent based on all information criterion indices, except for BIC. Selection of the two-class Mix2PL model was mostly better than for the two-dimensional and two-class MMixRIM for the two-class condition data sets. On the other hand, the two-dimensional and two-class MMixRIM model was better than the two-class Mix2PL model for the data sets of the two-dimensional and two-class conditions. Unlike other indices, BIC suggested the two-dimensional and two-class MMixRIM in fewer replications of the data sets for the two-dimensional and two-class conditions, particularly in D2C2R6D. Although the two-class Mix2PL model could perform well for the data sets of the two-dimensional and two-class, two-dimensional and two-class MMixRIM would

Table 6.7: Results of Comparisons between M2PL models for Two-dimensional and Two-class Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D2C2R0E	2PL	0	0	0	0	0
	2DM2PL	0	44	23	0	0
	3DM2PL	16	**55**	**73**	35	53
	4DM2PL	**84**	1	4	**65**	**47**
D2C2R3E	2PL	0	0	0	0	0
	2DM2PL	0	**59**	34	0	0
	3DM2PL	26	39	**62**	38	**62**
	4DM2PL	**74**	2	4	**62**	38
D2C2R6E	2PL	0	0	0	0	0
	2DM2PL	0	**56**	36	0	1
	3DM2PL	24	44	**64**	44	**66**
	4DM2PL	**76**	0	0	**56**	33
D2C2R0D	2PL	0	0	0	0	0
	2DM2PL	0	21	7	0	0
	3DM2PL	8	**77**	**85**	17	31
	4DM2PL	**92**	2	8	**83**	**69**
D2C2R3D	2PL	0	0	0	0	0
	2DM2PL	0	27	7	0	0
	3DM2PL	12	**73**	**89**	26	38
	4DM2PL	**88**	0	4	**74**	**62**
D2C2R6D	2PL	0	0	0	0	0
	2DM2PL	0	30	13	0	1
	3DM2PL	11	**68**	**83**	30	45
	4DM2PL	**89**	2	4	**70**	**55**

Note. 2DM2PL = Two-dimensional M2PL; 3DM2PL = Three-dimensional M2PL; 4DM2PL = Four-dimensional M2PL. The largest number is bold.

Table 6.8: Results of Comparisons between Mix2PL Models for Two-dimensional and Two-class Conditions

Condition	Model	AIC	BIC	CAIC	AIC_c	ABIC
D2C2R0E	2PL	0	0	0	0	2
	2CMix2PL	**61**	**100**	**99**	**73**	**85**
	3CMix2PL	35	0	1	27	15
	4CMix2PL	4	0	0	0	0
D2C2R3E	2PL	0	0	0	0	2
	2CMix2PL	**80**	**100**	**100**	**93**	**97**
	3CMix2PL	16	0	1	7	3
	4CMix2PL	4	0	0	0	0
D2C2R6E	2PL	0	0	0	0	2
	2CMix2PL	**92**	**100**	**100**	**96**	**98**
	3CMix2PL	6	0	1	3	1
	4CMix2Pl	2	0	0	1	0
D2C2R0D	2PL	0	0	0	0	2
	2CMix2PL	**71**	**99**	**99**	**82**	**87**
	3CMix2PL	23	1	1	18	13
	4CMix2PM	6	0	0	0	0
D2C2R3D	2PL	0	0	0	0	2
	2CMix2PL	**82**	**98**	**98**	**89**	**93**
	3CMix2PL	17	2	2	11	7
	4CMix2PL	4	0	0	0	0
D2C2R6D	2PL	0	0	0	0	2
	2CMix2PL	**94**	**99**	**99**	**97**	**99**
	3CMix2PL	6	1	1	3	1
	4CMix2PL	2	0	0	1	0

Note. 2CMix2PL = Two-class Mix2PL; 3CMix2PL = Three-class Mix2PL; 4CMix2PL = Four-class Mix2PL. The largest number is bold.

Table 6.9: Comparisons of 2CMix2PL and 2D2CMMixRIM

Condition	AIC		BIC		CAIC		AIC_c		ABIC	
	2CMix2PL	2D2CMMixRIM	2CMix2PL	2D2CMMixRIM	2CMix2PL	2D2CMMixRIM	2CMix2PL	2D2CMMixRIM	2CMix2PL	2D2CMMixRIM
D1C2E	**76**	24	**99**	1	**99**	1	**91**	9	**96**	4
D1C2D	**83**	17	**99**	1	**98**	2	**91**	9	**96**	4
D2C2R0E	0	**100**	3	**97**	0	**100**	0	**100**	0	**100**
D2C2R3E	0	**100**	16	**84**	0	**98**	2	**100**	0	**100**
D2C2R6E	0	**100**	48	**52**	11	**89**	0	**100**	0	**100**
D2C2R0D	0	**100**	12	**88**	0	**100**	0	**100**	0	**100**
D2C2R3D	0	**100**	29	**81**	8	**92**	0	**100**	0	**100**
D2C2R6D	0	**100**	**63**	37	22	**89**	0	**100**	1	**99**

Note. The largest value is bold.

be more appropriate for these data sets based on the model comparisons of these two models.

6.3 Recovery analysis

The recovery analyses of model parameters were conducted to assess the performance of the MMixRIM model compared with the performance of the M2PL and Mix2PL models. For the recovery of item parameters, BIAS, root mean square error (RMSE), and correlations between the true item parameters and the estimated item parameters were used. For the heterogeneous population conditions (i.e., D1C2E, D1C2D, D2C2R0E, D2C2R3E, D2C2R6E, D2C2R0D, D2C2R3D, and D2C2R6D), the recovery of class memberships was evaluated by using the percentage of cases assigned to the same latent classes to its generated latent class membership. Before the recovery analysis, label switching was monitored for the heterogeneous population conditions for MixIRT models applied to these conditions. In a MixIRT model, classes are latent, not manifest, and it is possible that the meaning of each latent class may differ in each replication. That is, the characteristics of Class 1 in the first replication might be observed as the characteristics of Class 2 in the second replication. This situation is called label switching. In this simulation study, label switching was monitored by comparing the true item parameters and the estimated item parameters, and corrected when observed. After correcting label switching, the estimated parameters were transformed to the scale of the true parameters as explained in the previous chapter.

6.3.1 Recovery analysis of item parameters

BIAS and RMSE were close to zero and correlations were close to one. The means and standard deviations for BIAS, RMSE, and correlations between the true (i.e., generating) item parameters and estimated item parameters over the 100 replications are reported in Tables 6.10 to 6.12 and Figures 6.7 to 6.9.

Table 6.10: Results of Recovery Analyses for Item Discrimination(a)/ Slopes Estimates for the First Dimension(a_1)

Model	Condition	BIAS		RMSE		Correlation	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
2PL	D1	0.000	(0.000)	0.067	(0.009)	**0.630**	(0.096)
Mix2PL	D1C2E	0.000	(0.000)	0.138	(0.014)	0.850	(0.031)
	D1C2D	0.000	(0.000)	0.147	(0.016)	0.830	(0.031)
M2PL	D2BR0	0.011	(0.018)	0.093	(0.013)	0.977	(0.006)
	D2BR3	0.002	(0.020)	0.110	(0.021)	0.968	(0.018)
	D2BR6	−0.002	(0.027)	0.164	(0.027)	0.933	(0.036)
	D2WR0	0.011	(0.019)	0.099	(0.021)	0.966	(0.022)
	D2WR3	0.004	(0.022)	0.115	(0.015)	0.956	(0.012)
	D2WR6	−0.001	(0.031)	0.182	(0.023)	0.902	(0.028)
MMixRIM	D2C2R0E	0.032	(0.038)	0.208	(0.026)	0.815	(0.044)
	D2C2R3E	0.017	(0.048)	0.241	(0.039)	0.809	(0.042)
	D2C2R6E	0.001	(0.077)	**0.389**	(0.085)	**0.780**	(0.050)
	D2C2R0D	0.035	(0.041)	0.241	(0.032)	0.824	(0.039)
	D2C2R3D	0.021	(0.072)	0.294	(0.119)	0.806	(0.042)
	D2C2R6D	−0.003	(0.087)	**0.424**	(0.090)	**0.759**	(0.048)

Note. When RMSE is larger than .3 or correlation is less than .8, the value is bold.

As reported in Table 6.10, BIASs for item discriminations of the unidimensional (i.e., D1) and the two-class conditions (i.e., D1C2E and D1C2D) were all zero. For estimated slopes for the first dimension by M2PL models for two-dimensional conditions (i.e., D2BR0, D2BR3, D2BR6, D2WR0, D2WR3, and D2WR6), BIASs were between 0.00 and 0.01, and decreased as correlations between dimensions increased. The last six rows in Table 6.10 present the recovery statistics for two-dimensional and two-class conditions (i.e., D2C2R0E, D2C3R3E, D2C2R6E, D2C2R0D, D2C2D3D, and D2C2R6D). BIASs for slopes were slightly larger than BIASs for other conditions. These values were all close to zero, ranging between 0.00 and 0.04. Additionally, RMSE for item discriminations (a) of D1 was about 0.07, those for D1C2E and D1C2D were about 0.14 and 0.15, respectively. For the slope estimates (a_1) of the two-dimensional conditions, RMSEs ranged between 0.09 and 0.18, and these RMSEs increased as the correlation increased contrary to BIASs. For the two-dimensional and two-class condition, RMSEs for the slopes were larger than other conditions, ranging between 0.21 and 0.42. Similar to RMSEs for two-dimensional conditions, these RMSEs increased as the correlation increased. The correlations were mostly larger than .8, except for some conditions. On the other hand, the correlation between true and estimated item discriminations was 0.63 for D1. It is important to note that this low correlation occurred for the very small variance of item discriminations. The variances for true and estimated item discriminations were only 0.01. It is likely that the low correlation for item discriminations did not necessarily mean poor estimation by the 2PL model. For the two-dimensional and two-class conditions, the range of correlations for slopes for the first dimension were between 0.76 and 0.82. These relatively lower correlations also might be caused by the smaller variances of parameters as observed in the unidimensional condition. For the two-dimensional and two-class conditions, Class 1 was assumed to be a group of students who did not learn about the latent trait measured on the second dimension, and Class 2 was assumed as a group of students who did not learn about the latent trait on the first dimension. Consequently, the slope parameters of items that measure the latent trait on the second dimension were close to zero with a

small variation for Class 1. On the other hand, for Class 2, the slope parameters of items that measure the latent trait on the first dimension were close to zero (see Figure 5.2).

Table 6.11: Results of Recovery Analyses for Slopes Estimates for the Second Dimension(a_2)

Model	Condition	BIAS		RMSE		Correlation	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
2PL	D1	-	-	-	-	-	-
Mix2PL	D1C2E	-	-	-	-	-	-
	D1C2D	-	-	-	-	-	-
M2PL	D2BR0	0.012	(0.019)	0.092	(0.013)	0.976	(0.007)
	D2BR3	−0.002	(0.020)	0.114	(0.022)	0.961	(0.025)
	D2BR6	−0.009	(0.028)	0.176	(0.029)	0.915	(0.042)
	D2WR0	0.009	(0.018)	0.095	(0.022)	0.966	(0.028)
	D2WR3	−0.001	(0.019)	0.111	(0.014)	0.956	(0.013)
	D2WR6	−0.006	(0.028)	0.180	(0.021)	0.897	(0.031)
MMixRIM	D2C2R0E	0.025	(0.037)	0.205	(0.026)	0.814	(0.048)
	D2C2R3E	0.011	(0.051)	0.240	(0.042)	0.812	(0.046)
	D2C2R6E	0.007	(0.078)	**0.392**	(0.089)	**0.782**	(0.048)
	D2C2R0D	0.046	(0.053)	0.230	(0.068)	**0.774**	(0.055)
	D2C2R3D	0.020	(0.081)	0.285	(0.159)	**0.775**	(0.054)
	D2C2R6D	0.018	(0.096)	**0.414**	(0.127)	**0.755**	(0.053)

Note. When RMSE is larger than .3 or correlation is less than .8, the value is bold.

Table 6.11 presents the recovery statistics for the slope estimates for the second dimensions. Thus, the unidimensional condition and the two-class conditions were not available to calculate the recovery statistics. All BIASs and RMSEs were very close to those for the slope estimates for the first dimensions in Table 6.10. Again, BIASs decreased, while

RMSEs increased as the correlation between dimensions increased for all conditions. BIASs and RMSEs for the two-dimensional and two-class conditions were slightly larger than those for the two-dimensional conditions. All correlations between true and estimated slopes for the second dimension were larger than .9 for the two-dimensional conditions. For the two-dimensional and two-class conditions, however, the correlations were between 0.76 and 0.81. As explained above, these relatively weaker correlations for these conditions might relate to the small variances of slope parameters.

As can be seen from Table 6.12, BIASs for the item difficulties of D1, D1C2E, and D1C2D were all zero. RMSE for the item difficulties of D1 was 0.08, and this was the smallest among the 15 conditions. For D1C2E and D1C2D, RMSEs were about 0.14 and 0.14, respectively. For the two-dimensional conditions, BIASs for intercepts were all negative value, ranging between -0.03 and -0.04. This means that the intercept parameters were underestimated by M2PL model regardless of the multidimensional structure and the degree of correlations between dimensions. Similar to BIASs, the multidimensional structure or correlations between dimensions did not affect RMSEs for the intercepts. Most RMSEs for intercepts were similarly close to 0.16. For the two-dimensional and two-class conditions, each recovery statistics for the intercepts did not vary across the six conditions. BIASs for intercepts were very close to zero, ranging between -0.01 and 0.00. RMSEs ranged between 0.24 and 0.28, and the correlations between the true and estimated item difficulties or intercepts for the all conditions were between 0.94 and 0.95.

BIASs, RMSEs, and correlations for item parameters are compared in Figures 6.7 to 6.9, respectively. Based on BIASs in Figure 6.7, the estimated intercept parameters by M2PL model tended to underestimate, while estimated slop parameters by MMixRIM tended to overestimate when dimensions were independent. According to RMSEs in Figure 6.8, the accuracy of estimation decreased as the complexity of the model increased. In particular, RMSEs for slope parameters increased as the correlation between dimensions decreased. As shown in Figure 6.9, the correlations for item difficulties and intercept parameters were

Table 6.12: Results Recovery Analyses for Item Difficulties(b)/ Intercepts(d)

Model	Condition	BIAS		RMSE		Correlation	
		Mean	(SD)	Mean	(SD)	Mean	(SD)
2PL	D1	0.000	(0.000)	0.080	(0.012)	0.993	(0.000)
Mix2PL	D1C2E	0.000	(0.000)	0.125	(0.014)	0.993	(0.002)
	D1C2D	0.000	(0.000)	0.137	(0.040)	0.991	(0.010)
M2PL	D2BR0	−0.036	(0.015)	0.162	(0.007)	0.972	(0.003)
	D2BR3	−0.036	(0.016)	0.163	(0.007)	0.972	(0.003)
	D2BR6	−0.036	(0.019)	0.163	(0.008)	0.972	(0.004)
	D2WR0	−0.035	(0.014)	0.162	(0.006)	0.972	(0.002)
	D2WR3	−0.034	(0.014)	0.162	(0.007)	0.972	(0.003)
	D2WR6	−0.034	(0.014)	0.162	(0.007)	0.972	(0.003)
MMixRIM	D2C2R0E	0.001	(0.018)	0.240	(0.013)	0.956	(0.004)
	D2C2R3E	−0.003	(0.017)	0.238	(0.014)	0.956	(0.005)
	D2C2R6E	−0.007	(0.018)	0.236	(0.014)	0.956	(0.005)
	D2C2R0D	0.004	(0.033)	0.284	(0.043)	0.943	(0.011)
	D2C2R3D	−0.002	(0.045)	0.283	(0.068)	0.941	(0.031)
	D2C2R6D	−0.007	(0.032)	0.268	(0.032)	0.946	(0.011)

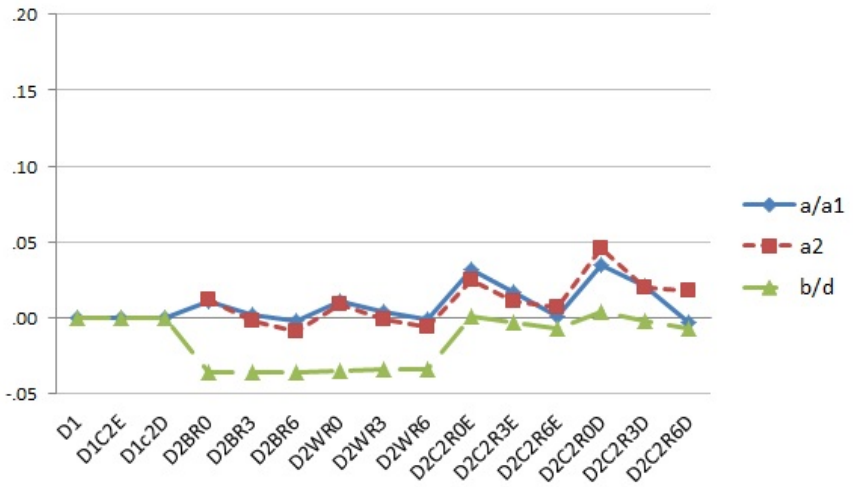

Figure 6.7: BIASs for item parameter estimates.

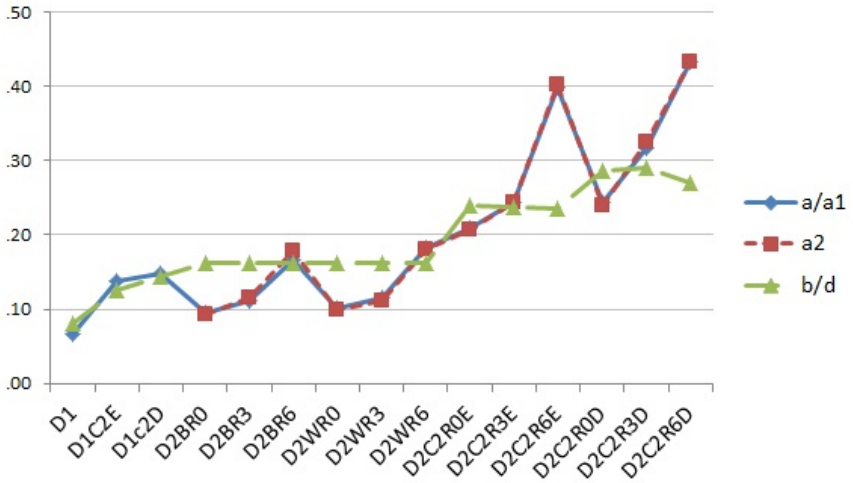

Figure 6.8: RMSEs for item parameter estimates.

similar across all conditions, but a more complicated model made a weaker correlation for slope parameters, except for D1, D1C2E and D1C2D. The relatively weak correlations for these three conditions results from the small variances of parameters rather than the

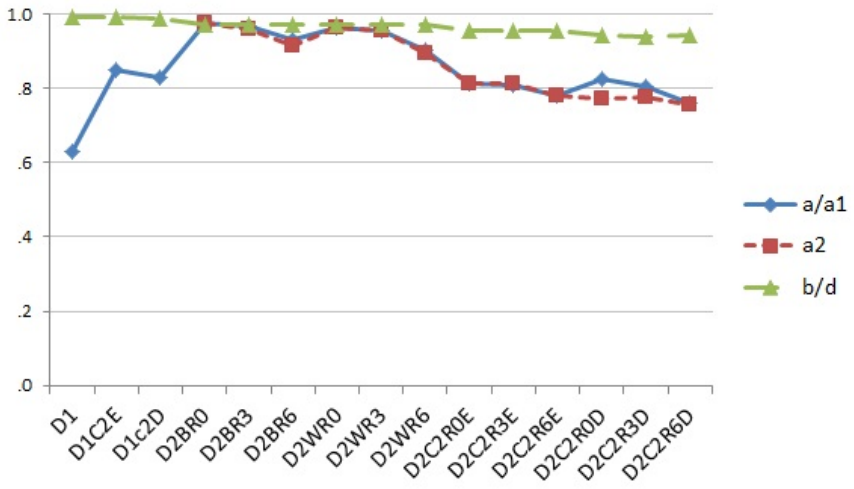

Figure 6.9: Correlations for item parameter estimates.

accuracy of estimation.

6.3.2 Recovery analysis of memberships

The recovery analysis of the latent class memberships was conducted for the heterogenous population conditions, and the results are presented in Table 6.13 and Figure 6.10. Table 6.13 reported the average percentages of correct detection of the class membership over 100 replications for each condition. For the two-class conditions, most cases were assigned to their true class by Mix2PL model, regardless of the size of latent classes. On average, the percentage of the correct detectiong of class membership was 98% for both D1C2E and D1C2D. The average percentages of the correct detection of class membership for the two-dimensional and two-class conditions were lower than the two-class conditions. Particularly, the percentages of correct detections differed depending on the size of latent classes. About 75% of cases were correctly assigned to their true class when the sizes of classes were equal (i.e., D2C2R0E, D2C2R3E, and D2C2R6E), and about 80% of cases were

correctly assigned to the true class when the size of one class was larger than another class (i.e., D2C2R0D, D2C2R3D, and D2C2R6D).

Table 6.13: Percentage of Correct Detect Class Membership

Model	Condition	Mean	(SD)
Mix2PL	D1C2E	.978	(.005)
	D1C2D	.982	(.005)
MMixRIM	D2C2R0E	.743	(.014)
	D2C2R3E	.746	(.015)
	D2C2R6E	.752	(.014)
	D2C2R0D	.801	(.013)
	D2C2R3D	.803	(.014)
	D2C2R6D	.806	(.014)

6.3.3 Effects of data structures on the estimation of model parameters

To examine the effect of different dimensional structure on the estimation of model parameters, a two-level linear model analysis was conducted because the 100 replications were nested within each condition. Moreover, the the variability of recovery statistics within each condition might be smaller than the variability of those between the conditions. BIASs, RMSEs, and Pearson correlations used as dependent variables in each two-level linear model. The type of distributions of ability, the type of multidimensionality, the size of latent class, and the correlation between dimensions were used as independent variables. In this simulation study, the 15 simulation conditions were based on different combinations of simulation factors. That is, the 15 conditions were basically based on the four different distributions of ability (i.e., unidimensional, two-dimensional, mixture with two latent classes, and two-dimensional and

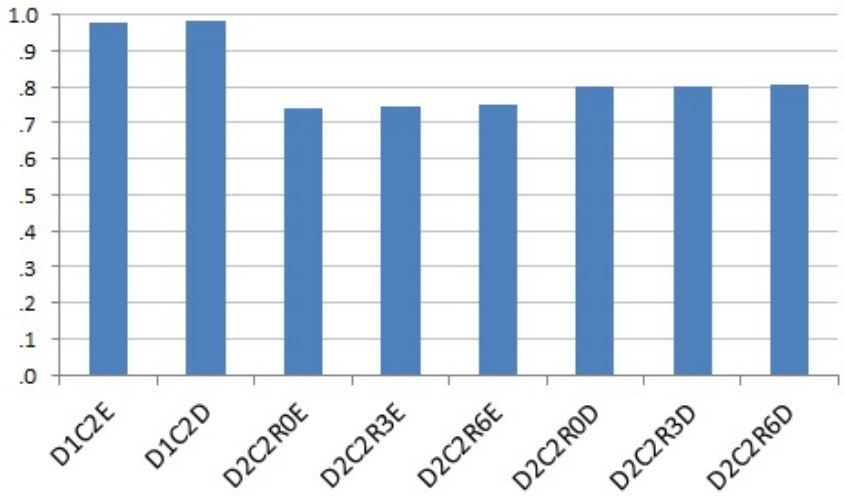

Figure 6.10: Percentages of correct eetection of class membership.

two-class), and the other three factors depended on the type of distributions of ability. For example, the size of latent class was applied to manipulate the two-class conditions (i.e., D1C2E versus D1C2D) and the two-dimensional and two-class conditions (i.e., D2C2R0E, D2C2R3E, and D2C2R6E versus D2C2R0D, D2C2R3D, and D2C2R6D), while the type of multidimensionality was applied only for the two-dimensional conditions (i.e., D2BR0, D2BR3, and D2BR6 versus D2WR0, D2WR3, and D2WR6). Accordingly, the type of distributions of ability was available to all replications, but type of multidimensionality, latent class sizes and correlation between dimensions were not available for some conditions. For these reasons, the type of distributions of ability was used as a level-1 independent variable, and other three simulation factors (i.e., latent class size, type of multidimensionality, and correlation between dimensions) were used as level-2 independent variables to explain differences between type of distributions of ability. The two-level structure of the relationship between factors is represented in Figure 6.11.

Variable Recording. Unlikely RMSEs and correlations, the possible range of BIAS is between a negative infinite to a positive infinite. Thus, for the easy interpretation, the abso-

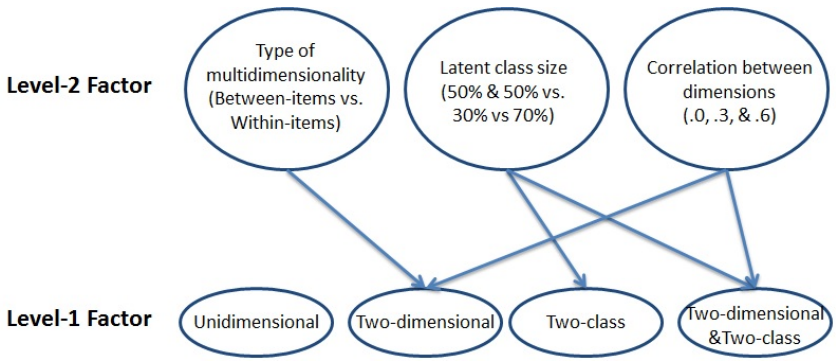

Figure 6.11: Two-level structure of simulation factors.

lute values of BIASs were used as a dependent variable. Since all independent variables are categorical variables, these variables were recoded by using dummy coding. Three dummy variables were used for the type of distributions of ability, and each dummy variable represents the two-dimensional ($2D$), two-class ($2C$), and two-dimensional and two-class ($2D2C$) conditions, respectively. Thus, the unidimensional condition was set as a reference group. One dummy variable ($SIZE$) was used for the latent class size because this had two categories. The case of equal size classes was recoded as zero, and the case of one dominant class was recoded as one so that the case of equal size classes can be a reference group. Similarly, the case of the between-items multidimensionality was recoded as zero, and the case of the within-items multidimensionality was recoded as one. Two dummy variables ($CORR_3$ and $CORR_6$) were generated for the correlation between dimensions. A total of 7 dummy variables (i.e., $2D$, $2C$, $2D2C$, $SIZE$, $STRUC$, $CORR_3$, and $CORR_6$) were used as independent in each two-level linear model. The two-level linear model can be formulated as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} \times 2D_{ij} + \beta_{2j} \times 2C_{ij} + \beta_{3j} \times 2D2C_{ij} + e_{ij}, \quad (6.1)$$

$$\beta_{0j} = \gamma_{00} + u_j, \quad (6.2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} \times STRUC_j + \gamma_{12} \times CORR_{3j} + \gamma_{13} \times CORR_{6j}, \quad (6.3)$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} \times SIZE_j, \quad (6.4)$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31} \times SIZE_j + \gamma_{32} \times CORR_{3j} + \gamma_{33} \times CORR_{6j}, \quad (6.5)$$

where Y_{ij} indicates the recovery statistics for the ith replication of the jth condition, β_{0j} is a random intercept, β_{1j}, β_{2j} , and β_{3j} are fixed slopes, e_{ij} indicates the level-1 residuals, and u_j is the level-2 residuals. Equation 6.2 is the equation for the random intercept, and Equations 6.2 to 6.5 are equations for the fixed slopes. In Equation 6.2, γ_{00} indicates the mean of a reference group, which means the mean of recovery statistics of the unidimensional condition (i.d., D1). In Equation 6.3, γ_{10} indicates the effect of the two-dimensional conditions, which means the average difference of recovery statistics of the two-dimensional conditions from those of the unidimensional condition. γ_{11} is the coefficient for the interaction between $2D$ and $STRUC$, which means the difference of the effect of the two-dimensional conditions by the type of multidimensionality. Similarly, γ_{12} and γ_{13} represent the coefficients of interactions between $2D$ and $CORR_3$ and between $2D$ and $CORR_6$, respectively. As mentioned above, $SIZE$ was not included in Equation 6.3 because the latent class size was not available for the two-dimensional conditions. In Equation 6.4, γ_{20} represents the average difference of the recovery statistics of the two-class conditions from those of the unidimensional condition, and γ_{21} indicates the coefficient of the interaction between $2C$ and $SIZE$. In Equation 6.5, γ_{30} means the average difference of recovery statistics of the two-dimensional and two-class conditions from those of the unidimensional condition, and γ_{31}, γ_{32}, and γ_{33} indicate the coefficients of interactions between $2D2C$ and $SIZE$, $CORR_3$, and $CORR_6$, respectively.

Assumption Test. For the correct inferences of the results from the multilevel linear model, a number of assumptions should be satisfied. First assumption is the linear depen-

dence of dependent variable on the independent variables. Another assumption is the homoscedastic normal distributions for the residuals (Sniger & Bosker, 2012). As reported in the previous section, correlations for the item difficulty (or intercept) estimates were very close to one in all simulation conditions. Also, the average percentages of correct detect class membership of 100 replications for each condition were not have a linear relationship with simulation conditions. Consequently, these two recovery statistics were not included to the multilevel linear model analyses. The homoscedastic normal distributions for the level-1 residuals was tested by using a normal probability plot of standardized residuals called a P-P plot and a scatter plot of the standardized residuals against the predicted values. Since raw BIASs, RMSEs, and correlations were violated the homoscedastic normal distribution assumption, these values were transformed. The natural logarithm transformation was applied to transform BIASs and RMSEs because these had a positive skewness (Howell, 2007). For correlations, the Fisher's logarithmic transformation was used because correlations also positively skewed, between .7 and 1.0 (Blommers & Forsyth, 1977). That is, $log(|BIAS|)$, $log(RMSE)$, and z_r were used in the two-level linear model analyses. Figures 6.12 to 6.14 are plots for testing the homoscedastic normal distribution of the level-1 residuals.

Figure 6.12 contains the plots for the assumption test of the item discriminations (a)/ slopes for the first dimension (a_1). Similarly, Figure 6.13 and Figure 6.14 are for the assumption test of the slops for the second dimension (a_2) and the item difficulties (b)/ intercepts(d), respectively. In Figure 6.12 and Figure 6.13, the two plots on the left-side are plots for BIASs, other two plots on the middle are for RMSEs, and another two plots are for the correlations. Because the correlations for the item difficulties and intercepts were not included to the two-level linear model analyses, the two plots on the left-side in Figure 6.14 are for BIASs, and another two-plots on the right-side in Figure 6.14 are for RMSEs.

The scatter plots were used to test the homogeneous variance of the level-1 residual by comparing the standardized residuals and the predicted values. The P-P plots indicates the

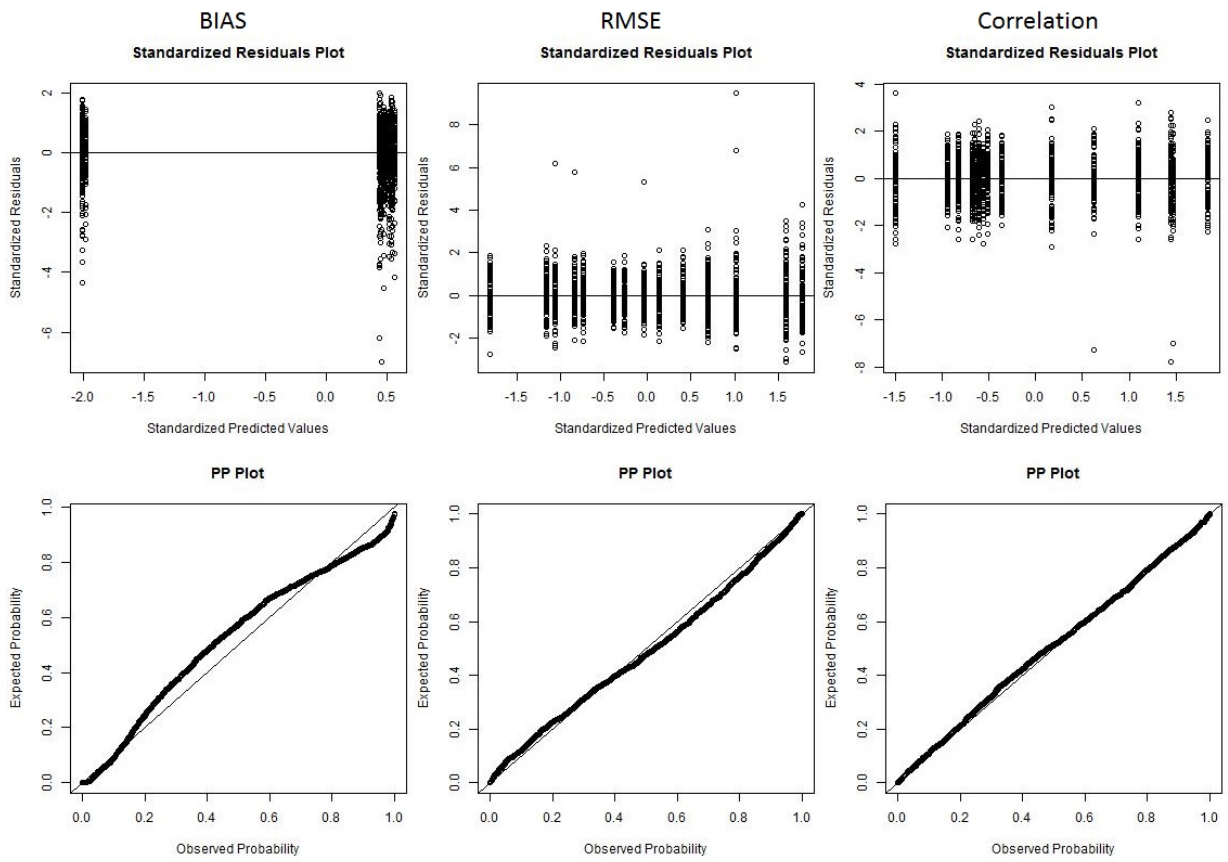

Figure 6.12: Scatter plots and P-P plots of standardized level-1 residuals of recovery statistics for item discrimination(a)/slopes for the first dimension(a_1).

normality of the level-1 residuals. As the pairs of the observed and expected probability of the standardized residual are close to the diagonal line, as the distribution of the level-1 residuals is close to normal. According to Figures 6.12 to 6.14, the the level-1 residuals were moderately homoscedastic normal.

Effects of Dimension Structure on Estimation of Item Discriminations and Slopes for the First Dimension. Table 6.14 presents the results of the two-level linear model analyses for the recovery statistics of the item discriminations (a)/ slopes for the first dimension (a_1). As explained previous, the reference group is the unidimensional condition. Accordingly, the intercept means the average recovery statistics of the unidimensional condi-

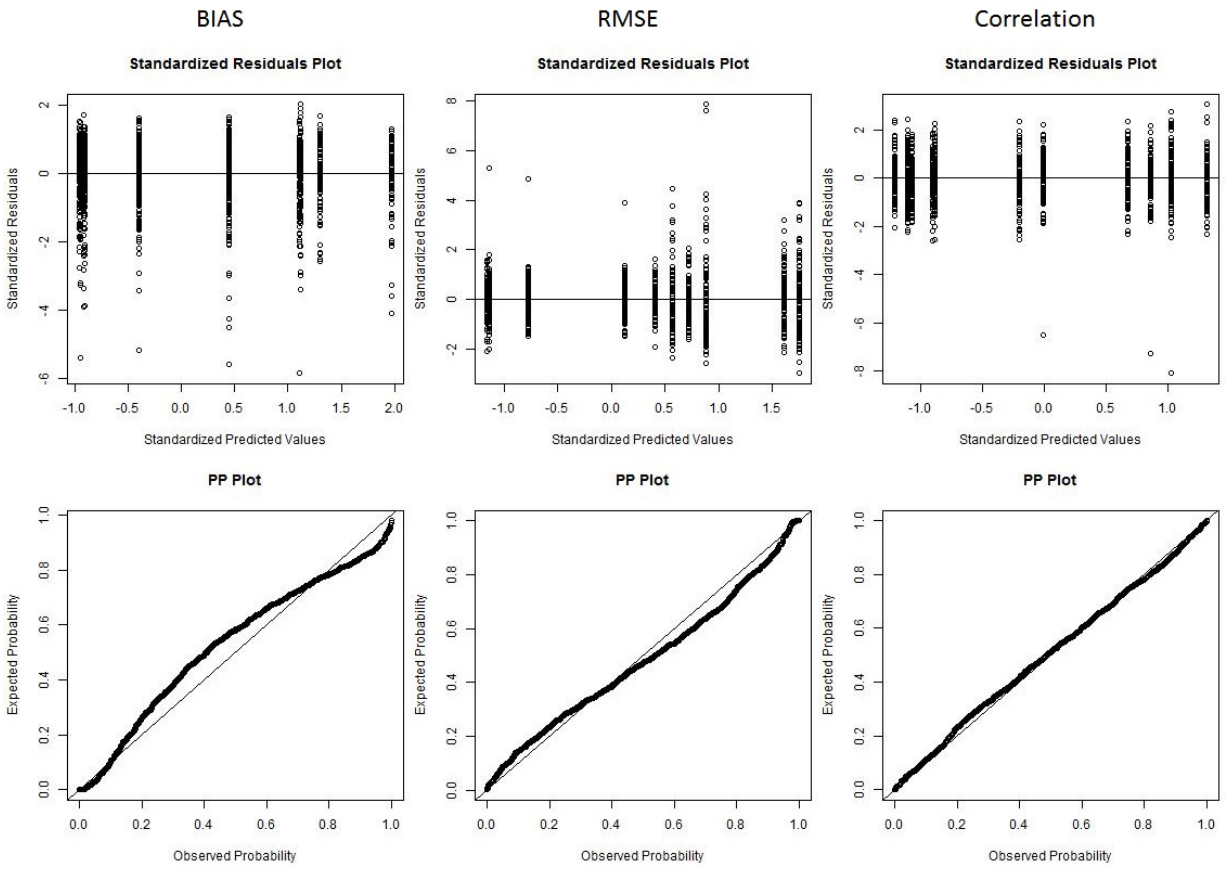

Figure 6.13: Scatter plots and P-P plots of standardized level-1 residuals of recovery statistics for slopes for the second dimension(a_2).

tion, and the fixed effects can be interpreted as the differences of the average recovery statistics of each condition from the the unidimensional condition. For example, the coefficient $2D$ indicates the difference between the average recovery statistics of the two-dimensional condition from the unidimensional condition. That is, the effect of the two-dimensional ability distribution on the estimation of the model parameters. Similarly, the coefficient of the interaction between the two-dimensional and two-class condition and the the latent class size (i.e., $2D2C \times SIZE$) indicates the difference of average recovery statistics when the sizes of latent classes were equal from those when there was a dominant class within six the two-dimensional

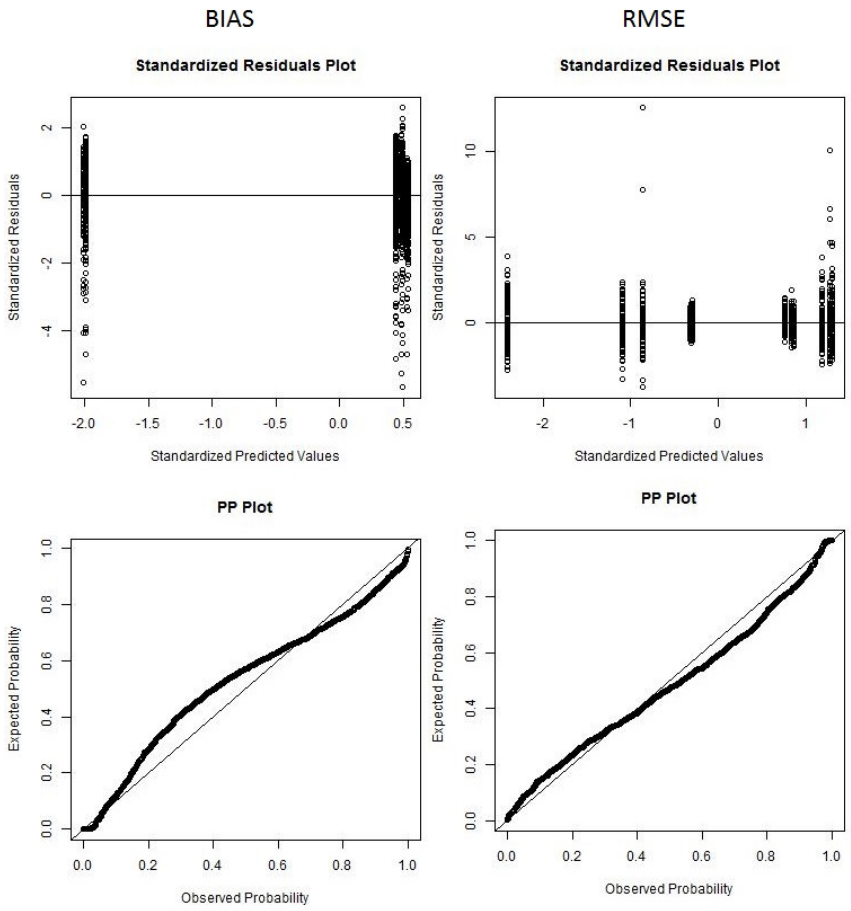

Figure 6.14: Scatter lots and P-P plots of standardized level-1 residuals of recovery statistics for item difficulty(b)/intercept(d).

and two-class conditions. That is, the effect of latent class size on the estimation of model parameter for the two-dimensional and two-class conditions.

According to Table 6.14, BIASs of the two-dimensional conditions and of the two-dimensional and two-class conditions were significantly larger than those of the unidimensional condition (Estimate for $2D = 31.67$, $p < .001$ and Estimate for $2D2C = 32.48$, $p < .001$), but BIASs of the two-class conditions were significantly smaller than those of the unidimensional condition (Estimate of $2C =$ -0.32, $p < .05$). The effect of the type of multidimensionality on BIASs of the two-dimensional conditions was not significant. Also, there was

Table 6.14: Results of Two-level Linear Model Analyses for Item Discriminations(a) and Slopes for the First Dimension(a_1)

	$\log(\lvert BIAS \rvert)$		$\log(RMSE)$		z_r	
	Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects						
Intercept	−36.043 ***	0.104	−2.718***	0.042	0.756***	0.077
$2D$	31.672***	0.138	0.335***	0.056	1.496***	0.102
$2C$	−0.324*	0.147	0.731***	0.060	0.509***	0.109
$2D2C$	32.481***	0.132	1.145***	0.055	0.418***	0.099
$2D \times STRUC$	−0.134	0.104	0.050	0.042	−0.174*	0.077
$2D \times CORR_3$	−0.103	0.104	0.166***	0.042	−0.161*	0.077
$2D \times CORR_6$	0.276	0.116	0.617***	0.045	−0.645***	0.086
$2C \times SIZE$	0.023	0.147	0.062	0.060	−0.065	0.109
$2D2C \times SIZE$	0.200*	0.085	0.137***	0.035	−0.014	0.063
$2D2C \times CORR_3$	0.053	0.104	0.155***	0.042	−0.037	0.077
$2D2C \times CORR_6$	0.359***	0.104	0.582***	0.042	−0.136	0.077
Random effects						
Intercept	0.000	0.000	0.002	0.001	0.006	0.000
Level-1	1.083***	0.040	0.024***	0.001	0.019***	0.000

Note. * $p < .05$; ** $p < .01$; *** $p < .001$. Reference group of $2D$, $2C$ and $2D2C$: Unidimensional ability; Reference group of $STRUC$: Between-items multidimensionality; Reference group of $SIZE$: Equal size two classes; and Reference group of $CORR_3$ and $CORR_6$: Independent dimensions.

no significant effect of the correlation between dimensions on BIASs of the two-dimensional conditions. The effect of latent class size on BIASs of the two-class conditions also was not statistically significant, while it had a significant influence on BIASs of the two-dimensional and two-class conditions (Estimate of $2D2C \times SIZE = 0.20$, $p < .05$). Additionally, BIASs of the two-dimensional and two-class conditions with highly correlated dimensions (Estimate of $2D2C \times CORR_6 = 0.36$, $p < .001$) were significantly larger than those with independent dimensions. For the random effects, the variance of intercept became zero after adding independent variables, whereas this differed significantly in the unconditional model.

Similar to BIASs, the effects of three distributions of abilities on RMSEs were significant (Estimate of $2D = 0.34$, $p < .001$; Estimate of $2C = 0.73$, $p < .001$; and Estimate of $2D2C = 1.14$, $p < .001$), but the direction of the effect of 2C was opposite. That is, RMSEs of the unidimensional condition were significantly smaller than RMSEs of other distributions of abilities. RMSEs of the two-dimensional conditions significantly increased as the correlations between dimensions increased (Estimate of $2D \times CORR_3 = 0.17$, $p < .001$ and Estimate of $2D \times CORR_6 = 0.62$, $p < .001$). RMSEs of the two-dimensional and two-class conditions also increased as the correlations between dimensions increased (Estimate of $2D2C \times CORR_3 = 0.16$, $p < .001$ and Estimate of $2D2C \times CORR_6 = 0.58$, $p < .001$). Moreover, RMSEs of two-dimensional and two-class conditions with one dominant class (Estimate of $2D2C \times SIZE = 0.14$, $p < .001$) were were significantly larger than those with equal size classes.

Three different types of distributions had significantly positive effects on the correlations (Estimate of $2D = 1.50$, $p < .001$; Estimate of $2C = 0.51$, $p < .001$; and Estimate of $2D2C = 0.42$, $p < .001$). The correlations for the two-dimensional data sets with the within-items multidimensionality were significantly lower than those for the two-dimensional data sets with the between items multidimensionality (Estimate of $2D \times STRUC$ = -0.17, $p < .05$). Moreover the correlations for the two-dimensional data sets with independent dimensions were significantly larger than those for the two-dimensional data sets with weakly or strongly correlated dimensions (Estimate of $2D \times CORR_3$ = -0.16, $p < .05$ and Estimate of $2D \times$

$CORR_6$ = -0.65, $p < .001$). There was no significant effect of other simulation factors on the correlations for the two-class data sets and the two-dimensional and two-class data sets.

Effects of Dimension Structure on Estimation of Slopes for the Second Dimension. Table 6.15 presents the results of the two-level linear model analyses for the slope parameters of the second dimension(a_2). Because only the two-dimensional and the two-dimensional and two-class conditions had the slopes of the second dimension, the main effect of $2C$ and the interactions between $2C$ and other possible independent variables were not included to the model. Therefore, the intercept in these models means the average recovery statistics of the between-items multidimensionality with two independent dimensions (i.e., D2BR0). BIASs of the two-dimensional and two-class conditions were significantly larger than BIASs of D2BR0 (Estimate of $2D2C = 0.79$, $p < .001$). There was no significant effect of the multidimensional type on BIASs of two-dimensional conditions, but the BIASs of the conditions with highly correlated two dimensions were significantly larger than those with independent two dimensions (Estimate of $2D \times CORR_6 = 0.30$, $p < .001$). Also, for the two-dimensional and two-class conditions, BIASs of the conditions with highly correlated dimensions were significantly larger than those with two independent dimensions (Estimate of $2D2C \times CORR_6 = 0.49$, $p < .001$). Additionally, the effect of a dominant class on BIASs of the two-dimensional and two-class conditions was significant (Estimate of $2D2C \times SIZE = 0.38$, $p < .001$).

For RMSEs, all fixed effects, except for $STRUC$, were significant at the alpha level .05. RMSEs of D2BR0 was significantly smaller than RMSEs of the two-dimensional and two-class conditions (Estimate of $2D2C = 0.80$, $p < .001$). RMSEs of the two-dimensional conditions increased as the correlation between dimensions increased (Estimate of $2D \times CORR_3 = 0.19$, $p < .001$ and $2D \times CORR_6 = 0.66$, $p < .001$). Similarly, RMSEs of the two-dimensional and two-class conditions increased as the correlation between dimensions increased (Estimate of $2D2C \times CORR_3 = 0.16$, $p < .001$ and $2D2C \times CORR_6 = 0.61$, $p < .001$). Moreover, RMSEs of the two-dimensional and two-class conditions with a dominant class were significantly

larger than those with equal size classes (Estimate of $2D2C \times SIZE = 0.08$, $p < .001$). The correlations of D2BR0 were significantly larger than the correlations of the two-dimensional and two-class conditions (Estimate of $2D2C$ = -1.06, $p < .001$). Similar to the slopes for the first dimension, the type of multidimensionality and the levels of correlation between dimensions had significant influence on the correlations for the two-dimensional data sets (Estimate of $STRUC$ = -0.11, $p < .01$; stimate of $2D \times CORR_3$ = -0.20, $p < .001$; and stimate of $2D \times CORR_6$ = -0.68, $p < .001$). For the two-dimensional and two-class data sets, the correlations for the two-dimensional and two-class data sets with one dominant class were significantly lower than those with equal sizes of classes (Estimate of $2D2C \times SIZE$ = -0.09, $p < .01$).

Effects of Dimension Structure on Estimation of Item Difficulties and Intercepts. As can be seen in Table 6.16, BIASs of the unidimensional condition were significantly smaller than those of the two-dimensional conditions (Estimate of $2D = 33.34$, $p < .001$) and of the two-dimensional and two-class conditions (Estimate of $2D2C = 32.12$, $p < .001$). Only a dominant class had a significant effect on BIASs of the two-dimensional and two-class conditions (Estimate of $2D2C \times SIZE$ =0.65, $p < .001$), and other simulation factors did not affect to BIASs. RMSEs of the unidimensional condition were smaller than those of other distributions of ability (Estimate of $2D = 0.71$, $p < .001$; Estimate of $2C = 0.44$, $p < .001$; and Estimate of $2D2C = 1.11$, $p < .001$). The effects of the dominant class on both RMSEs of the two-class conditions (Estimate of $2C \times SIZE = 0.08$, $p < .001$) and of the two-dimensional and two-class conditions(Estimate of $2D2C \times SIZE = 0.15$, $p < .001$) were significant. RMSEs of the two-dimensional and two-class with highly correlated dimensions were smaller than those with two independent dimensions (Estimate of $2D2C \times CORR_6$ = -0.04, $p < .05$).

Table 6.15: Results of Two-level Linear Model Analyses for Slopes for the Second Dimension(a_2)

	$\log(\lvert BIAS \rvert)$		$\log(RMSE)$		z_r	
	Estimate	SE	Estimate	SE	Estimate	SE
Fixed effects						
Intercept	-4.537^{***}	0.100	-2.389^{***}	0.019	2.209^{***}	0.036
$2D2C$	0.785^{***}	0.134	0.802^{***}	0.026	-1.064^{***}	0.050
$STRUC$	0.006	0.115	0.006	0.024	-0.114^{**}	0.042
$2D \times CORR_3$	-0.020	0.115	0.191^{***}	0.021	-0.196^{***}	0.042
$2D \times CORR_6$	0.302^{*}	0.129	0.655^{***}	0.024	-0.677^{***}	0.017
$2D2C \times SIZE$	0.382^{***}	0.094	0.077^{***}	0.018	-0.092^{**}	0.034
$2D2C \times CORR_3$	0.001	0.115	0.160^{***}	0.021	-0.003	0.041
$2D2C \times CORR_6$	0.492^{***}	0.115	0.608^{***}	0.021	-0.069	0.015
Random effects						
Intercept	0.000	0.000	0.000	0.000	0.002	0.000
Level-1	1.322^{***}	0.054	0.040^{***}	0.000	0.022^{***}	0.000

Note. * $p < .05$; ** $p < .01$; *** $p < .001$. Reference group of $2D2D$: Two-dimensional abilities; Reference group of $STRUC$: Between-items multidimensionality; Reference group of $SIZE$: Equal size two classes; and Reference group of $CORR_3$ and $CORR_6$: Independent dimensions.

Table 6.16: Results of Two-level Linear Model Analyses for Item Difficulties(b) and Intercepts(d)

	$\log(\lvert BIAS \rvert)$		$\log(RMSE)$	
	Estimate	SE	Estimate	SE
Fixed effects				
Intercept	-36.789 ***	0.095	-2.532***	0.011
$2D$	33.335***	0.125	0.712***	0.014
$2C$	0.174	0.134	0.444***	0.015
$2D2C$	32.124***	0.122	1.108***	0.014
$2D \times STRUC$	-0.026	0.095	-0.004	0.011
$2D \times CORR_3$	0.007	0.095	0.001	0.011
$2D \times CORR_6$	-0.065	0.106	-0.001	0.011
$2C \times SIZE$	0.074	0.134	0.080***	0.015
$2D2C \times SIZE$	0.645***	0.077	0.146***	0.009
$2D2C \times CORR_3$	0.030	0.095	-0.010	0.011
$2D2C \times CORR_6$	-0.060	0.095	-0.035*	0.011
Random effects				
Intercept	0.000	0.000	0.000	0.000
Level-1	0.894***	0.033	0.010***	0.000

Note. * $p < .05$; ** $p < .01$; *** $p < .001$. Reference group of $2D$, $2C$ and $2D2C$: Unidimensional ability; Reference group of $STRUC$: Between-items multidimensionality; Reference group of $SIZE$: Equal size two classes; and Reference group of $CORR_3$ and $CORR_6$: Independent dimensions.

Chapter 7

DISCUSSION

The main purpose of this study was to propose a multidimensional mixture random item model (MMixRIM) as an alternative model for analysis of multidimensional. When a MIRT model is applied to multidimensional data, the multidimensionality of data can be observed as a characteristics of an assessment. On the other hand, the multidimensionality can be understood based on a characteristics of a group of examinees in MixIRT model. For some multidimensional data, however, the characteristics of an assessment or a group of examinees might not be enough to explain the dimensional structure. It is possible that an assessment might have been developed to measure two latent traits but some groups of examinees did not have a chance to learn one of latent traits measured. This could have happened for several reasons such as because of different course schedules between schools or teachers. For the data set collected from this situation, MMixRIM would be better because this model can provide information based on characteristics of both person and assessment.

An empirical multidimensional data set from a test designed to measure middle school teachers' understanding about computation of fractions was analyzed using M2PL, MRM, Mix2PL, MMixRIM models. The results from MIRT, MRM, and Mix2PL models were inconsistent because these IRT models capture the multidimensionality of data as the characteristics of either persons or items. For some data, however, the multidimensionality might result from the variabilities of both persons and items. Therefore, each type of IRT model might not be enough to figure out the multidimensionality of this empirical data sets. Moreover, the MMixRIM detected that some items measured different traits between latent classes, while these items measured the trait on the same dimension based on the M2PL model. The results

of the empirical study suggested some problems that needed to be studied for MIRT and MixIRT models used for analyzing multidimensional data. These results were used to motivate the present study and the development of a MMixRIM for analysis of multidimensional data.

7.1 Summary of Simulation Study

In the simulation study, the performances of the MMixRIM were compared to those of M2PL and Mix2PL models under several types of dimensional structures. Four different ability distributions, two types of multidimensionality, three levels of correlation between dimensions, and two different sizes of latent classes were manipulate. There was a total of 15 conditions. Each condition was replicated 100 times. Exploratory M2PL models, Mix2PL models, and two-dimensional MMixRIMs analyses were applied to understand how different dimensional structures were represented in the results for each IRT model. All generated data sets were analyzed by nine different IRT models: 2PL model; two- to four-dimensional M2PL (i.e., 2DM2PL, 3DM2PL, and 4DM2PL models) models; two-to four-class Mix2PL (i.e., 2CMix2PL, 3CMix2PL, and 4CMix2PL models) models; and two- to three-class two-dimensional MMixRIMs (i.e., 2D2CMMixRIM and 2D3CMMixRIM). To determine the best-fitting model for the exploratory model analyses, five kinds of model information criterion indices (AIC, BIC, CAIC, AIC_c, and ABIC) were applied.

The exploratory M2PL model analysis was conducted to the unidimensional data sets, and all model information criterion indices failed to detect the true model (i.e., 2PL model). AIC, CAIC, AIC_c, and ABIC mostly suggested 4DM2PL model, and BIC mostly suggested 3DM2PL model as the best-fitting model. The results of the exploratory M2PL model analysis for the two-dimensional data sets showed that all model information criterion indices also failed to detect the true model regardless of the type of multidimensionality and the level of correlation between dimensions. These indices tended to suggest 3DM2PL model instead of 2DM2PL model, which is the true model for these data sets. When the exploratory M2PL

model analysis was conducted for the two-class data sets, AIC, AIC_c, and ABIC mostly suggested 4DM2PL model as the best-fitting model, BIC mainly suggested 2DM2PL model, and CAIC preferred 3DM2PL model. Except by BIC, M2PL model with a larger number of dimensions than the number of latent classes was favored by model information criterion indices. Similarly, the results of the exploratory M2PL model analysis for the two-dimensional and two-class data sets were that 4DM2PL model was mostly favored by AIC and AIC_c. BIC and CAIC suggested either 2DM2PL or 3DM2PL model. ABIC tended to suggest 3DM2PL model when the sizes of latent classes were equal, while it suggested mainly 4DM2PL model when the size of one class was larger than another.

The results of the exploratory Mix2PL model analysis for the unidimensional data sets showed that AIC, BIC and CAIC successfully detected the true model as the best-fitting model in most replications, but that AIC_c and ABIC failed to detect the true model. When the exploratory Mix2PL model analysis was applied to the two-dimensional data sets, the suggestions by five model information criterion indices were consistent. In most replications, the 2CMix2PL model was selected as the best-fitting model regardless of the type of multidimensionality and level of correlation between dimensions. Furthermore, when the exploratory Mix2PL model analysis was applied to the two-class data sets, the five indices successfully detected the true model, the 2CMix2PL model. According to the results of the exploratory Mix2PL model analysis for the two-dimensional and two-class data sets, the 2CMix2PL model was also suggested most often as the best-fitting model.

Based on the results of the exploratory MMixRIM analyses for the two-dimensional data sets, the two-dimensional MMixRIM succussed to detect the true model regardless of the type of multidimensionality and the level of correlation between dimensions. All model information criterion indices suggested 2D1CMMixRIM, which is equivalent to M2PL model and the true model for the two dimensional data sets, as the best-fitting model in the most replications. The results of the exploratory MMixRIM analysis for the two-class data sets differed depending on the model information criterion index. AIC, AIC_c, and ABIC generally

suggested 2D2CMMixRIM as the best fitting model, while BIC and CIAC mostly selected 2D1CMMixRIM. These patterns were observed regardless of the size of latent class. The same pattern was observed in the exploratory MMixRIM analysis for the two-class data sets were shown in the exploratory MMixRIM analyses for the two-dimensional and two-class data sets. AIC, AIC_C, and ABIC successfully detect 2D2CMMixRIM, which is the true model for these data sets, as the best-fitting model, whereas BIC and ABIC failed to detect the true model regardless of the latent class size and the levels of the correlation between dimensions.

Unlike expectation, the Mix2PL model did not need additional classes for the two-dimensional and two-class data sets, even though this model did not assume a multidimensional ability distribution within latent classes. For comparison of the 2D2CMMixRIM and 2CMix2PL models, however, the information criterion indices for the 2D2CMMixRIMs were much smaller than those for the 2CMix2PL models for the two-dimensional and two-class data sets. On the other hand, information criterion indices for the 2CMix2PL model for the two-class data sets were smaller than those of 2D2CMMixRIM. That is, 2CMix2PL was more appropriate for the two-class data sets, and the 2D2CMMixRIM was more appropriate for the two-dimensional and two-class data sets, in spite of the results of the exploratory Mix2PL model analyses that suggested the 2CMix2PL model was the best-fitting model for both the two-class data sets and the two-dimensional and two-class data sets.

Recovery of the estimated item parameters was monitored by calculating BIAS, RMSE, and correlation between the true and estimated parameters. For the item discriminations or the first dimensions slopes, BIASs were very close to zero across all conditions, but RMSEs differed depending on the conditions. Average of RMSEs for the unidimensional data sets was the smallest, and averages of RMSEs for the two-dimensional and two-class data sets were larger than other conditions. The level of correlation between dimensions had a negative influence on RMSEs for the slope estimated by M2PL models and MMixRIMs. Correlations for the two-dimensional data sets were almost one, and those for the two-class data sets and for the two-dimensional and two-class data sets were moderate. For the unidimensional data

sets, however, the correlations were relatively lower than other conditions. That is because of the very small variance of item discriminations, not due to poor estimation. RMSEs, BIASs, and correlations for the second dimension slopes were similar to those for the first dimension slopes. According to BIASs, RMSEs, and correlations, item difficulties and intercepts were recovered well for all conditions.

Recovery of the class memberships was also checked by calculating the percentages of correct detections of class memberships for the two-class data sets and for the two-dimensional and two-class data sets. Class memberships for the two-class data sets were almost perfect compared to their true membership. For the two-dimensional and two-class data sets, however, the percentages of correct detect class memberships were slightly lower than those for the two-class data sets.

The effects of different dimensional structures on the estimation of model parameters were examined by the two-level linear model analyses. Recovery statistics for the estimated item parameters of the unidimensional data sets were significantly smaller than those of other data sets. The type of multidimensionality had a significant effect on the correlations for the slopes, but not on BIASs and RMSEs. The degree of correlation between dimensions appeared to be related to recovery of some statistics. In particular, a highly correlated dimensions had a significantly negative effect on BIASs and RMSEs for the slopes of the two-dimensional and two-class data sets, except for the absolute values of BIASs for the intercepts. A dominant latent class also had a significantly negative effect on the recovery statistics for the item difficulties of the two-class data sets and for the intercepts of the two-dimensional and two-class data sets.

7.2 Conclusion

Based on the finings from the simulation study, four main conclusions were drawn. First, the performances of Mix2PL model and MMixRIM were better than those of M2PL model for the multidimensional data analysis considered in this study. According to the results of

the exploratory M2PL model analyses, more than a three dimensional M2PL model was mostly suggested as the best fitting model for all data sets, even for the unidimensional data sets. As expected, however, 2D1CMMixRIM was mainly suggested as the best-fitting model based on the results of the exploratory MMixRIM analyes for the two-dimensional data sets. The results of the exploratory MMixRIM analyses for the two-dimensional and two class data sets also suggested that 2D2CMMixRIM was appropriate for these data sets. According to the results of the exploratory Mix2PL model analyses for the two-dimensional data sets and for the two-dimensional and two-class data sets, 2CMix2PL model was favored by five model information criterion indices regardless of conditions. These results might be supported by the relationship between a between-items multidimensional IRT models and a mixture Rasch model proved by Rijmen and De Boeck (2005). Rijmen and De Boeck (2005) have demonstrated that the between items multidimensional IRT models is equivalent to a mixture Rasch model. Additionally, the results from the exploratory Mix2PL analyses for the two-dimensional and two-class data sets might be an evidence of the robustness to the weak multidimensionality of Mix2PL model. In the simulation study, the two-dimensional and two-class data sets were generated based on the assumption that Class 1 represented the group of examinees who did not learn the latent trait on the second dimension, and Class 2 represented the group of examinees who did not learn the latent trait on the first dimensions. Accordingly, the multidimensionality within each latent class became weak, and this weak multidimensionality within each latent class did not require additional latent classes in the Mix2PL model.

Second, the recovery statistics of the estimated slopes were larger than those of intercepts of M2PL models and MMixRIMs. Additionally, the recovery statistics of intercepts increased as the correlations between dimensions increased. These results were consistent with Bolt and Lall (2003). A plausible explanation for these patterns is that the model identification for rotational indeterminacy affected the estimation of slope parameters. In this study, the independent dimensions were assumed to hold for all data sets, even though some data

sets were generated based on the correlated dimensions. Although the estimated parameters transformed to the scale of the true parameters before the recovery analyses, the bias of the estimate slopes could not correct perfectly.

Third, the estimation of model parameters tended to be less accurate as the number of parameters increased. Additionally, the percentages of correct detections of class memberships in simpler model (i.e., Mix2PL model) were also larger than in more complicated models (i.e., MMixRIM). It might be due to the sample size relative to the number of model parameters. MMixRIM has a relatively smaller sample size than other models because this model has a larger number of model parameters than others. Accordingly, the sample size might not be sufficient to get stable estimation of parameters for the MMixRIM.

Finally, the five information criterion indices used in the simulation study performed differently depending on the type of IRT model. AIC, AIC_c, and ABIC detected well the true model for the exploratory Mix2PL model and MMixRIM analyses, but BIC and CAIC performed well only for the exploratory Mix2PL model analysis. For the exploratory M2PL model analysis, all indices failed to detect the true model. The different performances of these indices could be due to the differences in penalties applied for each information criterion index. As can be seen in Figure 7.1, the penalty of AIC, AIC_c, and ABIC were close each other. BIC and CAIC have similar penalties but these are much larger than those of AIC, AIC_c, and ABIC. Additionally, the difference in penalty functions between M2PL models was smaller than those between Mix2PL models or MMixRIMs. That is, the effect of penalty function on model selection might differ depending on the combinations of the kinds of the model information criterion index and the type of IRT models. Consequently, it would be helpful to consider the various types of indices for the model comparisons when several types of IRT models are considered, like this simulation study.

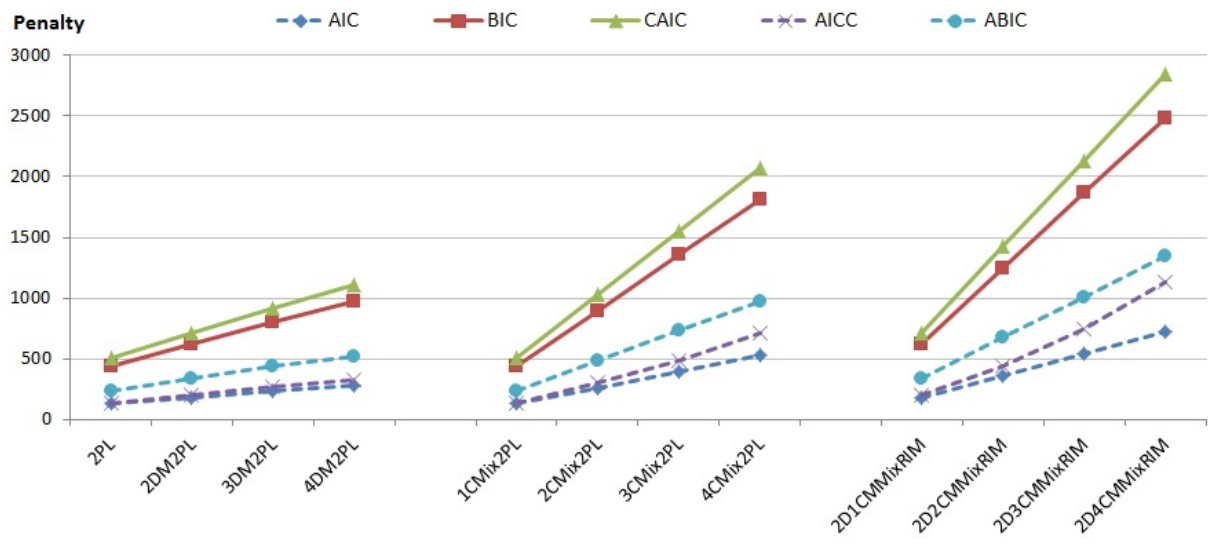

Figure 7.1: Penalty terms of five information criterion indices

7.3 Limitations and Future Research

The main advantage of the random item model is that it is available to include not only person covariates but also item covariates. In this study, covariates were not included in the models although the proposed model was an extension of a random item model. It would be useful to explore the effect of covariates on explaining the multidimensionality and the characteristics of latent classes by including covariates to the MMixRIM.

To avoid complicating simulation conditions, the sample size and the test length were fixed. Although the sample size and test length (i.e., 1,000 examinees and 30 items) used in this simulation study are commonly used in the simulation study for the multidimensional data analyses (e.g., Béguin & Glas, 2001; Bolt & Lall, 2003), these might not have been large enough for the MMixRIM. It is possible that the relatively larger recovery statistics of MMixRIM than those of other models were caused by relatively smaller sample size than other models when the number of model parameters were considered. Moreover, it could affect the performance of information criterion indices. Therefore, further study with

a larger sample size and a longer test length would be useful to more accurately explore the performance of MMixRIM for multidimensional data analysis.

In addition to the relative sample size, the way of counting the number of estimated parameters or the type of likelihood values might make an impact on the performance of information criterion indices. For example, Yao and Schwarz (2006) counted both item parameters and ability parameters as the number of parameters for calculating AIC, while ability parameters were not be counted as the number of parameters in this study. Meanwhile, only parameters of item distributions (i.e., means and variances) were counted in some studies with the random item model (e.g., Cho, Gilbert, & Goodwin, 2013). It might be valuable to examine the effects of these differences on the model selection.

Finally, only two-dimensional conditions were considered in the simulation study, but sometimes an assessment could be designed to measure more than three traits. Additionally, the situation assumed for the simulation study yielded weak multidimensionality within each latent class. The application of MMixRIM to the data sets with a strong multidimensionality within each latent class might improve the performance of the MMixRIM.

Bibliography

[1] Adams, R. J., Wilson, M., & Wang, W.-C.(1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control, 19*, 716-723.

[3] Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement, 48*, 313-332.

[4] Assessment Systems Corporation . (1996). XCALIBRE: Marginal maximum likelihood estimation program, Version 1.10 [Computer program]. St. Paul MN: Author.

[5] Batley, R.-M., & Boss, M. W. (1993). The effects on parameters estimation of correlated dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17*, 131-141.

[6] Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models.*Psychometrika, 66*, 541-562.

[7] Blommers, B. J., & Forsyth, R. A. (1997). *Elementary statustucal methods in psychology and education*. Lanham, MD: University Press of America.

[8] Bock, R. D, Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4.0 [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

[9] Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335-352.

[10] Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Marcov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395-414.

[11] Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Journal of Educational Measurement, 33*, 2-14.

[12] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extension. *Psychometrika, 52*, 345-370.

[13] Cho, S.-J., Gilbert, J. K., & Goodwin, A. P. (2015). Explanatory multidimensional multilevel random item response model: An application to simultaneous investigation of word and person contributions to multidimensional lexical representations. *Psychometrika, 78*, 830-855.

[14] Choi, I.-H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational Psychological Measurement, 75*, 78-101.

[15] Choi, Y.-J. (2014). *Metric identification in Mixture IRT models.* (Unpublished doctoral dissertation). The University of Georgia, Athens, GA.

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

[17] Cohen , A. S., & Cho, S.-J. (2016). Information criteria. In W. J. van der Linden (Ed.), *Handbood of item response theoty, Volumn two: Statistical Tools* (pp. 363-378). Boca Raton, FL: CRC Press.

[18] Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, nehavioral, and health sciences*. Hoboken, NJ: John Wiley & Sons.

[19] Congdon, P. (2003). *Applied Bayesian modelling*. New York, NY: John Wiley.

[20] Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 883-904.

[21] Dai, Y. (2013). A mixture Rasch model with a covariate: A simulation study via Bayesian Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 37*, 375-396.

[22] De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guildford Press.

[23] De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

[24] De Jong, M. G., & Steenkamp, J. -B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*, 3-32.

[25] De la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement, 34*, 264-285.

[26] 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*, 249-262.

[27] Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

[28] Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (Tech. Rep. No. 12-119). University Park, PA: The Pennsylvania State University, The Methodology Center. Retrieved from http://methodology.psu.edu.

[29] Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 259-374.

[30] Fox, J. -P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

[31] Fraser, C. (1988). NOHARM: A computer programing for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer program]. Armidale, Australia: University of New Egnland.

[32] Fredericks, S., Tuerlincks, F., De Bock, P., & Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. *Journal of Educational Measurement, 47*, 432-457.

[33] Gamerman, D., & Lopes, H. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference* (2nd ed.), Boca Raton: FL, Chapmen & Hall/CRC.

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

[35] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds). *Bayesian Statistics 4* (pp. 169-194). Oxford: Oxford University Press.

[36] Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

[37] Hambelton, R. K., Swaminathan, H. S., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

[38] Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91-115.

[39] Heidelberger, P., & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*, 1109-1144.

[40] Howell, D. C. (2007). *Statistical methods for psychology* (6th ed.). Belmont, CA: Thomson Wadswoth.

[41] Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck, & M. Wilson (Eds). *Explanatory item response models: A Generalized linear and nonlinear approach* (pp. 189-212). New York, NY: Springer.

[42] Jeon, M., Draney, K., & Wilson, M. (2015). A general saltus LLTM-R for cognitive assessments. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W. -C. Wang (Eds.), *Quantitative psychology research: Springer Proceedings in Mathematics and Statistics* (pp. 73-90). New York, NY: Springer.

[43] Kang, T. (2006). *Model selection methods for unidimensional and multidimensional IRT models*. (Unpublished doctoral dissertation). The University of Wisconsin-Madison, Madison, WI.

[44] Kirisci, L,. & Hsu, T., & Kaohsiung, L. Y. (2001). Robustness of item parameter estimation programs to assumptions to unidimensionality and normality. *Applied Psychological Measurement, 25*, 146-162.

[45] Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Social and Behaviorl Sciences, 46*, 135-140.

[46] Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*, 232-244.

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

[48] Li, Y., Jiao, H., & Lissitz, W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K-12 large-scale science assessment, *Journal of Applied Testing Technology, 13*, 1-27.

[49] Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, No.7.

[50] Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[51] Lord, F. R., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

[52] McDonald, R. P. (1967). Non-linear factor analysis. *Psychometric Monographs*, No.15.

[53] McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-107.

[54] McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York, NY: Springer.

[55] McKinley, R. L., & Reckase, M. R. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82-1). Iowa City, IA: American College Testing Program.

[56] McKinley, R. L., & Reckase, M. R. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR 83-2). Iowa City, IA: American College Testing Program.

[57] McKinley, R. L., Way, W. D. (1992). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models* (Technical Report TR-5). Retrieved from Educational Testing Service website: https://www.ets.org/Media/Research/pdf/RR-92-16.pdf

[58] Mislevy, R. J., & Bock, R. K. (1990). BIILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items [Computer program]. Chicago, IL: Scientific Software International.

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

[60] Mulaik, S. T. (2010). *Foundations of factor analysis*. Boca Raton, FL: Chapman & Hall/CRC.

[61] Muthén, L. K., & Muth'en, B. O. (1998-2013). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthn & Muthn.

[62] Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17*, 29-38.

[63] Oshima, T.C., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional IRT-based internal measurement of differential functioning of items and tests. *Journal of Educational Measurement, 34*, 253-272.

[64] Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC, *R News, 6*, 7-11.

[65] Raftery, A. E., & Lewis, S. (1992). How many iterations in the Gibbs sampler. In J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. Smith (Eds.), *Bayesian statistics* 4 (pp. 763-773). Oxford, England: Oxford University Press.

[66] Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 321-333). Berkeley, CA: University of California Press.

[67] Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

[68] Reckase, M. D. (1990, April). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

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

[70] Reckase, M. D. (1997b). A linear logistic multidimensional model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.

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

[72] Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271-285.

[73] Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika, 70*, 481-496.

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

[75] Sahlin, K. (2011). *Estimating convergence of Markov chain Monte Carlo simulations* (Master's thesis). Retrieved from http://www2.math.su.se/matstat/reports/master/2011/rep2/report.pdf

[76] Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 38*, 111-121.

[77] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

[78] Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

[79] Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online, 4*, 19-32.

[80] Smit, A., Kelderman, H., & van der Flier, H. (2000). The mixed Birnbaum model: Collateral information. *Methods of Psychological Research Online, 5*, 31-43.

[81] Snijders, T. A. B. & Bosker R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, (2nd ed.), Los Angeles, Sage.

[82] Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods, 7*, 13-26.

[83] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Joutnal of the Royal Statistical Society, Series B, 64*, 583-639.

[84] Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. J. (2007). OpenBUGS [Computer program]. Robinson Way, Cambridge CB2 2SR, UK: MRC Biostatistics Unit, Institute of Public Health.

[85] Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

[86] Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.

[87] Stout, W. (2006). Nonparametric dimensionality assessment package DIMPACK (Version 1.0) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

[88] Svetina, S., & Levy, R. (2013). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment, 19*, 35-57.

[89] Sympson, J. B. (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, MN: University of Minnesota.

[90] Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*, 159-203.

[91] Thissen, D. (1991). *MULTILOG Version 6.0 user's guide*. Chicago: Scientific Software International.

[92] Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369-386.

[93] Von Davier, M. (2001). WINMIRA [Computer program]. St.Paul: Assessment Systems Corporation.

[94] Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement, 66*, 721-738.

[95] Wang. A. (2011). *A mixture cross-classification IRT model for test speediness* (Unpublished doctoral dissertation). The University of Georgia, Athens, GA.

[96] Whitely, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18*, 67-84.

[97] Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck, & M. Wilson (Eds). *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43-74). New York, NY: Springer.

[98] Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGITST: A computer program for estimating examinees ability and item characteristic curve parameters*. Princeton, NJ: Educational Testing Service.

[99] Yao, L. (2003). BMIRT: Bayesian multivariate item response model [Computer software]. Monterey ,CA: DMDC DoD Center.

[100] Yan. L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-forman tests., *Apploied Psychological Measurement, 30*, 1-25.

Appendix A

OpenBugs Code Used for Two-dimensional and Two-class MMixRIM

```
## 2-dimensional & 2-class MMixRIM
## NI=30 items, NE=1000 examinees
## r2: responses
## d2: intercept parameter
## a2.1: slope parameter for the first dimension
## a2.2: slope parameter for the second dimension
## gmem: class membership
## theta2: two-dimensional ability parameters

model{
  for (j in 1:NE) {
   for (k in 1:NI) {
          r2[j,k]<-resp[j,k]
}}

## 2 class model
for ( j in 1:NE){
for (k in 1:NI){
r2[j,k]~dbern(p2[j,k])
logit(p2[j,k])<-a2.1[gmem2[j],k]*theta2[j,1]+a2.2[gmem2[j],k]*theta2[j,2]
                +d2[gmem2[j],k]
```

```
l2[j,k]<-log(p2[j,k])*r2[j,k]+log(1-p2[j,k])*(1-r2[j,k])
}}

# Priors for thetas
for (j in 1:NE){
theta2[j,1:2]~dmnorm(mut2[1:2],taut2[1:2,1:2])
gmem2[j]~dcat(pi2[1:2])
}
pi2[1:2]~ddirch(alpha2[])

alpha2[1]<-.5
alpha2[2]<-.5

# Priors for a and d
for(g in 1:2){
a2.1[g,1]<-1
a2.2[g,1]<-0
d2[g,1]<-0

for(k in 2:20){
a2.1[g,k]~dnorm(mua2.1[g],taua2.1[g])I(0,)
a2.2[g,k]~dnorm(mua2.2[g],taua2.2[g])I(0,)
d2[g,k]~dnorm(mud2[g],taud2[g])
}
a2.1[g,21]<-0
a2.2[g,21]<-1
d2[g,21]<-0
```

```
for(k in 22:30){
a2.1[g,k]~dnorm(mua2.1[g],taua2.1[g])I(0,)
a2.2[g,k]~dnorm(mua2.2[g],taua2.2[g])I(0,)
d2[g,k]~dnorm(mud2[g],taud2[g])
}}

# Hyper prior for theta
mut2[1]<-0
mut2[2]<-0
taut2[1,1]<-1
taut2[1,2]<-0
taut2[2,1]<-0
taut2[2,2]<-1

#Hyper prior for a and d
for(g in 1:2){
mua2.1[g]~dnorm(0,.5)I(0,)
mua2.2[g]~dnorm(0,.5)I(0,)
mud2[g]~dnorm(0,.5)

taua2.1[g]~dgamma(1,1)
taua2.2[g]~dgamma(1,1)
taud2[g]~dgamma(1,1)

vara2.1[g]<-1/taua2.1[g]
vara2.2[g]<-1/taua2.2[g]
```

```
vard2[g]<-1/taud2[g]
}

# calculate loglikelihood
loglik2<-sum(l2[1:NE, 1:NI])
}
```

Appendix B

Monitoring Convergence based on Heidelberger and Welch's (1983) Convergence Diagnostics and MC error-standard deviation ratio

Table B.1: Percentage of Passed Item Parameters of M2PL and Mix2PL Models Based Heidelberger and Welch's (1983) Convergence Diagnostics

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
1	1.00	.98	1.00	.99	1.00	1.00	.99	.99	1.00
2	1.00	.91	1.00	1.00	.99	.86	.99	1.00	1.00
3	.88	.98	.73	.99	.98	1.00	1.00	1.00	.98
4	.98	.73	1.00	.99	.96	1.00	1.00	1.00	1.00
5	.95	.91	.96	.98	.96	.99	1.00	1.00	.92
6	.95	.65	.73	.99	1.00	1.00	.98	.85	.92
7	.98	.99	1.00	1.00	1.00	.96	.95	1.00	.98
8	.98	.77	1.00	1.00	.98	.99	.95	.99	.96
9	.93	.75	1.00	.99	1.00	.98	1.00	1.00	.88
10	1.00	.83	.98	.99	.98	1.00	.99	1.00	.98
11	1.00	.85	1.00	1.00	.99	1.00	1.00	1.00	1.00
12	1.00	.78	.82	.99	1.00	.98	1.00	.98	.95
13	.98	.73	.99	1.00	.96	1.00	.87	.99	1.00
14	.97	.72	.90	.99	.83	.99	.99	.98	.98
15	.93	.88	.99	1.00	1.00	1.00	.94	.94	1.00
16	.97	.95	.91	.98	.99	.73	1.00	1.00	1.00
17	.90	.75	.97	.98	1.00	.99	1.00	1.00	.98

Continued on next page

Table B.1 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
18	.98	.95	.98	.99	1.00	.89	.99	.99	1.00
19	.88	.97	.78	.99	1.00	.96	.98	1.00	.99
20	1.00	.74	.94	.99	.99	1.00	.99	.99	.88
21	1.00	.77	1.00	1.00	.99	.98	1.00	.96	.99
22	1.00	.81	1.00	1.00	1.00	.86	1.00	1.00	.99
23	.92	.72	1.00	.99	1.00	.96	1.00	.98	1.00
24	.97	.81	.97	.96	1.00	.99	1.00	.98	.95
25	.95	.73	.75	.99	.96	.99	1.00	.99	.79
26	.95	.88	.73	.96	.98	.98	.86	.86	1.00
27	.63	.70	.98	1.00	1.00	.99	1.00	1.00	.99
28	1.00	.69	1.00	1.00	.95	.99	.99	.98	.99
29	.92	.70	.99	1.00	.98	1.00	1.00	1.00	1.00
30	.98	.92	.96	1.00	1.00	1.00	1.00	.99	.99
31	1.00	.81	.79	1.00	1.00	.99	1.00	.99	1.00
32	.95	.98	.88	1.00	1.00	1.00	.96	.98	.98
33	1.00	.88	.99	1.00	1.00	1.00	1.00	1.00	.80
34	.95	.75	.74	1.00	1.00	1.00	.99	.71	1.00
35	.97	.98	.99	.94	.98	.94	1.00	.98	.98
36	.92	.83	.94	1.00	1.00	.90	1.00	1.00	1.00
37	.98	.98	.99	.99	.98	1.00	.98	.99	.99
38	.98	1.00	.75	.98	.69	.99	1.00	.99	1.00
39	1.00	.81	1.00	1.00	.99	1.00	.99	.99	1.00
40	.87	.89	.98	.98	.99	1.00	1.00	.95	.94
41	.98	.80	1.00	1.00	.93	.93	1.00	1.00	.94
42	1.00	.99	.99	1.00	1.00	.96	1.00	.99	.98
43	.78	.81	1.00	1.00	1.00	.74	1.00	.93	1.00

Continued on next page

Table B.1 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
44	.98	.91	1.00	1.00	1.00	.96	1.00	.99	.98
45	.98	.97	.99	1.00	.95	.98	1.00	1.00	1.00
46	.97	.72	.98	.98	.96	.80	1.00	1.00	1.00
47	.98	.92	.99	.99	1.00	1.00	.98	.96	1.00
48	.82	.63	.98	.99	1.00	1.00	.99	1.00	1.00
49	.95	.92	.96	.98	.99	.93	.99	1.00	.89
50	.93	1.00	.99	1.00	.75	1.00	.95	1.00	1.00
51	.93	.92	.98	1.00	.99	1.00	.99	1.00	1.00
52	1.00	.87	1.00	1.00	1.00	1.00	.95	.95	1.00
53	.85	.69	1.00	.99	.99	1.00	1.00	1.00	.99
54	.98	.95	.99	.99	.99	.83	1.00	.95	1.00
55	1.00	.88	.75	.99	.98	.99	1.00	.95	.87
56	.98	.75	.95	1.00	1.00	.93	1.00	1.00	1.00
57	1.00	.86	.74	1.00	1.00	1.00	1.00	1.00	.96
58	.93	.96	1.00	.99	.98	.98	1.00	1.00	.89
59	1.00	.64	1.00	.96	1.00	1.00	.98	1.00	1.00
60	.95	.75	.78	1.00	.99	1.00	1.00	.99	.99
61	1.00	.81	1.00	.96	1.00	.62	1.00	.99	1.00
62	.93	.96	.74	1.00	.99	1.00	1.00	.96	1.00
63	.98	.81	1.00	.99	.67	1.00	.99	.85	.99
64	.98	.98	1.00	.99	.99	1.00	.98	.98	.96
65	.73	.82	1.00	1.00	1.00	.95	1.00	.92	1.00
66	1.00	.93	.98	.99	.98	.96	1.00	.76	1.00
67	1.00	.67	1.00	.99	.99	.99	1.00	1.00	.96
68	.97	.83	.91	1.00	.99	.99	1.00	1.00	.99
69	.98	.84	.90	1.00	.99	1.00	.99	.93	1.00

Continued on next page

Table B.1 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
70	.97	.98	.86	1.00	.98	.98	.99	1.00	.94
71	.97	.84	.94	.98	.96	.85	1.00	.98	.98
72	1.00	.86	.74	1.00	1.00	.96	.98	.96	.99
73	1.00	.83	.97	.99	.99	.98	1.00	1.00	1.00
74	.97	.68	.99	.99	.99	.98	1.00	.98	.99
75	1.00	.67	1.00	.98	.98	.98	.94	1.00	1.00
76	.98	.78	.99	.95	.99	.96	1.00	1.00	.82
77	.92	.83	1.00	1.00	1.00	.86	.89	.95	.89
78	1.00	.96	.99	.99	1.00	.96	.93	1.00	1.00
79	.98	.95	.98	.99	1.00	.85	1.00	.99	1.00
80	.75	.77	.98	1.00	1.00	1.00	.99	.95	1.00
81	1.00	1.00	.94	.99	1.00	1.00	1.00	.99	1.00
82	.97	.93	.84	.81	.99	1.00	.95	.99	.98
83	1.00	.88	.99	1.00	.94	.98	.99	1.00	1.00
84	1.00	.96	.99	1.00	1.00	.98	1.00	.99	1.00
85	.73	.88	.73	.99	1.00	1.00	.99	1.00	.96
86	.92	.95	1.00	.99	1.00	1.00	.99	.99	1.00
87	.98	.78	.99	.96	1.00	1.00	.99	.96	.99
88	.98	.68	.98	.99	.99	.92	1.00	.99	1.00
89	.97	.96	.88	.99	.96	.99	1.00	.98	1.00
90	.90	.95	.98	1.00	.99	.99	1.00	.95	1.00
91	.98	.74	.98	1.00	.98	.93	.99	.93	.98
92	.95	.74	.96	1.00	.99	.99	1.00	1.00	1.00
93	1.00	.82	.97	1.00	.82	.99	.99	.98	1.00
94	.97	.80	.99	1.00	1.00	.99	1.00	.96	.99
95	1.00	.77	.98	.99	.99	.95	1.00	.98	.82

Continued on next page

Table B.1 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
96	1.00	.93	.96	1.00	.96	1.00	.99	1.00	.99
97	1.00	.75	.75	1.00	.98	.86	1.00	.98	.99
98	1.00	.92	.82	.99	.98	1.00	.95	.88	1.00
99	1.00	.68	.73	1.00	.90	1.00	.99	.99	.85
100	1.00	.94	1.00	1.00	.92	1.00	1.00	.99	.99

Table B.2: Percentage of Passed Item Parameters of MMixRIM Based Heidelberger and Welch's (1983) Convergence Diagnostics

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
1	1.00	.99	.99	.99	.99	.96
2	**.34**	.99	.99	.99	.99	.98
3	.96	1.00	1.00	.99	.95	.76
4	.93	.96	1.00	1.00	.96	.95
5	.99	.89	.98	.93	1.00	.95
6	.98	.96	.95	.99	.99	.96
7	1.00	.99	.99	.98	.98	.99
8	.96	1.00	1.00	1.00	.99	.99
9	.97	.71	.98	.99	1.00	.95
10	1.00	.93	.99	.96	.98	1.00
11	.99	.99	.99	.96	.99	.97
12	.99	.99	1.00	.99	.97	.99
13	.99	1.00	.91	.98	.99	.98
14	.88	.98	.99	.97	.99	.99
15	.92	.99	.99	.96	.92	1.00
16	.98	1.00	.99	.96	.98	.99
17	.93	1.00	.99	.99	**.35**	.81
18	.98	.92	.98	.97	.93	.99
19	.99	.99	.99	1.00	.95	1.00
20	.98	.98	1.00	.99	.99	1.00
21	.99	1.00	.99	.97	.98	.93
22	.63	.99	.96	.99	.99	.99
23	.99	.99	.97	.99	.99	1.00
24	.95	1.00	.99	.99	.99	.95

Continued on next page

Table B.2 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
25	.95	1.00	1.00	.99	.95	.98
26	.81	.99	1.00	.95	1.00	1.00
27	.99	1.00	.99	.99	1.00	.98
28	.99	.96	.99	1.00	.99	.97
29	.96	1.00	.99	.95	.85	1.00
30	.99	.98	.98	.98	.99	.99
31	1.00	.97	.96	.99	.88	.96
32	.99	.99	.85	1.00	.99	1.00
33	1.00	.96	.96	.98	1.00	.97
34	.99	.97	1.00	.98	1.00	1.00
35	.89	.91	1.00	.97	1.00	1.00
36	1.00	.97	.96	.95	1.00	.98
37	.99	.98	.98	.99	.81	.96
38	.99	.96	.96	.97	.96	.98
39	.98	1.00	.99	.99	.95	1.00
40	.92	.99	.99	.99	.98	1.00
41	.93	1.00	.96	.98	.99	1.00
42	.95	.87	.97	.98	.96	.99
43	.95	.99	.95	.97	.98	1.00
44	.94	.99	.96	.99	1.00	.99
45	1.00	.98	.93	.99	.98	.99
46	1.00	.95	.99	.98	.99	.99
47	.90	.93	.98	.98	.99	1.00
48	.96	.99	.99	.98	.99	.99
49	.92	.98	.96	.96	.90	.98
50	.99	.98	.92	.99	.99	.99

Continued on next page

Table B.2 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
51	.92	.99	.99	1.00	.97	1.00
52	.86	.98	1.00	.99	.95	.99
53	.98	.97	.98	.99	.99	1.00
54	.99	1.00	.98	.95	.98	.98
55	.97	.92	.99	.93	1.00	.97
56	.99	.98	.99	.99	1.00	.96
57	.99	1.00	.99	.98	.98	.99
58	.90	.99	.98	.99	.92	.98
59	1.00	**.40**	.99	.99	.98	1.00
60	.99	1.00	.99	1.00	.99	.95
61	.99	.99	.94	.93	.99	1.00
62	1.00	1.00	.99	.96	.91	.99
63	1.00	.86	1.00	.98	.94	.99
64	.99	.91	1.00	.97	.95	1.00
65	.99	.98	1.00	.96	.94	.99
66	.70	.99	.96	.99	.97	.98
67	.89	.97	.96	.96	.99	.97
68	.99	.70	.99	.95	.98	.99
69	.99	.88	1.00	.96	.98	.99
70	1.00	.95	.91	.98	.95	.99
71	.92	.93	.99	.71	.98	.80
72	.99	.97	.99	.99	.99	.98
73	.99	.99	.99	.99	.99	.93
74	.98	.99	.95	.99	.92	1.00
75	.98	.97	.98	.99	.98	.88
76	.96	.99	.92	.98	1.00	.99

Continued on next page

Table B.2 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
77	.79	.79	.99	.98	.99	.99
78	.98	1.00	.98	.89	.98	.95
79	.98	.99	.95	.99	.96	.99
80	.99	.96	.99	1.00	.98	.99
81	.99	.83	1.00	.99	.99	.96
82	.98	.98	.96	.98	.99	.98
83	.96	.99	.97	.99	.99	.99
84	1.00	.99	.98	.99	.99	1.00
85	.98	.99	.97	.99	.80	.98
86	.99	.93	.92	.99	.92	1.00
87	.83	.98	.96	.97	.98	1.00
88	.98	.99	.95	.98	.95	.99
89	.78	1.00	.99	.99	1.00	.99
90	.95	.99	.95	.99	1.00	.99
91	.99	.97	1.00	.99	.98	1.00
92	.87	.96	.75	1.00	.99	.92
93	.99	.99	1.00	.96	.99	1.00
94	.95	.98	1.00	.99	.96	.95
95	.99	.99	.99	.99	1.00	.99
96	.99	.99	.99	.99	.99	1.00
97	.96	.98	.99	1.00	.94	.99
98	.95	.99	.98	1.00	.99	.99
99	.94	.98	.98	.99	.98	1.00
100	.99	1.00	.95	.99	1.00	.99

Table B.3: Percentage of Passed Item Parameters of M2PL and Mix2PL Models Based the Ratio of MC Error to Standard Deviation

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
1	1.00	**.03**	.75	1.00	1.00	.85	1.00	1.00	.90
2	1.00	**.01**	.75	1.00	1.00	.95	1.00	1.00	.86
3	1.00	**.01**	.75	1.00	1.00	.87	1.00	1.00	.62
4	1.00	**.02**	.75	1.00	1.00	.81	1.00	1.00	.87
5	1.00	**.00**	.75	1.00	1.00	.96	1.00	1.00	.65
6	1.00	**.00**	.75	1.00	1.00	.96	1.00	1.00	.88
7	1.00	**.03**	.75	1.00	1.00	.56	1.00	.89	.62
8	1.00	**.01**	.75	1.00	1.00	.98	1.00	1.00	.83
9	1.00	**.01**	.75	1.00	1.00	.69	1.00	1.00	**.48**
10	1.00	**.02**	.75	1.00	1.00	.99	1.00	1.00	.77
11	1.00	**.00**	.75	1.00	1.00	.74	1.00	.99	.60
12	1.00	**.03**	.75	1.00	1.00	.99	1.00	1.00	.80
13	1.00	**.01**	.75	1.00	1.00	.94	1.00	.95	.54
14	1.00	**.00**	.75	1.00	1.00	1.00	1.00	1.00	.85
15	1.00	**.00**	.75	1.00	1.00	1.00	1.00	1.00	.96
16	1.00	**.02**	.75	1.00	1.00	.86	1.00	.96	.69
17	1.00	**.03**	.75	1.00	.99	.69	1.00	1.00	.69
18	1.00	**.00**	.75	1.00	1.00	.99	1.00	.99	.77
19	1.00	**.01**	.75	1.00	1.00	.79	1.00	.96	.64
20	1.00	**.01**	.75	1.00	1.00	.89	1.00	1.00	.99
21	1.00	**.03**	.75	1.00	.87	.71	1.00	1.00	.61
22	1.00	**.01**	.75	1.00	1.00	.96	1.00	1.00	.90
23	1.00	**.01**	.75	1.00	1.00	.85	1.00	1.00	.82
24	1.00	**.02**	.75	1.00	1.00	.89	1.00	1.00	.87

Continued on next page

Table B.3 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
25	1.00	**.01**	.75	1.00	1.00	.90	1.00	.89	.57
26	1.00	**.01**	.75	1.00	1.00	.99	1.00	1.00	.76
27	1.00	**.01**	.75	1.00	1.00	.86	1.00	1.00	.56
28	1.00	**.02**	.75	1.00	1.00	.77	1.00	1.00	.79
29	1.00	**.00**	.75	1.00	1.00	1.00	1.00	1.00	.98
30	1.00	**.02**	.75	1.00	1.00	.90	1.00	1.00	.68
31	1.00	**.02**	.75	1.00	1.00	.90	1.00	1.00	.70
32	1.00	**.01**	.75	1.00	1.00	.85	1.00	1.00	.69
33	1.00	**.00**	.75	1.00	1.00	.77	1.00	.99	.56
34	1.00	**.01**	.75	1.00	1.00	.71	1.00	.98	.67
35	1.00	**.02**	.75	1.00	1.00	.87	1.00	1.00	.96
36	1.00	**.02**	.75	1.00	1.00	1.00	1.00	.93	.55
37	1.00	**.03**	.75	1.00	1.00	.99	1.00	1.00	.71
38	1.00	**.01**	.75	1.00	1.00	.71	1.00	1.00	.92
39	1.00	**.01**	.75	1.00	.96	.83	1.00	1.00	.50
40	1.00	**.00**	.75	1.00	1.00	.81	1.00	1.00	.54
41	1.00	**.00**	.75	1.00	.93	.74	1.00	1.00	.89
42	1.00	**.00**	.75	1.00	.89	.73	1.00	1.00	.65
43	1.00	**.00**	.75	1.00	.93	.77	1.00	1.00	.83
44	1.00	**.01**	.75	1.00	.99	.83	1.00	1.00	.65
45	.98	**.01**	.75	1.00	1.00	.85	1.00	.86	.74
46	1.00	**.01**	.75	1.00	1.00	.89	1.00	1.00	.87
47	1.00	**.00**	.75	1.00	1.00	.96	1.00	1.00	.92
48	1.00	**.01**	.75	1.00	1.00	1.00	1.00	1.00	.61
49	1.00	**.02**	.75	1.00	1.00	.71	1.00	1.00	.73
50	1.00	**.03**	.75	1.00	1.00	.85	1.00	1.00	.74

Continued on next page

Table B.3 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
51	1.00	**.01**	.75	1.00	1.00	.85	1.00	1.00	.79
52	1.00	**.00**	.75	1.00	1.00	.76	1.00	1.00	.88
53	1.00	**.01**	.75	1.00	1.00	.87	1.00	1.00	.79
54	1.00	**.03**	.75	1.00	1.00	.88	1.00	1.00	.63
55	1.00	**.01**	.75	1.00	1.00	.98	1.00	1.00	.95
56	1.00	**.01**	.75	1.00	1.00	.99	1.00	1.00	.92
57	1.00	**.02**	.75	1.00	1.00	.93	1.00	1.00	.61
58	1.00	**.00**	.75	1.00	1.00	.95	1.00	1.00	1.00
59	1.00	**.00**	.75	1.00	1.00	.99	1.00	1.00	.70
60	1.00	**.03**	.75	1.00	.96	.93	1.00	.98	.71
61	1.00	**.03**	.75	1.00	1.00	.68	1.00	.99	.82
62	1.00	**.02**	.75	1.00	1.00	1.00	1.00	.92	.56
63	1.00	**.00**	.75	1.00	1.00	.93	1.00	1.00	.57
64	.98	**.00**	.75	1.00	1.00	.92	1.00	1.00	.77
65	1.00	**.00**	.75	1.00	1.00	.86	1.00	1.00	.93
66	1.00	**.02**	.75	1.00	1.00	1.00	1.00	.99	.68
67	1.00	**.01**	.75	1.00	1.00	1.00	1.00	.95	.82
68	1.00	**.01**	.75	1.00	1.00	1.00	1.00	1.00	.95
69	1.00	**.00**	.75	1.00	1.00	.93	1.00	1.00	.58
70	1.00	**.02**	.75	1.00	1.00	.98	1.00	1.00	.90
71	1.00	**.06**	.75	1.00	1.00	.64	1.00	.93	**.42**
72	1.00	**.02**	.75	1.00	1.00	.76	1.00	1.00	.75
73	1.00	**.03**	.75	1.00	1.00	.81	1.00	1.00	.98
74	1.00	**.00**	.75	1.00	1.00	.96	1.00	.96	.56
75	1.00	**.02**	.75	1.00	1.00	.71	1.00	1.00	.63
76	1.00	**.02**	.75	1.00	.99	.70	1.00	1.00	.68

Continued on next page

Table B.3 – continued from previous page

Replication	D1	D1C2D	D1C2E	D2BR0	D2BR3	D2BR6	D2WR0	D2WR3	D2WR6
77	1.00	**.01**	.75	1.00	1.00	.90	1.00	1.00	.87
78	1.00	**.02**	.75	1.00	1.00	.98	1.00	.99	.54
79	1.00	**.01**	.75	1.00	1.00	1.00	1.00	1.00	.96
80	1.00	**.00**	.75	1.00	1.00	1.00	1.00	.96	**.42**
81	1.00	**.00**	.75	1.00	1.00	.82	1.00	.99	.68
82	1.00	**.02**	.75	1.00	1.00	.98	1.00	.99	.51
83	1.00	**.02**	.75	1.00	1.00	.87	1.00	1.00	.71
84	1.00	**.01**	.75	1.00	1.00	1.00	1.00	.98	.62
85	1.00	**.04**	.75	1.00	1.00	.79	1.00	1.00	.56
86	1.00	**.00**	.75	1.00	1.00	.99	1.00	1.00	.55
87	1.00	**.00**	.75	1.00	1.00	.87	1.00	1.00	.83
88	1.00	**.03**	.75	1.00	1.00	.73	1.00	1.00	.50
89	1.00	**.02**	.75	1.00	1.00	.83	1.00	1.00	.87
90	1.00	**.01**	.75	1.00	1.00	.87	1.00	1.00	.99
91	1.00	**.02**	.75	1.00	1.00	.98	1.00	1.00	.77
92	1.00	**.02**	.75	1.00	1.00	1.00	1.00	1.00	.80
93	1.00	**.00**	.75	1.00	1.00	.99	1.00	1.00	.96
94	1.00	**.02**	.75	1.00	1.00	.75	1.00	1.00	.96
95	1.00	**.01**	.75	1.00	1.00	.96	1.00	.99	.58
96	1.00	**.00**	.75	1.00	1.00	.71	1.00	1.00	.99
97	1.00	**.02**	.75	1.00	1.00	.98	1.00	1.00	.86
98	1.00	**.00**	.75	1.00	1.00	.98	1.00	1.00	.75
99	1.00	**.02**	.75	1.00	.99	1.00	1.00	1.00	.95
100	1.00	**.02**	.75	1.00	1.00	.81	1.00	.98	.93

Table B.4: Percentage of Passed Item Parameters of MMixRIM Based the Ration of MC Error to Standard Deviation

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
1	.96	.88	.90	1.00	.81	.90
2	.76	.90	.86	.98	.99	.95
3	.95	.89	.86	.90	.80	.76
4	.70	.74	.70	.80	.97	.99
5	.77	.81	.82	.94	.87	.93
6	.90	.71	.92	.96	.85	.96
7	.89	.76	.85	.93	.79	.72
8	.93	.90	.88	.92	.82	.97
9	.99	.61	.68	.99	.90	.79
10	.98	.56	**.47**	.98	.98	.94
11	.79	.72	.76	**.40**	.92	.66
12	.68	.82	.94	.98	.98	.97
13	.86	.82	.82	.79	.93	.89
14	.80	.90	.99	.81	.77	.74
15	.76	.71	.60	1.00	.73	.91
16	.88	.82	.86	.98	.92	.89
17	.77	.84	.89	.91	**.32**	.73
18	.83	.88	.80	.93	.96	.96
19	.80	.73	.80	.82	.89	.92
20	.94	.85	.96	.91	.77	.73
21	.93	.74	.85	.78	1.00	.96
22	.76	.86	.85	.96	.79	.92
23	.89	.95	.99	.89	.74	.61
24	.89	.89	.89	.80	.86	.73

Continued on next page

Table B.4 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
25	.79	.74	.90	.97	.97	.90
26	.80	.85	.82	.98	.84	.95
27	.94	.79	.86	.80	.95	.92
28	.93	.80	.86	.82	.88	.95
29	.82	.85	.94	.91	.97	.96
30	.92	.89	.80	.99	.99	1.00
31	.82	.83	.82	1.00	.82	.79
32	1.00	.96	.68	.74	.82	.93
33	.95	.87	.89	.83	.83	.88
34	.74	.81	.89	.96	.98	.67
35	.82	.78	.79	1.00	.93	.80
36	.93	.99	.93	.85	1.00	.95
37	.78	.98	.93	.98	1.00	.90
38	.98	.94	.82	1.00	.96	.92
39	.96	.94	**.11**	.96	.72	.93
40	.80	.85	.80	.99	.81	.77
41	.87	.84	.92	.99	.94	1.00
42	.80	.82	.88	.99	.98	.86
43	.88	.86	.73	.98	.96	.99
44	.82	.88	.98	.88	.99	.93
45	1.00	.98	.89	.94	.99	.95
46	.90	.82	.93	.99	.87	.89
47	.82	.60	.75	.84	.89	.93
48	.88	.90	.81	.96	.98	.99
49	.74	.76	.71	.88	.80	.79
50	.92	.73	.64	.89	.90	.86

Continued on next page

Table B.4 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
51	.72	.80	.53	1.00	.99	.96
52	.82	.90	.80	1.00	.93	.92
53	.98	.79	.91	.94	.87	.94
54	.87	.92	.69	.92	.91	.90
55	.85	.76	.78	.99	.93	.92
56	.90	.98	.89	.96	.93	.99
57	.97	.96	.93	.99	.95	.97
58	.96	.86	.73	.93	.95	.93
59	.91	.67	.98	.99	1.00	.93
60	.91	.65	.73	1.00	.99	.99
61	.92	.85	.90	.98	.80	.97
62	.77	.83	.68	.92	1.00	.72
63	.84	.60	.86	.95	.95	.73
64	.85	.67	.79	.99	.96	.91
65	.90	.85	.90	.95	.92	.89
66	.72	.99	.78	.93	.97	.99
67	.76	.88	.86	.95	.99	.77
68	.78	.61	.88	.97	.99	.93
69	.99	.85	.96	.92	.80	.93
70	.89	.92	.88	.98	.98	.90
71	.73	.83	.74	.73	.87	.74
72	.99	.95	.88	.96	.97	.99
73	.88	.87	.93	.88	.86	.86
74	.90	.86	.76	.96	.96	.86
75	.77	.92	.95	.98	.95	.73
76	.91	.90	.94	.86	.90	.82

Continued on next page

Table B.4 – continued from previous page

Replication	D2C2R0D	D2C2R3D	D2C2R6D	D2C2R0E	D2C2D3E	D2C2R6E
77	.63	.59	.74	.99	.95	.89
78	.66	.82	.86	.80	.74	.93
79	.86	.87	.90	.99	1.00	.85
80	.92	.80	.95	.89	.86	.86
81	.82	**.11**	.87	1.00	.92	.86
82	.87	**.30**	.88	.85	.86	.99
83	.89	.88	.79	.96	.99	.97
84	.99	.94	.87	.92	.90	.92
85	1.00	.97	.82	.98	.74	1.00
86	.83	.71	.69	.98	.89	.89
87	.84	.86	.82	.99	.93	.95
88	.95	.71	.92	.96	.88	.98
89	.83	.96	.91	1.00	.95	.99
90	.83	.90	.89	1.00	.86	.86
91	.79	.85	.95	.99	.90	.85
92	.84	.79	.79	.97	.84	.85
93	.98	.92	.93	.77	.88	.97
94	1.00	.80	.72	.70	.71	.70
95	.77	.93	.77	.99	.89	.79
96	.73	.77	.86	1.00	.95	.98
97	.88	.93	.86	.95	.95	.93
98	.84	.91	.90	.89	.79	.79
99	.98	.76	.78	.99	.85	.96
100	.92	.87	.90	1.00	.96	.98