2018

A Rotatable Asymmetric Variable Compensation MIRT Model

Xinchu Zhao
University of South Carolina - Columbia

Recommended Citation

Zhao, X.(2018). *A Rotatable Asymmetric Variable Compensation MIRT Model.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/5090

A Rotatable Asymmetric Variable Compensation MIRT Model

by

Xinchu Zhao

Bachelor of Education
University of Macau 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2018

Accepted by:

Brian Habing, Major Professor

David Hitchcock, Committee Member

Lianming Wang, Committee Member

Terry Ackerman, Outside Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Dedication

Dedicated

to

my grand parents,

Zhuxiu Zhao, Lanying Xia, Liansheng Liu, and Chuanfang Deng,

my parents,

Yong Zhao and Jiang Liu,

my wife,

Weizhou Tang

Acknowledgments

When I look back and think of my last five years, I find that I actually had a very joyful PhD journey (without loss of hair), as I was supported by many persons. Here, I write to thank all of them.

First, I would like to thank my advisor and committee chair, Dr. Brian Habing, for your valuable guidance and great flexibility. You always gave me the freedom to do whatever I wanted and provided me with great feedback. Your encouragement gave me a lot of confidence to present at conferences and finish my dissertation. I really enjoy working with you.

I would like to thank my doctoral committee members, Dr. Terry Ackerman, Dr. David Hitchcock, and Dr. Lianming Wang, who spent a lot of time on my dissertation and provided me with so many great comments to help me improve this work. I also want to thank Dr. Louis Roussos for providing me with the real data used in this dissertation.

I am deeply thankful to my family. I want to thank my grandparents and parents for fostering me, and giving me a lot of freedom to walk my own path. Thank you for always supporting me from the other side of the earth. Especially, I want to thank my wife, Weizhou Tang. We met, fell in love, and got married during our PhD programs. I really appreciate you for being a part of my life! I want to thank our pet dog, Lurker (the name comes from the PC game Starcraft), who became our family

member since I was in high school. Thank you for still remembering me when I came back to China. Do not eat too much!

I want to thank my friends in China and North America. Thank you for always being supportive and not asking “When do you graduate?” frequently. I wish you much happiness, and every success in your future.

Finally, I would like to thank the other persons who helped me pursue the PhD degree. Sorry for not listing all of you, but I will always remember your kindness.

As I still have some space here, I want to say “Hi!” to Dr. Habing’s future students. If you are seeking for acknowledgment samples, I hope this one can help you.

Abstract

The purpose of this study is to develop, estimate, and interpret a new variable compensation multidimensional item response theory (MIRT) model, named the Rotatable Asymmetric Variable Compensation Model (RAVCM), that allows for transformation between different correlation structures. Since the model is rotatable like the common compensatory models (CM), it is not necessary to specify or estimate the correlation of abilities to recover the model. Also, it can approximate the existing MIRT models well. In simulation, the RAVCM is shown to estimate the parameters with small error, especially when the non-compensatory model (NCM) is the true model and the correlation of abilities is misspecified, and when the test has a mixture of compensatory and non-compensatory items. In a real data study, the RAVCM demonstrates better fit, and provides an additional way to interpret the latent abilities from the compensatory model. The angle between item vectors in the RAVCM can be considered as a measure of compensation. Its effectiveness at distinguishing the CM and the NCM is evaluated and compared to some other common goodness of fit statistics in MIRT. Via a simulation study, it is shown that the RAVCM works well at both the test-level and the item-level in many cases. Two forms of the high dimensional RAVCM are proposed and discussed. Simulations show that the simple form is estimable. When the number of items and examinees are large, the estimates have smaller error, compared to the small-sample cases.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Item response theory (IRT) models have been widely used in educational measurement. These models describe the relationship between the probability of correct responses given by examinees to each item and examinees' abilities (usually expressed by θ). Traditionally, IRT entails three assumptions: unidimensionality, local independence, and monotonicity. Unidimensionality assumes that all items in a test measure the same latent trait of the examinees. Local independence means that given examinees' abilities, their responses to different items are independent. Monotonicity means that the probability of answering an item correctly is a monotone increasing function of ability.

The unidimensionality assumption is common in many classical IRT models such as the Rasch model (Rasch, 1966) or three-parameter logistic (3PL) model (Birnbaum, 1968). In practice, however, this assumption is often violated since educational and psychological tests usually do not strictly measure only one latent trait. For example, a GRE quantitative test measures math ability as well as English reading ability, especially for those examinees who are not native English speakers. Violation of the unidimensionality assumption could cause biased ability and item parameter estimation (e.g., Ackerman, 1989; Kirisci & Hsu, 1995). For these cases, multidimensional item response theory (MIRT) models that involve two or more latent traits may be more appropriate.

Traditionally, the MIRT models can be classified as compensatory models (CM) or non-compensatory models (NCM) (e.g., Ackerman, 1989; Way, Ashley, & Forsyth, 1988). The compensatory models allow one ability to compensate for lack of another, while the non-compensatory models require examinees to be high in all the required abilities to have a high probability of answering a question correctly. Variable compensation models (VCM) were developed to provide the middle ground between the CM and NCM (Ackerman & Bolt, 1995).

The compensatory MIRT models have been extensively developed and studied (e.g., Ackerman, Gierl, & Walker, 2003; Bolt & Lall, 2003; De La Torre & Patz, 2005; Li, Jiao, & Lissitz, 2012). Estimation of Non-compensatory models are studied in some research (e.g., Babcock, 2011; Bolt & Lall, 2003; Chalmers & Flora, 2014), but only Bolt and Lall (2003) give an example of real data application. The variable compensation model is studied in Simpson (2005). Beyond the challenge of estimating them, one disadvantage of the NCM and VCM is that they do not allow for transformation of the correlation structure of abilities. Without specifying the true correlation, it is hard to recover and interpret the parameters in those models.

This study develops a new model that allows for variable compensation and transformation between correlation structures. Compared to the existing MIRT models, the new model showed more flexibility. The detail of the new model, the problem it can solve, and some comparison between it and the other models and statistics are included in the following chapters. Chapter 2 includes introduction to the background of this research. In Section 2.1, the three types of existing MIRT models, their properties and estimation are reviewed. Section 2.2 introduces transformation between correlation structures. In Chapter 3, a new model is proposed to solve the transformation problem in section 2.2. Section 3.1 introduces the rotatable asymmetric variable compensation model (RAVCM) and its properties. Section 3.2 provides procedure and details of estimation of the RAVCM via MCMC methods. In Section 3.3 and 3.4, simulation and real data study results are shown and discussed. Section 3.5 discusses the definition of compensation and model interpretation. In Chapter 4, the effectiveness of goodness of fit statistics at distinguishing between the compensatory model and the non-compensatory model are compared, at the test-level and the item-level. Chapter 5 talks about estimation and interpretation of the high-dimensional RAVCM. Some simulation studies about the three-dimensional RAVCM are given.

Chapter 2

Background

2.1 Existing MIRT Models

To simplify estimation and interpretation, all models discussed below are two dimensional, two-parameter models (without a guessing parameter). There are many possible parameterizations of compensatory models and non-compensatory models, and this article only focuses on one version for each type of models.

A two dimensional, two-parameter compensatory MIRT model (CM; Ackerman, 1994; Reckase, 1985) can be written as:

$$P(X_{ij} = 1 | \boldsymbol{a_i}, \boldsymbol{\theta_j}, b_i) = \frac{1}{1 + e^{-1.7\boldsymbol{a_i}(\boldsymbol{\theta}_j - b_i)}}.$$

For convenience, in this dissertation we re-parameterize the model as

$$P(X_{ij} = 1 | \boldsymbol{a_i}, \boldsymbol{\theta_j}, b_i) = \frac{1}{1 + e^{-1.7(\boldsymbol{a_i}\boldsymbol{\theta_j} - b_i)}}, \tag{2.1}$$

where

X_{ij} is the 0-1 score on item i by person j;

$\boldsymbol{a_i} = (a_{1i}, a_{2i})$ is the vector of discrimination parameters of item i;

b_i is the difficulty parameter of item i;

$\boldsymbol{\theta_j} = (\theta_{1j}, \theta_{2j})$ is the vector of ability parameters of examinee j; and

1.7 is the scaling constant that minimizes the difference between normal distribution function and logistic function. Note that the b_i's in the two parameterizations are different.

In this model, $a_{1i}\theta_{1j} + a_{2i}\theta_{2j}$ can be represented by $\boldsymbol{a_i^t\theta_j}$, the inner product of item vector and the ability vector. The direction and length of the vector $\boldsymbol{a_i}$ can be interpreted as the composite of abilities the item measures and the degree of multidimensional discrimination, respectively (Ackerman, 1994). This item vector can also be considered as the vector of factor loadings in factor analysis. Due to the indeterminacy of the difficulty parameter from each dimension, only a single difficulty parameters (b_i in the model) can be estimated for each item.

Another parameterization of the CM (Reckase, 1997) is

$$P(X_{ij} = 1|\boldsymbol{a_i}, \boldsymbol{\theta_j}, b_i) = \frac{1}{1 + e^{-1.7(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} - ||a_i||b_i)}},$$

which will not be used in this dissertation.

A two dimensional, two-parameter non-compensatory MIRT model (NCM; Sympson, 1978) can be written as:

$$P(X_{ij} = 1|a_{1i}, a_{2i}, b_{1i}, b_{2i}, \theta_{1j}, \theta_{2j}) = \left\{\frac{1}{1 + e^{-1.7a_{1i}(\theta_{1j} - b_{1i})}}\right\}\left\{\frac{1}{1 + e^{-1.7a_{2i}(\theta_{2j} - b_{2i})}}\right\}.$$

Similar to the CM, in this dissertation we re-parameterize the NCM as:

$$P(X_{ij} = 1|a_{1i}, a_{2i}, b_{1i}, b_{2i}, \theta_{1j}, \theta_{2j}) = \left\{\frac{1}{1 + e^{-1.7(a_{1i}\theta_{1j} - b_{1i})}}\right\}\left\{\frac{1}{1 + e^{-1.7(a_{2i}\theta_{2j} - b_{2i})}}\right\}, \tag{2.2}$$

where

X_{ij} is the 0-1 score on item i by person j,

a_{1i} and a_{2i} are the discrimination parameters from each dimension of item i,

b_{1i} and b_{2i} are the difficulty parameters from each dimension of item i, and

θ_{1j} and θ_{2j} are the ability parameters of examinee j.

In contrast to the compensatory model, the non-compensatory model has a different difficulty parameter for each ability dimension, and the probability of a correct response is the product of probabilities for each dimension. For this model, the discrimination parameters are denoted by scalars, since the model cannot be simply expressed by the inner product of item vector and ability vector.

The difference between the CM and the NCM is that the CM allows for compensation between abilities. Since the CM is a function of the summation of $a_{1i}\theta_{1j}$ and $a_{2i}\theta_{2j}$, when one ability parameter is small, an examinee can still have large probability for giving a correct response if the other ability parameter is large enough, assuming non-zero a_{1i} and a_{2i}. This compensation is not allowed by the NCM. Since each item in the NCM can be viewed as a product of two unidimensional items, the probability of correct response cannot exceed the value given by each "unidimensional item". Therefore, when the NCM is the true model, to have a large probability of correct response, an examinee need to have large value in all θ's, and there is no compensation. The non-compensatory model can also be considered as a multi-step model. Each unidimensional part can be treated as an independent step in answering the item. Failing to correctly answer one or more steps leads to an incorrect response.

To allow for more flexibility in compensation, the two-parameter, two dimensional variable compensation model (VCM; Ackerman & Bolt, 1995) has been introduced:

$$\begin{aligned} &P(X_{ij} = 1|a_{1i}, a_{2i}, b_{1i}, b_{2i}, \theta_{1j}, \theta_{2j}) \\ = &\frac{e^{1.7(a_{1i}\theta_{1j}+a_{2i}\theta_{2j}-b_{1i}-b_{2i})}}{1+e^{1.7(a_{1i}\theta_{1j}+a_{2i}\theta_{2j}-b_{1i}-b_{2i})}+\lambda\big[e^{1.7(a_{1i}\theta_{1j}-b_{1i})}+e^{1.7(a_{2i}\theta_{2j}-b_{2i})}\big]}, \end{aligned} \tag{2.3}$$

where λ, which takes value between 0 and 1, can be considered as the level of non-compensation the VCM has. The VCM is the CM if $\lambda = 0$, and is the NCM if $\lambda = 1$. The other parameters have the same notation as in the NCM.

In unidimensional IRT, the probability of correctly answering an item can be graphically expressed by an item characteristic curve (ICC; e.g., Hambleton & Swaminathan, 2013). In MIRT, this probability is expressed by item response surface (IRS), which can be graphically displayed by probability contour plots in two-dimensional case (Ackerman, 1996). An example of contour plots of these three models are shown in Figure 2.1. The CM has linear and parallel equal probability contours. The NCM and VCM have curved equal probability contours.

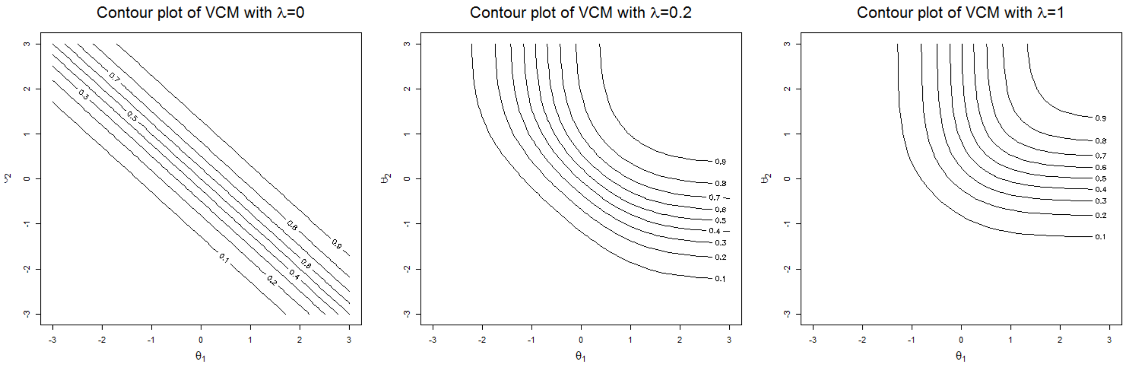

Figure 2.1 The contour plots of compensatory ($\lambda = 0$), variable compensation ($\lambda = 0.2$) and non-compensatory ($\lambda = 1$) model,with $a_1 = a_2 = 1$ and $b_1 = b_2 = 0$.

2.2 Transformation Between Different Correlation Structures

For the compensatory model, an item can be graphically represented as an item vector $\boldsymbol{a_i}$ (Reckase & McKinley, 1991). The length of the vector is the multidimensional discrimination, MDISC$_i$. In two dimensional case, it can be written as:

$$\text{MDISC}_i = (a_{1i}^2 + a_{2i}^2)^{1/2}.$$

The direction of the item vector, measured by the angle of this vector from the positive θ_1 axis, can be written as:

$$\alpha_i = arccos\left(\frac{a_{1i}}{\sqrt{a_{1i}^2 + a_{2i}^2}}\right)^{1/2}.$$

Each angle represents a different composites of abilities in the (θ_1, θ_2) space. When this angle is 0 or $\pi/2$, the item only measures one ability in the space. Otherwise, it measures both the two abilities in the space to some degree.

The compensatory model has no assumption on the correlation between θ_1 and θ_2 (Ackerman, 1994). In this model, the space, or axes system, of the latent abilities θ_1 and θ_2 can be arbitrarily rotated, simply using matrix multiplication. The transformation from correlated to orthogonal abilities can be expressed by matrix multiplication (e.g., Oshima, Davey, & Lee, 2000; Smith, 2009):

$$\boldsymbol{\theta_j = \Sigma\theta_j^*,} \tag{2.4}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$ are abilities values in correlated and uncorrelated spaces. The transformation (rotation) matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_2 & \psi_1 \end{bmatrix},$$

where $\psi_1 = \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2}$ and $\psi_2 = \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}$, with the correlation $0 \leqslant \rho \leqslant 1$. To insure the probability in (2.1) remains the same, a transformation of the discrimination parameters is needed:

$$\boldsymbol{a_i^t} = (\boldsymbol{a_i^*})^t\boldsymbol{\Sigma}^{-1}. \tag{2.5}$$

Since $\boldsymbol{a_i^t\theta_j} = (\boldsymbol{a_i^*})^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma\theta_j^*} = (\boldsymbol{a_i^*})^t\boldsymbol{\theta_j^*}$, the probability of correct response given by transformed and untransformed parameters are the same. Therefore, this transformation has no influence on model fitting.

Since $\boldsymbol{\Sigma}$ is an invertible matrix, this transformation is invertible. One can transform the parameter back to correlated space, simply exchanging the roles of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$. Moreover, by applying both the original transformation and inverse transformation, one can transform parameters in a correlation structure to those in uncorrelated space, then to parameters under any other correlation structures. Therefore, transformations between any correlation structures are possible.

For graphical convenience, the θ_1 axis is typically represented as orthogonal to the θ_2 (Ackerman, 1994). In plots, The effect of transformation is reflected in the item vectors. With the transformation (Equation 2.4), the discrimination MDISC$_i$ and the reference angle α_i will change. Figure 2.2 demonstrates the transformation between correlated space and uncorrelated space. In the correlated (θ_1^*, θ_2^*) space, the red and the blue item only measure θ_1^* and θ_2^*, respectively. The black item (with item vector in the middle) equally measures two abilities. As the abilities are transformed to uncorrelated space, the red (on θ_1 axis) and the blue (on θ_2 axis) item vector are rotated towards the black item vector, thus measure both the two uncorrelated abilities (θ_1, θ_2). The direction of the black item remains the same, while its MDISC$_i$ becomes greater in this space.

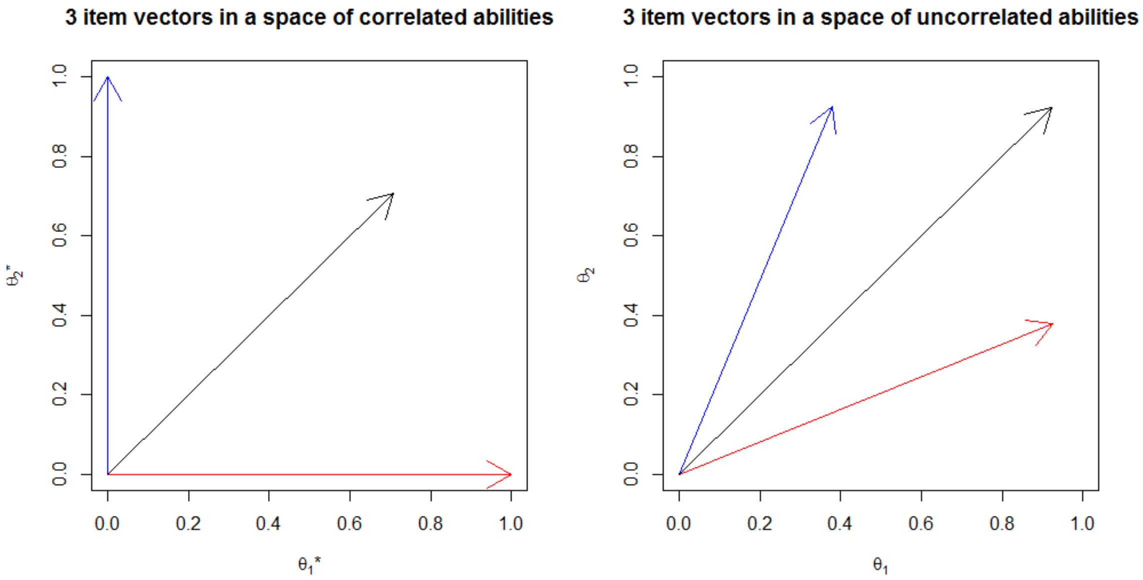

Figure 2.2 Three items in correlated and uncorrelated axes systems.

Since the transformation is one-to-one, parameters sets from two different correlation structures are also linked one-to-one. This transformation allows for estimation using an ability distribution with an arbitrary correlation. When interpreting the latent traits, one can transform the ability estimates using (Equation 2.4) to ensure that they have a certain correlation structure. This is not the case for the non-compensatory model and the variable compensation model. For those models, any transformation involving rotation changes the probabilities in the NCM (Equation 2.2) and VCM (Equation 2.3) due to their multiplicative structure. Therefore, without specifying the correlation structure between the latent traits, it is difficult to estimate and interpret the parameters in (Equation 2.2) and (Equation 2.3). As such, neither model has found wide use for analyzing real data sets. Similarly, the common rotations in factor analysis and multivariate statistics, such as Varimax (Kaiser, 1958) and Procruste (Dean, 2000) rotation can be applied on the CM (e.g., Bolt & Lall, 2003), but not the NCM and VCM.

The motivation of this study is to develop a model allows for different compensation levels, and also for transformation between correlation structures.

Chapter 3

Rotatable Asymmetric Variable Compensation MIRT Model

3.1 The New Model and its Properties

To develop a non-compensatory MIRT model that allows for transformation between different correlation structures, we add additional discrimination parameters to the NCM (Equation 2.2). Because of the rotatability and flexibility of the new model, it is named the Rotatable Asymmetric Variable Compensation Model (RAVCM). It can be written as:

$$\begin{aligned}&P(X_{ij}=1|a_{1i},a_{2i},a_{3i},a_{4i},b_{1i},b_{2i},\boldsymbol{\theta_j})\\=&\left\{\frac{1}{1+exp[-1.7(a_{1i}\theta_{1j}+a_{2i}\theta_{2j}-b_{1i})]}\right\}\left\{\frac{1}{1+exp[-1.7(a_{3i}\theta_{1j}+a_{4i}\theta_{2j}-b_{2i})]}\right\}.\end{aligned} \tag{3.1}$$

The item response function of this new model can be viewed as a product of item response functions of two compensatory items. To avoid indeterminacy, it is assumed that the angle of the first item vector from the θ_1 axis is smaller than the angle of the second item vector. That is, $a_{1i}/(a_{1i}^2+a_{2i}^2) \geqslant a_{3i}/(a_{3i}^2+a_{4i}^2)$. Similarly to the compensatory model, the correlation structure in the RAVCM can be transformed via Equation 2.4, and the probability in (Equation 3.1) can be kept by transforming the two item vectors (a_{1i}, a_{2i}) and (a_{3i}, a_{4i}) by multiplying the inverse transformation matrix in (5) separately.

When a_{2i} and a_{3i} in the RAVCM (Equation 3.1) are 0, the new model (Equation 3.1) and the non-compensatory model (Equation 2.2) are the same. If the RAVCM (Equation 3.1) is non-compensatory with correlated abilities, its two item vectors are exactly on the θ_1 and θ_2 axes, respectively. After transforming the abilities to uncorrelated space, the two items vectors are rotated towards the $\theta_1 = \theta_2$ line, and thus the RAVCM (Equation 3.1) is no longer a standard non-compensatory model in this uncorrelated space. Therefore, the NCM (Equation 2.2) is also a special case of the RAVCM (Equation 3.1), under a specific correlation structure. In the NCM, the true item parameters are unique, and they can be obtained only if this correlation is correctly specified. In the RAVCM, the item parameters can be recovered under any correlation structure. In Figure 3.1, the different representations of the same item in spaces with different correlation structures are shown. The item response function of the item is the NCM when $\rho = 0.6$.

The RAVCM cannot be exactly matched up with the compensatory model (Equation 2.1) or the variable compensation model (Equation 2.3). However, like the VCM (Equation 2.3), the RAVCM (Equation 3.1) allows for different compensation levels. One extreme case is that when $a_1/a_2 = a_3/a_4$, it is compensatory with linear equal probability contours. With proper parametrization, the item response surface of the CM (Equation 2.1) can be approximated well by the RAVCM (Equation 3.1). For any item in the CM with parameters $a_{1,CM}$, $a_{2,CM}$ and b_{CM}, let the parameters in the RA-VCM be $a_1 = a_3 = 0.739a_{1,CM}$, $a_2 = a_4 = 0.739a_{2,CM}$ and $b_1 = b_2 = 0.739b_{CM} - 0.559$, then the maximum difference of the item response functions of the CM and the RA-VCM is approximately 0.037. (This difference is calculated numerically.) Figure 3.2 shows the contour plots of a CM and a RAVCM approximates it. Figure 3.3 shows the contour plot of the probability difference between the CM and the RAVCM. In section 3.3, it is shown that when the CM is the true model, the average errors of the fitted CM and RAVCM are very similar.

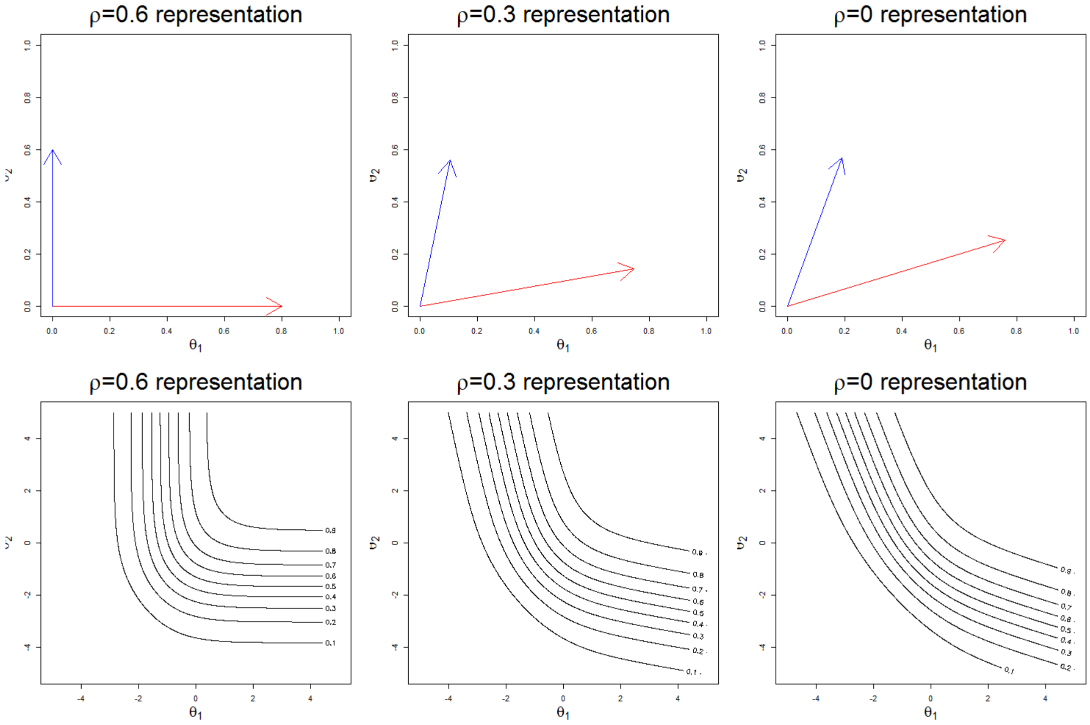

Figure 3.1 The different representation of item vectors and contour plots of a single item under spaces as viewed using different correlation structures.

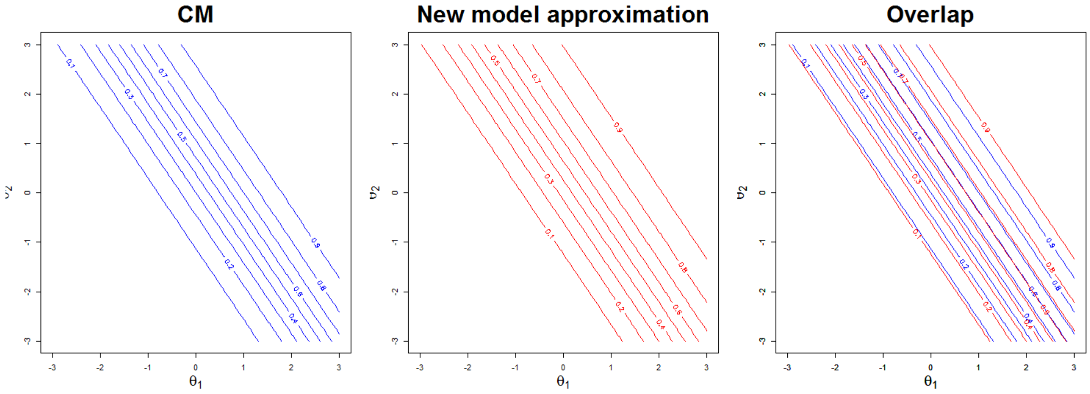

Figure 3.2 The contour plots of: (a) CM with $a_{1,CM} = 1$, $a_{2,CM} = 0.7$ and $b_{CM} = 0.5$; (b) New model approximation with $a_1 = a_3 = 0.739 * a_{1,CM}$, $a_2 = a_4 = 0.739 * a_{2,CM}$ and $b_1 = b_2 = 0.739 * b_{CM} - 0.559$; (c) The overlapped contour plots.

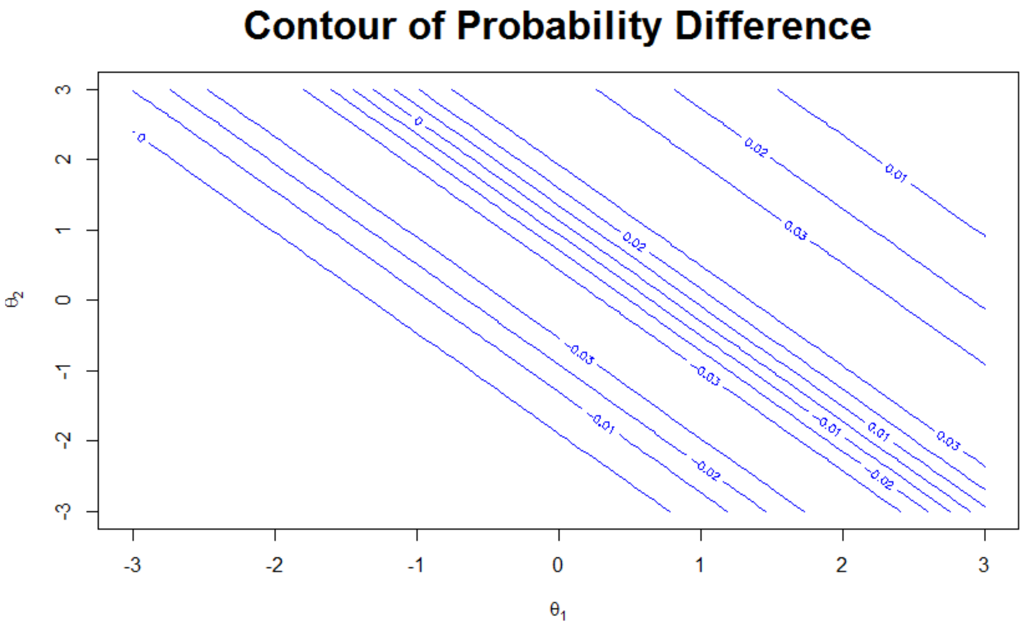

Figure 3.3 The contour plots of the probability difference of CM with $a_{1,CM} = 1$, $a_{2,CM} = 0.7$ and $b_{CM} = 0.5$ and RAVCM approximation with $a_1 = a_3 = 0.739 * a_{1,CM}$, $a_2 = a_4 = 0.739 * a_{2,CM}$ and $b_1 = b_2 = 0.739 * b_{CM} - 0.559$;

When $0 < a_2 < a_1$ and $0 < a_3 < a_4$, the contour plot of the RAVCM is similar to the variable compensation model. Moreover, the RAVCM allows for asymmetric compensation, which means that the compensation levels can be different for different portions of the ability distribution. When one of the item vectors is close to an axis, for example, the θ_1 axis, and the other one is close to the line $\theta_1 = \theta_2$, the compensation level of "θ_1 to θ_2" is larger than "θ_2 to θ_1". In other words, when θ_2 is small, a large θ_1 can raise the probability in (Equation 3.1), but when θ_1 is small, a large θ_2 cannot do the same thing effectively. One extreme example is that, if $a_1 = 1$, $a_2 = 0$, $a_3 = 1$, $a_4 = 1$, then θ_1 can compensate for θ_2 but θ_2 cannot compensate for θ_1. In this case, one half of the contour plot is similar to the CM and the other half is like the NCM (Figure 3.4). In this case, θ_1 can compensate for θ_2 but θ_2 cannot compensate for θ_1. An example of asymmetric compensation could be a math test with an item requiring reading skills: reading skill helps only when an examinee has enough math ability.

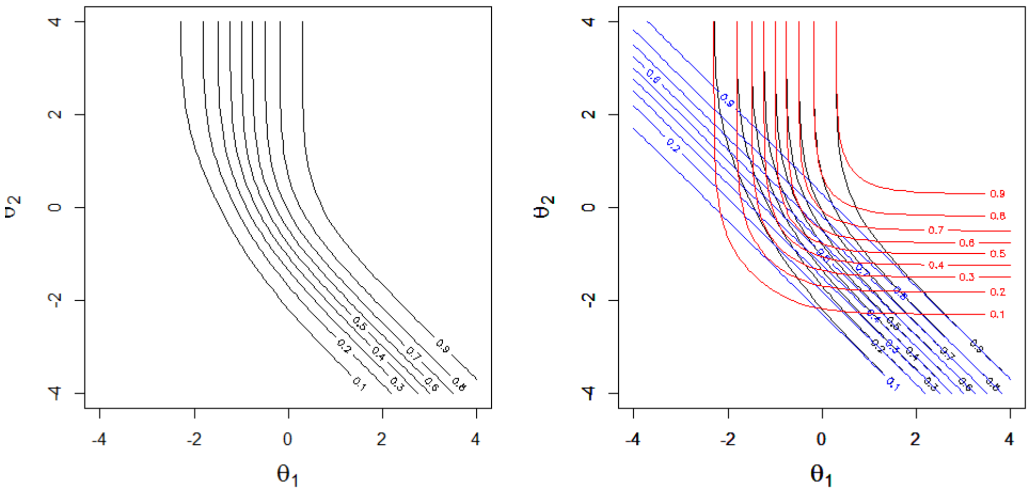

Figure 3.4 θ_1 compensates θ_2 only. The second plot is overlapped with CM (blue) and NCM (red).

The RAVCM (Equation 3.1) can also be considered as a multi-step model since it is multiplicative. Its a parameters can be interpreted as the amount of ability used in each step. In terms of item response surface, (a_1, a_2) and (a_3, a_4) determines the angle and density of the upper and lower half of the contours. As in Figure 3.4, when the first item vector $(a_1, a_2) = (1, 0)$ is on the θ_1 axis, the top half of the contours are perpendicular to the θ_1 axis. The angle between the second item vector $(a_3, a_4) = (1, 1)$ and the θ_1 axis is 45 degree, so the angle between the lower half of the contours and the θ_1 axis is 45 degree. The second item vector is longer than the first, therefore the lower half of the contours is denser.

In the RAVCM (Equation 3.1), a measure of the total compensation level can be defined by the angle between the two component item vectors. This angle is 0 when the model is a compensatory model (RAVCM with $a_1/a_2 = a_3/a_4$), and $\pi/2$ when the model is the NCM (Equation 2.2). Since the compensation level is not stable under rotations that change the correlation structure, to compare those of different items, these angles should be obtained in the uncorrelated space.

3.2 Estimation of the RAVCM

3.2.1 Markov Chain Monte Carlo for Estimating MIRT Models

In recent research, Markov chain Monte Carlo (MCMC) methods have been widely used to recover parameters of IRT models (e.g., Albert, 1992; Babcock, 2009, 2011; Baker, 1998; Beguin & Glas, 2011; Bolt & Lall, 2003; Kim & Bolt, 2007; Patz & Junker, 1999; Simpson, 2005). MCMC methods attempt to directly draw samples from the parameters' joint Bayesian posterior distribution, and thus do not require any complicated calculations such as derivatives and integrals. Therefore, as model complexity increases, MCMC methods can still be easily applied, while the other methods such as marginal maximum likelihood (MML) with the EM algorithm (e.g., Bock & Aitkin, 1981) becomes too messy (Wollack, Bolt, Allan, & Lee, 2002).

The non-compensatory model is more difficult to estimate than the compensatory model due to its multiplicative structure and separate difficulty parameters for each dimension. However, in past decade, several researchers have shown that the NCM was estimable using MCMC methods. For examples, Bolt and Lall (2003) used the Metropolis-Hastings algorithm to estimate the multicomponent latent trait model (MLTM; Whitely, 1980), which is a version of NCM with fixed discrimination parameters. Babcock used a modified Metropolis-Hastings algorithm (2009) and Metropolis-Hastings within Gibbs sampling (2011), and Chalmers and Flora (2014) used the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm to estimate the NCM. They showed that the correlation between abilities can be well recovered with enough unidimensional items on each dimensions. In their simulation, Babcock used 2, 6, or 10 unidimensional items and 50 multidimensional items; Chalmers and Flora had 5, 10, or 15 unidimensional items and 10 non-compensatory items. Wang and Nydick (2015) compared the performance of the MCMC and MH-RM on a different version of NCM (restricted C-MIRT) and concluded that the MCMC gives better estimates when the correlation between abilities are large. For the variable compensation model, Simpson (2005) estimated a simplified version of it (the G-MIRT model) with fixed discrimination parameters using MCMC through WinBUGS. The estimation results were not very satisfactory since many MCMC chains did not converge. The problem might be caused by the λ parameter. Note that, even with small λ (e.g., 0.2), the contour plot of the VCM can still be very similar to that of the NCM (e.g., Figure 2.1). Therefore, it is not easy to find an appropriate prior of λ. Fu (2015) suggested using ad-hoc prior distributions for λ to make its Markov chain converges.

The RAVCM (Equation 3.1) has a similar mathematical structure to the NCM (Equation 2.2). It does not contain the compensation parameter λ, and thus can avoid any problems due to it. The estimation challenge comes from its large number of parameters and great flexibility. Here, the estimation of the RAVCM is implemented via the Metropolis-Hastings algorithm within Gibbs sampling (e.g., Babcock, 2011; Patz & Junker, 1999). The parameters are first divided into several Gibbs sampling subsets, then generated iteratively by Metropolis-Hastings algorithm, with proper prior and proposal distributions.

WinBUGS and openBUGS are software designed for Bayesian analysis. They were used in several past studies to recover parameters (e.g., Bolt & Lall, 2003; Patz & Junker, 1999; Simpson, 2005). However, they take too much time if the chain is long and the number of parameter is large. In this study, the MCMC procedure is implemented using the Rcpp, a package for writing and running C++ code in R. It can use the CPU more efficiently, and thus greatly increase computational speed. For 30 items and 1000 examinees, it only takes the Rcpp program about 20 minutes to run 30000 iterations on a normal personal computer with 3.20 GHz CPU, which is many times shorter than BUGS.

3.2.2 Settings and Details

An estimation challenge of the RAVCM (Equation 3.1) is its indeterminacy. For the NCM (Equation 2.2), MCMC estimations may switch between the two dimensions over stages of the chains (Bolt & Lall, 2003). This means that at some point of the Markov chain, the parameters of one dimension come to represent those of the other dimension, and vice versa (e.g., Figure 3.5). An additional problem happens on the RAVCM is that the item parameters (a_{1i}, a_{2i}, b_{1i}) and (a_{3i}, a_{4i}, b_{2i}) can be exchanged without changing the probability in (Equation 2.1), and without changing ability parameters θ_{1j}, θ_{2j}. Thus, two types of "label switching" may happen on the RAVCM: dimension label switching of all the parameters, and item parameter switching of parameters of a single item. This may cause instability of the MCMC chains. To avoid the former type of "label switching", identification constraints should be added. For the CM, this study uses a common method that is to fix the first item vector on the θ_1 axis (e.g., Fraser & McDonald, 1988; Bolt & Lall, 2003). For the models containing two item vectors, such as the RAVCM and its special case, the NCM, this study fixes only one vector on the θ_1 axis, and let the other one be free, since in practice it is not possible to fix more than one vector unless there exist many known unidimensional items. The difficulty parameter(s) of the first item in each model is also fixed at 0 (Bolt & Lall, 2003). To avoid the later type of "label switching", it is assumed that $a_{1i}/(a_{1i}^2 + a_{2i}^2) \geqslant a_{3i}/(a_{3i}^2 + a_{4i}^2)$ to ensure that the first item vector has smaller angle from the θ_1 axis. (When this assumption is violated by the generated a values in the chain, these values are rejected with probability 1.) Additionally, the other settings and details of the MCMC estimation need to be adjusted carefully.

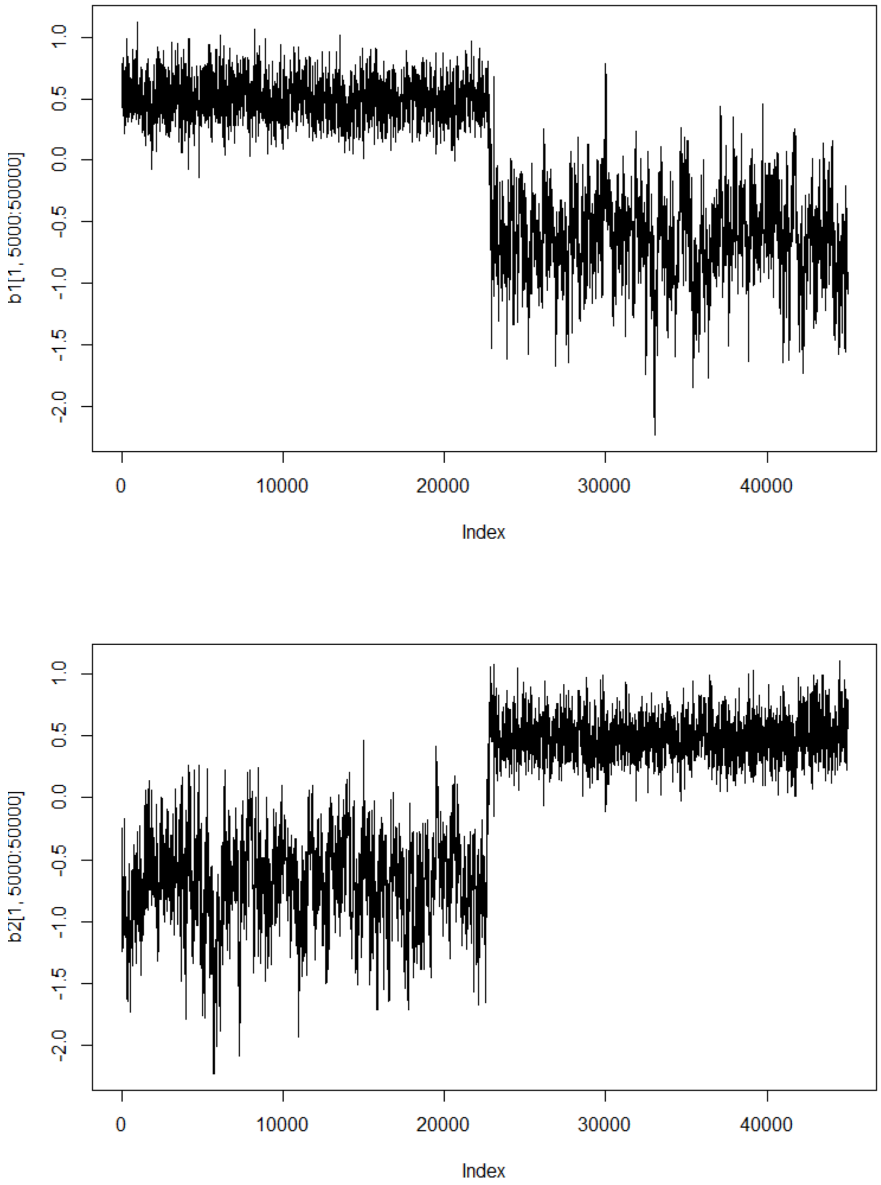

Figure 3.5 An example of label switching in MCMC. The top and bottom graph show the generated value of a b_1 and b_2 (for the same item), respectively.

Gibbs Sampling Subsets

It is important to choose proper parameter subsets of Gibbs sampling to keep the balance between calculation speed and convergence speed. Although having fewer subsets lead to faster calculation in one iteration, more iterations are needed to stabilize the chain. Patz and Junker (1999) suggested putting all of the ability parameters in one subset and putting all of item parameters in the other subset. In this study, the ability parameters, discrimination parameters and difficulty parameters are divided into three different subsets. One reason is that, compared to the NCM, it takes longer to get converging chains due to the additional discrimination parameters in the RAVCM. With more subsets, the acceptance rate of the Metropolis-Hastings can be adjusted by changing the variance of the proposal distribution separately for each group of parameters. This can make the chains converge faster.

Prior Distributions

As mentioned previously, it is assumed that the first item vector has smaller angle from the θ_1 axis. The prior distributions of a_{3i} and a_{2i} should be set to be stochastically smaller than, or equal to the priors of a_{1i} and a_{4i}, respectively. On the other hand, to allow for more flexibility, the variance of the prior distributions of a_{2i} and a_{3i} should not be too small, unless it is known that the true model is close to the NCM. This study takes the truncated normal distribution with mean 1 and variance 1 as the prior distribution for a_{1i} and a_{4i}. (It is truncated such that the a parameters are greater than 0.) For a_{2i} and a_{3i}, it takes Exponential (1.5) when the true model is the CM (1) or the tests have complicated structures, and takes the Exponential (2) as the prior distribution when the true model is the NCM (Equation 1.2) or the VCM (3). These priors are non-negative, to avoid the negative estimates for discrimination parameters. For the ability parameters, Bolt and Lall (2003) used multivariate normal priors with covariance matrix equal to the identity matrix for the CM, and estimated the covariance matrix for the NCM. In this study, since the RAVCM (Equation 2.1) allows for transformation between correlation structures like CM, the covariance of the normal priors are set to be 0. Normal priors with mean 0 and variance 2 are assigned to the difficulty parameters.

Proposal distributions

The proposal distributions decide how the values are generated in the chain. It is usual to set the mean of proposal distribution to the last value generated, while the variances of the proposal distributions should be carefully selected for each step so that the acceptance rate of the proposed value is at a reasonable level. When the variance is too large, the acceptance rate is usually low since the proposal distribution tends to generate values far away from the former accepted one, then the chain hardly moves. If the variance is too small, the chain will move very slowly thus the whole algorithm will be inefficient. Following Gelman, Roberts, and Gilks (1996), the optimal acceptance rate for normal proposal distribution is about 0.352 for two parameters, and 0.279 for 4 parameters. In this study, the variances of the proposal distributions are selected such that the acceptance rates are generally between 20% and 30% for a parameters and 30% to 40% for θ and b parameters. 3000 preliminary iterations were performed to adjust proposal distribution variances. If the acceptance rates are out of the desired range, the procedure can automatically adjust the variance of the proposal distribution of the subset, and repeat this procedure until the acceptance rates of all subsets are ideal.

3.2.3 Algorithm

Notations of this section

The whole test result is recorded by a matrix denoted by $\boldsymbol{X}$. The result of item i answered by responser j is denoted by X_{ij}, the element on row i and column j. It is 1 when the answer is correct, 0 when the answer is incorrect. Then the row X_i is a sequence of all examinees' results for item i and the column X_j is the results of all items for examinee j.

The ability parameters of examinee j are denoted by the vector $\boldsymbol{\theta_j}$=(θ_{1j}, θ_{2j}). The difficulty parameters are denoted by the vector $\boldsymbol{b_i}$=(b_{i1}, b_{i2}). Then discrimination parameters are denoted by the vector $\boldsymbol{a_i}$=(a_{i1}, a_{i2}, a_{i3}, a_{i4}). The vector of all item parameters of item i is denoted by $\boldsymbol{\gamma}_i$. Combining the parameters of all items or examinees, we have the matrices $\boldsymbol{\theta}$, $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\gamma}$. The variances of proposal distributions of θ, a, and b parameters are denoted by c_θ^2, c_a^2 and c_b^2, respectively.

Algorithm

1. Generate θ_j^k from $p(\boldsymbol{\theta}_j|\boldsymbol{\gamma}^{k-1}, \boldsymbol{X})$ for each j:

a. Generate $\boldsymbol{\theta}_j^* = (\theta_{1j}^*, \theta_{2j}^*)$ from $N(\theta_{1j}^{k-1}, c_\theta^2)$ and $N(\theta_{2j}^{k-1}, c_\theta^2)$ for each j.

b. For each j, accept $\boldsymbol{\theta_j^k} = \boldsymbol{\theta}_j^*$ with probability

$$\begin{aligned}
&\alpha(\boldsymbol{\theta}_j^{k-1}, \boldsymbol{\theta}_j^*) \\
=& \frac{p(X_j|\boldsymbol{\theta}_j^*, \boldsymbol{\gamma}_j^{k-1})\pi(\boldsymbol{\theta}^*)}{p(X_j|\boldsymbol{\theta}_j^{k-1}, \boldsymbol{\gamma}_j^{k-1})\pi(\boldsymbol{\theta}^{k-1})} \wedge 1 \\
=& \frac{[\prod_i P_{ij}(\boldsymbol{\theta}_j^*, \boldsymbol{\gamma}_i^{k-1})^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}_j^*, \boldsymbol{\gamma}_i^{k-1}))^{1-X_{ij}}]exp(-\frac{(\theta_{1j}^*)^2+(\theta_{2j}^*)^2}{2})}{[\prod_i P_{ij}(\boldsymbol{\theta}_j^{k-1}, \boldsymbol{\gamma}_i^{k-1})^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}_j^{k-1}, \boldsymbol{\gamma}_i^{k-1}))^{1-X_{ij}}]exp(-\frac{(\theta_{1j}^{k-1})^2+(\theta_{2j}^{k-1})^2}{2})} \wedge 1,
\end{aligned}$$

where P_{ij} is given by Equation 2.1 in section 2.1. π is the prior distribution. When $\boldsymbol{\theta}_j^*$ is rejected, take $\boldsymbol{\theta}_j^k = \boldsymbol{\theta}_j^{k-1}$.

2. Generate $\boldsymbol{a}_i^k$ from $p(a|\boldsymbol{\theta^k}, \boldsymbol{b}^{k-1}, \boldsymbol{X})$ for each i:

a. Generate discrimination parameters (a_{il}^*) from $N(a_{il}^{k-1}, c_a^2)$ separately for each i, where $l = 1, 2, 3, 4$.

b. For each i, accept $\boldsymbol{a}_i^k = \boldsymbol{a}_i^*$ with probability

$$\begin{aligned}&\alpha(\boldsymbol{a}_i^{k-1}, \boldsymbol{a}_i^*)\\ &= \frac{[\prod_j P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^*, \boldsymbol{b}_i^{k-1})^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^*, \boldsymbol{b}_i^{k-1}))^{1-X_{ij}}] \prod_{l=1}^4 \pi(a_{il}^*)}{[\prod_j P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^{k-1}, \boldsymbol{b}_i^{k-1})^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^{k-1}, \boldsymbol{b}_i^{k-1}))^{1-X_{ij}}] \prod_{l=1}^4 \pi(a_{il}^{k-1})} \wedge 1.\end{aligned}$$

When $\boldsymbol{a}_i^*$ is rejected, take $\boldsymbol{a}_i^k = \boldsymbol{a}_i^{k-1}$

3. Generate $\boldsymbol{b}_i^k$ from $p(b|\boldsymbol{\theta^k}, \boldsymbol{a}^k, \boldsymbol{X})$ for each i:

a. Generate difficulty parameters (b_{il}^*) from $N(b_{il}^{k-1}, c_b^2)$ separately for each i, where $l = 1, 2$.

b. For each i, accept $\boldsymbol{b}_i^k = \boldsymbol{b}_i^*$ with probability

$$\begin{aligned}&\alpha(\boldsymbol{b}_i^{k-1}, \boldsymbol{b}_i^*)\\ &= \frac{[\prod_j P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^k, \boldsymbol{b}_i^*)^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^k, \boldsymbol{b}_i^*))^{1-X_{ij}}] \prod_{l=1}^2 \pi(b_{il}^*)}{[\prod_j P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^k, \boldsymbol{b}_i^{k-1})^{X_{ij}}(1 - P_{ij}(\boldsymbol{\theta}^k, \boldsymbol{a}_i^k, \boldsymbol{b}_i^{k-1}))^{1-X_{ij}}] \prod_{l=1}^2 \pi(b_{il}^{k-1})} \wedge 1.\end{aligned}$$

When $\boldsymbol{b}_i^*$ is rejected, take $\boldsymbol{b}_i^k = \boldsymbol{b}_i^{k-1}$.

4. Repeat step 1, 2 and 3 with a large number of iterations (e.g., 30000).

3.3 Simulation

Simulation studies are used to examine the parameter recovery of the RAVCM under different true models, and to check the effect of rotation between correlation structures. To compare the estimates of the parameters in different models, all models are estimated by the same method, i.e., Metropolis Hasting within Gibbs sampler. The choices of prior distributions are also similar. When estimating the CM, NCM, and VCM, the prior distributions of a, b, and θ are set to be normal, which is the same as the priors of the RAVCM. In different simulations, the ability parameters (θ_1, θ_2) are all generated from multivariate normal, with different correlation. The a parameters are generated from the truncated normal distribution with mean 1 and variance 1. To keep the monotonicity of the model, it is assumed that a parameters are not smaller than 0. If a generated value is negative, it will be re-generated. The difficulty parameters (b) are generated from normal distribution with variance 1.5 and mean 0 for the CM, mean -1 for the NCM, due to the large overall difficulty of the NCM (Bolt & Lall, 2003).

The data sets are simulated according to the generated parameters and one of the existing MIRT models. The Cronbach's alpha of all data sets are greater than 0.9. The proportion of correct answers of all data sets are approximately 0.5. Four sample sizes are considered: (a) 500 examinees, 15 items; (b) 500 examinees, 30 items; (c) 1000 examinees, 15 items; (d) 1000 examinees, 30 items. In the first simulation study, to compare the performance of the CM, NCM and RAVCM under different true models, the data sets are generated from the CM, NCM or the mixture of the CM and NCM. In the second study, to assess ability to rotate, the data sets are generated from the NCM under different correlations of abilities. To compare the estimates given by different models under the same scale and direction, the estimated abilities are scaled or Procrustes rotated and scaled according to the shape of generated abilities. For rotatable models such as the CM and RAVCM, the Procrustes rotation does not change the fitted probabilities when the item vectors are inversely rotated. However, for the NCM, since rotation changes the fitted probabilities, Procrustes rotation cannot be applied.

The following statistics are included in the results. The first one is root mean square errors (RMSEs) of ability estimates, defined as

$$RMSE(\theta_{.}) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\theta_{.j} - \hat{\theta}_{.j})^2},$$

where N is the total number of examinees; $\theta_{.j}$ and $\hat{\theta_{.j}}$ are the true ability and estimates ability of the j-th examinees, respectively. The point "." can be replaced by any dimension number, such as 1 or 2 in two-dimensional case. Second, RMSEs of the estimated response surfaces, which is

$$RMSE(\hat{P}(\boldsymbol{\theta})) = \sqrt{\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}(P_{ij}(\boldsymbol{\theta}_j) - \hat{P}_{ij}(\boldsymbol{\theta}_j))^2},$$

where n and N are the number of items and examinees, respectively. $P_{ij}(\boldsymbol{\theta}_j)$ and $\hat{P}_{ij}(\boldsymbol{\theta}_j)$ are probabilities that examinee j answers item i correctly, given by the true and estimated item response surface of item i, respectively. The two probabilities are both evaluated using true abilities, and thus only measure how well the item response surfaces are recovered. Unlike the integral over the whole space, this statistic measures the differences between the true and estimated surfaces where the abilities exist. The last statistic is log-likelihood of the estimated model, which is a measure of goodness of fit. The statistics in the results are based on an average of 50 replications, for each setting.

In the first simulation study (Table 3.1, 3.2, and 3.3), the true correlation (0) was set to be known. The purpose is to compare parameter recovery of the CM, NCM and RAVCM under different true models. The data sets are generated using the CM, NCM, or a mixture of them. As expected, when the CM is the true model, the NCM gives the largest RMSEs and the worst fitting statistic among the three models. Although the RAVCM cannot recover the abilities and item response surfaces as well as the CM (It might be caused by the difference between the CM and RAVCM surfaces.), it gives much smaller RMSEs than the NCM. Also, the ability RMSEs of the RAVCM is very similar to those of the CM. When the NCM is the true model, the log-likelihoods given by the three models are close, but the RMSEs of ability estimates and surfaces given by the CM are generally larger than the NCM and RAVCM. When half of items are generated from the CM and the other half are generated using the NCM, the RAVCM gives smaller RMSEs of ability and surface estimates than the CM and the NCM.

Table 3.1 Performance of the CM, NCM and RAVCM when the true model is the CM. ("R" is the abbreviation of "RMSE". "Fitted" means the fitted model.)

Sample Size	Fitted	R(θ_1)	R(θ_2)	SE(θ_1)	SE(θ_2)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees 15 items	CM	0.58	0.59	0.59	0.81	0.04	-2218
	NCM	0.67	0.67	0.69	0.79	0.24	-3637
	RAVCM	0.59	0.59	0.69	0.72	0.08	-2335
500 examinees 30 items	CM	0.48	0.55	0.46	0.75	0.03	-4739
	NCM	0.57	0.65	0.67	0.73	0.16	-5026
	RAVCM	0.50	0.57	0.59	0.68	0.08	-4915
1000 examinees 15 items	CM	0.63	0.65	0.55	0.91	0.03	-4749
	NCM	0.73	0.73	0.80	0.83	0.16	-5313
	RAVCM	0.65	0.66	0.48	0.99	0.07	-4997
1000 examinees 30 items	CM	0.52	0.53	0.40	0.72	0.03	-10107
	NCM	0.61	0.62	0.63	0.64	0.14	-10478
	RAVCM	0.52	0.53	0.55	0.60	0.07	-10663

In the second simulation (Table 3.4, 3.5), the data were generated according to the NCM (2) with correlated abilities (small correlation, 0.4 and moderate correlation, 0.7), and estimated using both the NCM (2) and the RAVCM (6) as if the abilities were uncorrelated. Because of its rotatability, the RAVCM is true under any correlation structures, while the NCM is no longer the true model when the correlation of abilities are misspecified. To check the effect of rotation, the parameter recovery was evaluated after transforming the estimates back onto the simulated scales where

Table 3.2 Performance of the CM, NCM and RAVCM when the true model is the NCM. ("R" is the abbreviation of "RMSE". "Fitted" means the fitted model.)

Sample Size	Fitted	R(θ_1)	R(θ_2)	SE(θ_1)	SE(θ_2)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees	CM	0.79	0.81	0.63	0.91	0.10	-3007
15 items	NCM	0.72	0.74	0.75	0.84	0.07	-3014
	RAVCM	0.71	0.74	0.78	0.80	0.08	-3036
500 examinees	CM	0.65	0.69	0.55	0.83	0.09	-5937
30 items	NCM	0.47	0.56	0.54	0.63	0.06	-5870
	RAVCM	0.47	0.56	0.55	0.61	0.07	-5914
1000 examinees	CM	0.73	0.75	0.85	0.91	0.11	-5441
15 items	NCM	0.63	0.67	0.74	0.83	0.06	-5093
	RAVCM	0.63	0.68	0.69	0.72	0.08	-5163
1000 examinees	CM	0.68	0.70	0.49	0.75	0.12	-11359
30 items	NCM	0.55	0.58	0.58	0.63	0.04	-11045
	RAVCM	0.56	0.59	0.60	0.61	0.07	-11132

the abilities are correlated. Under all conditions, the RAVCM gives smaller RMSEs of abilities than the NCM. The RMSEs of surfaces are similar. It indicates that after transforming the estimates to correlated space, the RAVCM gives surfaces close to the NCM, which means that the true surfaces in uncorrelated space are well recovered. The log-likelihood of fitted RAVCM is smaller than the NCM when the number of items are small (15). With 30 items, the log-likelihoods are similar.

Note that in Table 3.4 and 3.5, the RMSEs of surfaces are compared under the spaces with true ability correlations. The true model under those spaces should be the NCM, and this is the reason why the RMSEs of surfaces given by the two models are similar. For the RAVCM, the estimated abilities and surfaces under uncorrelated space can be rotated to the other spaces with different correlation structure. The RMSE's of $\hat{P}(\boldsymbol{\theta})$ and likelihood for the RAVCM in Table 3.4 and 3.5 do not change. For the RAVCM, if the item response surfaces are well recovered in the uncorrelated space, the rotated surface will also be good in the correlated spaces. In Figure 3.6, the correlation was misspecified as 0 while the true correlation for the NCM was 0.7. The estimated RAVCM is very close to the true model under 0 correlation representation

Table 3.3 Performance of the CM, NCM and RAVCM when the true model is the mixture of the CM and NCM. (Half of items are generated using the CM, and the other half are generated using the NCM). ("R" is the abbreviation of "RMSE". "Fitted" means the fitted model.)

True Model: 7 CM items, 8 NCM items or 15 CM items, 15 NCM items.							
Sample Size	Fitted	R(θ_1)	R(θ_2)	SE(θ_1)	SE(θ_2)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees	CM	0.78	0.79	0.74	0.76	0.09	-2920
15 items	NCM	0.86	0.87	0.90	0.99	0.16	-3506
	RAVCM	0.76	0.78	0.78	0.82	0.11	-2983
500 examinees	CM	0.67	0.70	0.59	0.63	0.08	-5204
30 items	NCM	0.62	0.64	0.64	0.64	0.10	-5263
	RAVCM	0.54	0.58	0.62	0.61	0.06	-5254
1000 examinees	CM	0.65	0.70	0.72	0.78	0.08	-5239
15 items	NCM	0.67	0.75	0.62	0.88	0.14	-5299
	RAVCM	0.62	0.69	0.71	0.75	0.07	-5331
1000 examinees	CM	0.63	0.68	0.38	0.80	0.09	-11037
30 items	NCM	0.51	0.55	0.54	0.62	0.08	-11075
	RAVCM	0.47	0.49	0.50	0.58	0.05	-11077

Table 3.4 Performance comparison between the NCM and RAVCM when abilities are correlated. Parameter recovery of the RAVCM was evaluated after transforming the estimates back onto the simulated scales where the abilities are correlated. The true correlation is 0.4.

Sample Size	Fitted	R(θ_1)	R(θ_2)	SE(θ_1)	SE(θ_2)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees	NCM	0.73	0.75	0.94	0.95	0.12	-3282
15 items	RAVCM	0.67	0.68	0.87	0.94	0.12	-2789
500 examinees	NCM	0.52	0.55	0.59	0.62	0.05	-5933
30 items	RAVCM	0.48	0.50	0.59	0.58	0.06	-5908
1000 examinees	NCM	0.65	0.69	0.90	0.88	0.12	-7403
15 items	RAVCM	0.62	0.63	0.82	0.87	0.12	-6648
1000 examinees	NCM	0.52	0.53	0.60	0.66	0.05	-11811
30 items	RAVCM	0.46	0.49	0.61	0.64	0.06	-11874

and can be rotated to the 0.7 representation. Both representations looks close to the true model. For the NCM, the problem is that it gives similar item response surfaces under different specified correlation spaces (see the last column of Figure 3.6). If the specified correlation is far away from the true correlation, the error of estimated item response surfaces given by the NCM will be large.

Table 3.5 Performance comparison between the NCM and RAVCM when abilities are correlated. Parameter recovery of the RAVCM was evaluated after transforming the estimates back onto the simulated scales where the abilities are correlated. The true correlation is 0.7.

Sample Size	Fitted	R(θ_1)	R(θ_2)	SE(θ_1)	SE(θ_2)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees	NCM	0.65	0.68	0.94	0.95	0.18	-4182
15 items	RAVCM	0.51	0.55	0.69	0.79	0.18	-3699
500 examinees	NCM	0.60	0.61	0.80	0.87	0.11	-6582
30 items	RAVCM	0.49	0.52	0.79	0.83	0.11	-6372
1000 examinees	NCM	0.56	0.60	0.82	0.85	0.08	-6418
15 items	RAVCM	0.50	0.54	0.77	0.85	0.09	-5758
1000 examinees	NCM	0.55	0.57	0.71	0.72	0.06	-12781
30 items	RAVCM	0.45	0.47	0.66	0.73	0.07	-12754

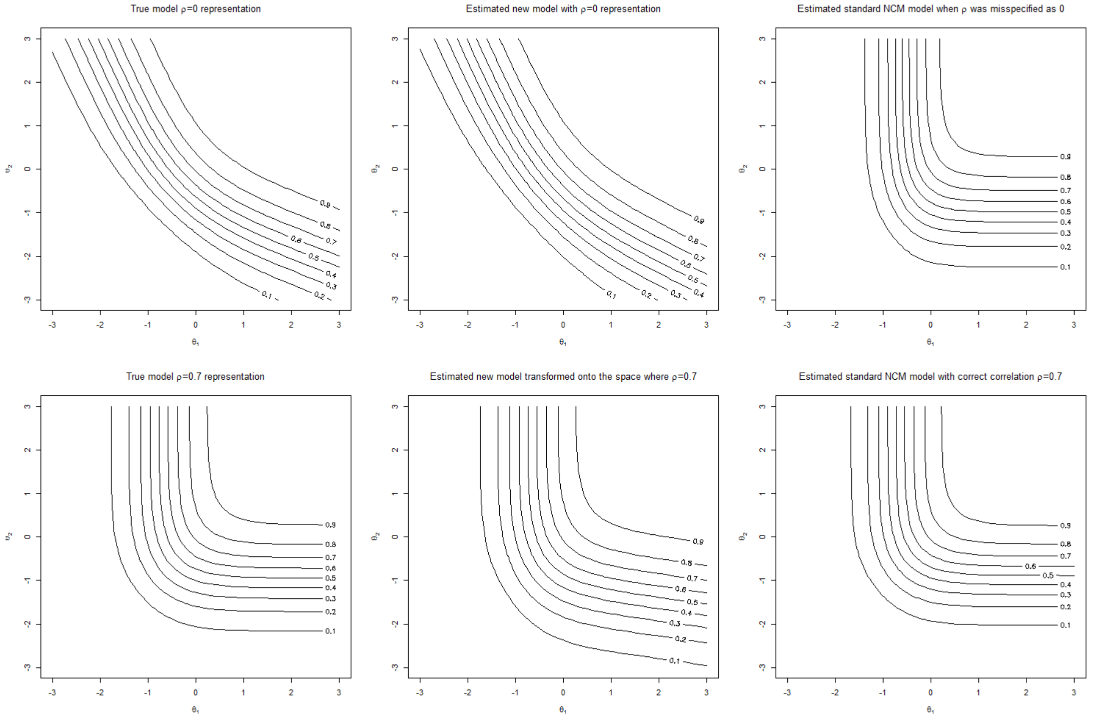

Figure 3.6 Contour plots of the true and estimated RAVCM under different spaces.

Note: Figure 3.6 shows the contour plots of (a) the true model under 0 and 0.7 correlation representation (the fisrt column); (b) the estimated RAVCM under misspecified correlation 0 and the transformed estimated RAVCM under the space where the correlation is 0.7 (the second column); (c) estimated standard NCM when correlation was specified as 0 and 0.7 (the third column). The label of the horizontal and vertical axis are θ_1 and θ_2, respectively. The true correlation of the generated ability parameters was 0.7. All plots are for the same item.

3.4 A Real Data Study: Diagnostic Geometry Assessment

3.4.1 Overall Fitting and Interpretation

The CM, NCM and RAVCM were fit to a real data set from the Diagnostic Geometry Assessment project. This test is used by Measured Progress to identify three common misconceptions that their students hold in geometry (Shape Properties, Transformations, and Geometric Measurement). The geometric measurement part of the test (1995 examinees, 10 items) is analyzed in the current study since it is used to measure two abilities: 1) the process of mentally structuring space; and 2) the connection between mental structuring and measurement formulas. Each item is about measuring the area of a figure using "unit figures" and/or formulas. Most items are multiple choice questions, while item 847 has three true-false sub-items. All items are related to structuring space, while items 851, 855 and 859 do not require knowledge of formula. A hierarchical cluster analysis (Roussos, Stout, & Marden, 1998) of the test also suggests that the test is related to two latent traits (Figure 3.7).

Since the data is real, the true model and parameters are unknown. Also, it is not clear that there should exist a unique true item type (CM, NCM or any other) for all items. For example, in item 879, the figure consists of two rectangles. To measure the area of the figure, one can use formula to calculate the areas of the two rectangles and add them up, or cover the shape with unit square tiles and count the

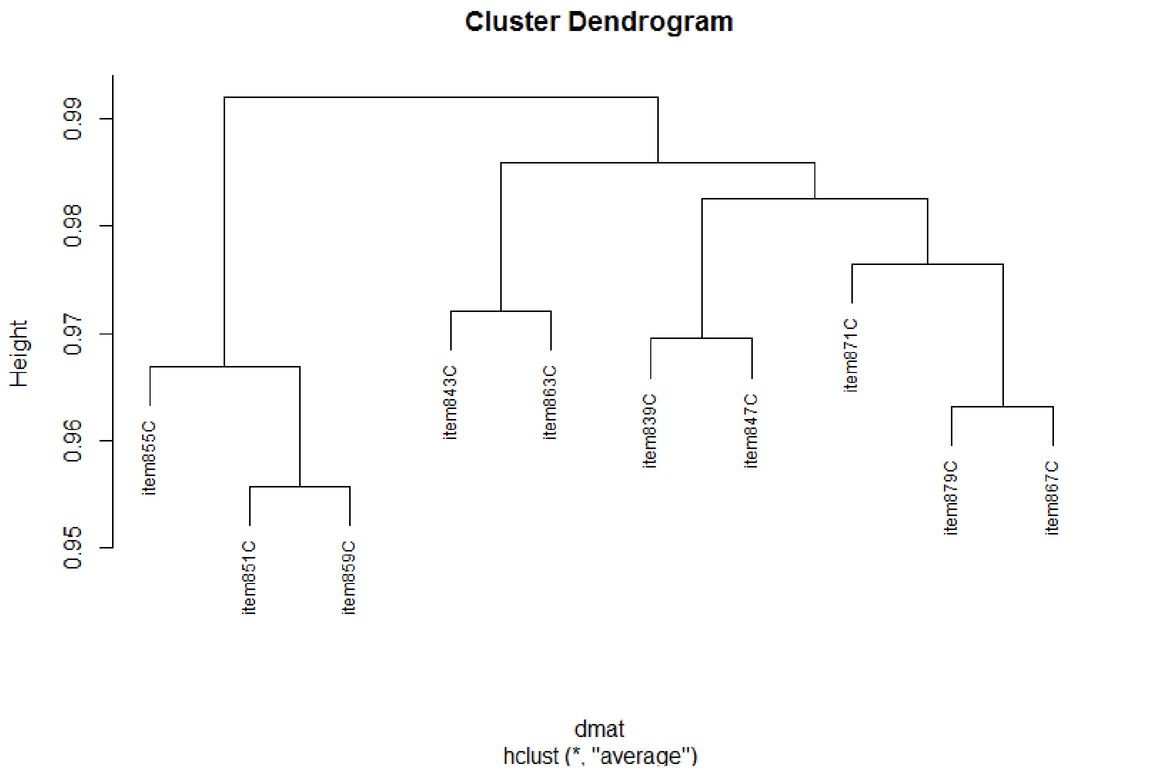

Figure 3.7 Dendrogram obtained from hierarchically clustering the DGAM data with complete linkage and Euclidean distance. Items 851, 855, and 859 are the only ones that lack formulas, but all of the items are related to structuring space.

number of tiles. To answer the question correctly, one only needs to know one of the two methods. Thus, the item seems compensatory. One the other hand, item 847 can be considered as non-compensatory, since one needs both abilities to answer all three sub-items correctly.

The CM, NCM and RAVCM are fit to the data. For the CM, according to some preliminary fitting results and the cluster analysis, item 851 fixed on the θ_1 axis since it is the most extreme. The prior distribution of θ's is set to standard multivariate normal with correlation 0. b parameters has multivariate normal prior with variance 1.5 and mean 0 (for the CM) or -0.5 (for the NCM and RAVCM). a parameters has truncated (assume $a > 0$) normal prior with mean 1 and variance 2.

Table 3.6 reports item estimates given by the three models. From the CM solution, the items 851, 855 and 859 tend to measure the first ability only, which can be interpreted as the ability of mentally structuring space. The other items tends to measure both abilities, i.e., structuring space and connection between structuring and formulas. For the RAVCM, the pattern of estimated b parameters are generally similar to the NCM estimates, while some estimated a parameters are not. Some items, such as item 851, 855, and 871 appears to be non-compensatory. The others appear similar to compensatory, or even asymmetrically compensatory items. For example, item 847 has similar $\hat{a_1}$ and $\hat{a_2}$, small $\hat{a_3}$ and large $\hat{a_4}$. It means that θ_2 can compensate for θ_1 but θ_1 has difficulty compensating for θ_2. For item 859, a discrimination parameter estimated by the RAVCM is very large ($\hat{a_1} = 3.67$). The NCM also gives a large estimate ($\hat{a_1} = 4.97$), while the CM gives small estimates ($\hat{a_1} = 0.67$, $\hat{a_2} = 0.03$). The item response surfaces given by the three models are very different (Figure 3.8). The CM is almost not related to θ_2. The NCM requires examinees to have enough θ_1 and also θ_2 to answer the question correctly. The estimated surface given by the RAVCM is asymmetrically compensatory. How large θ_1 is required depends on the value of θ_2. To have high probability of correct response, large θ_1 is required when θ_2 is very small. For example, when θ_2 is -3, θ_1 needs to be about 2 to have 90% probability of correct response. When θ_2 is large, less θ_1 is required.

The distribution of ability estimates given by the three models are different (Figure 3.9). NCM and RAVCM almost imply two latent classes of examinees. It shows that the NCM and the RAVCM can give different interpretation of the abilities from the CM.

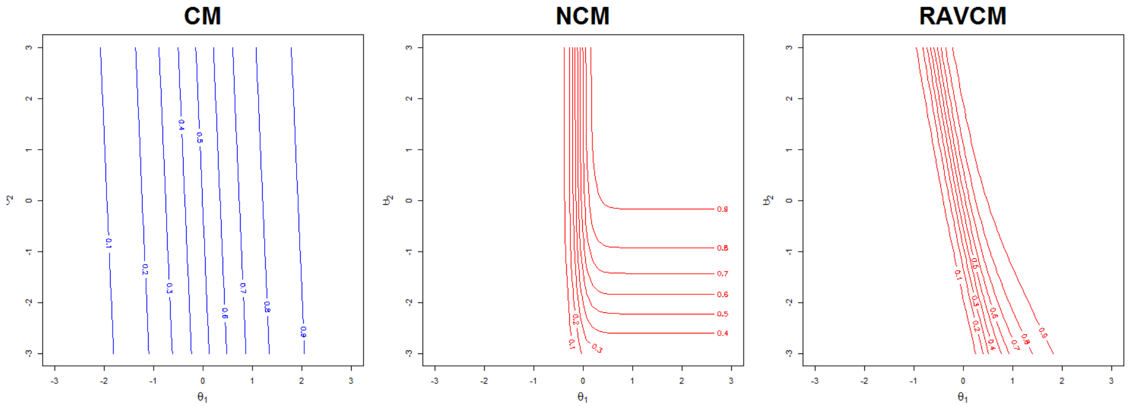

Figure 3.8 The contour plots of the item response surfaces estimated by the CM, NCM and RAVCM for item 859.

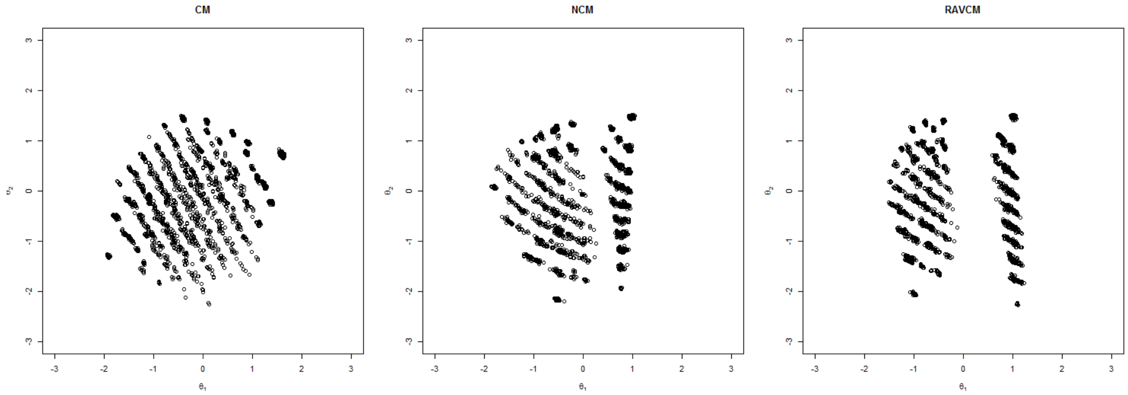

Figure 3.9 Estimated ability distribution of the CM, NCM and RAVCM.

3.4.2 Item Fit and Person Fit

As an item type can be true for a single item but not the whole test, model fit needs to be evaluated at the item level. In this section, the log-likelihood given by each fitted model is evaluated at each iterate of the MCMC chains. The average of these log-likelihoods are compared at item level (Table 3.7). The NCM and RAVCM fit item 859 better than the CM. The other log-likelihoods are similar.

Table 3.7 Item log-likelihoods given by CM, NCM, and RAVCM fitted to the DGAM data.

	851	855	859	839	843	847	863	867	871	879
CM	-910	-1007	-1078	-985	-960	-901	-1013	-917	-1033	-877
NCM	-1055	-1033	-641	-960	-934	-886	-1019	-946	-1030	-906
RAVCM	-1059	-999	-414	-960	-956	-882	-1021	-956	-1038	-941

	Total
CM	-9681
NCM	-9410
RAVCM	-9226

Drasgow, Levine and Williams (1985) proposed a standardized person fit statistic for unidimensional 3PL model, which is $z_3 = \frac{l-E(l)}{(Var(l))^{1/2}}$. l is the log-likelihood for a person's response. It and its moments are calculated using estimated parameters. Asymptotically, z_3 is standard normal, and its square has chi-squared distribution with degree of freedom 1. To use z_3 as an item fit statistic, one can replace person log-likelihood by item log-likelihood. Table 3.8 shows the squared z_3 statistics at the item-level. Generally the RAVCM has the smallest z_3^2. Figure 3.11 gives the quantile-quantile plot of z_3^2 given by the three models. High z_3^2 indices indicate poor fit. For the DGAM data, the CM and RAVCM gives better person fit than the NCM. Note that all points are below the q-q line. It might be caused by the small number of items (Molenaar & Hoijtink, 1990). In this case person fit statistics tend to be conservative.

Table 3.8 Drasgow's z_3^2 for each item given by CM, NCM, and RAVCM fitted to the DGAM data.

	851	855	859	839	843	847	863	867	871	879
CM	11.11	6.86	8.34	5.07	5.08	4.45	5.28	7.35	6.22	6.52
NCM	5.93	6.18	9.88	4.18	2.66	3.85	3.78	5.57	5.67	4.79
RAVCM	4.67	5.76	7.36	4.41	2.46	1.67	4.53	4.99	5.38	1.04

Table 3.6 The CM, NCM and RAVCM item parameter estimates for the DGAM data. Parameters are scaled such that the variance of $\hat{\theta}_1$ and $\hat{\theta}_2$ is 1.

Item	CM			NCM				RAVCM					
	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{b}_1$	$\hat{b}_2$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{b}_1$	$\hat{b}_2$
851	0.99	0.00	0.00	0.89	0.75	-0.50	-0.50	0.87	0.00	0.16	0.63	-0.50	-0.50
855	0.76	0.14	0.11	0.94	0.73	-0.88	-0.46	1.11	0.04	0.03	0.77	-1.18	-0.35
859	0.67	0.03	-0.01	4.97	0.63	-0.59	-1.40	3.67	0.60	1.13	0.56	-0.43	-0.93
839	0.65	0.34	-0.41	1.39	0.81	-2.78	-0.41	0.52	0.44	0.04	0.80	-1.48	-0.71
843	0.71	0.27	-0.56	1.54	0.70	-2.39	-0.75	0.69	0.46	0.10	0.67	-1.27	-1.06
847	0.64	0.38	-0.77	0.78	0.92	-2.27	-0.78	0.35	0.54	0.07	0.75	-1.45	-1.22
863	0.55	0.38	-0.38	1.05	0.71	-2.13	-0.29	0.60	0.31	0.07	0.72	-1.62	-0.54
867	0.53	0.57	-0.55	0.88	0.84	-2.31	-0.42	0.44	0.45	0.05	0.77	-1.62	-0.69
871	0.54	0.65	-0.05	0.68	0.82	-1.48	-0.08	0.60	0.08	0.02	0.83	-1.54	-0.20
879	0.48	0.55	-0.77	0.13	0.90	-1.32	-1.19	0.24	0.58	0.08	0.56	-1.19	-1.31

Sinharay (2006) suggested using a Bayesian item fit diagnostic plot to check model fitting, based on the posterior predictive model-checking (PPMC) method (Rubin, 1984). The method generates replicated responses using parameters drawed at each iterate of the MCMC chain. In an item fit plot, the solid line connects points that indicates observed proportion correct of each raw score group. A box represents the distribution of replicated proportions correct. For any item, too many observed proportions lying far from the center of the replicated values indicate a failure of the model to explain the responses to the item. Figure 3.10 shows inadequate fit of the CM for item 859 since some boxes are far away from the line. For the NCM and RA-VCM, the observed proportions lie closer to the center of the boxes, which indicates better fit. For the other items, the fit plots of the three models are similar.

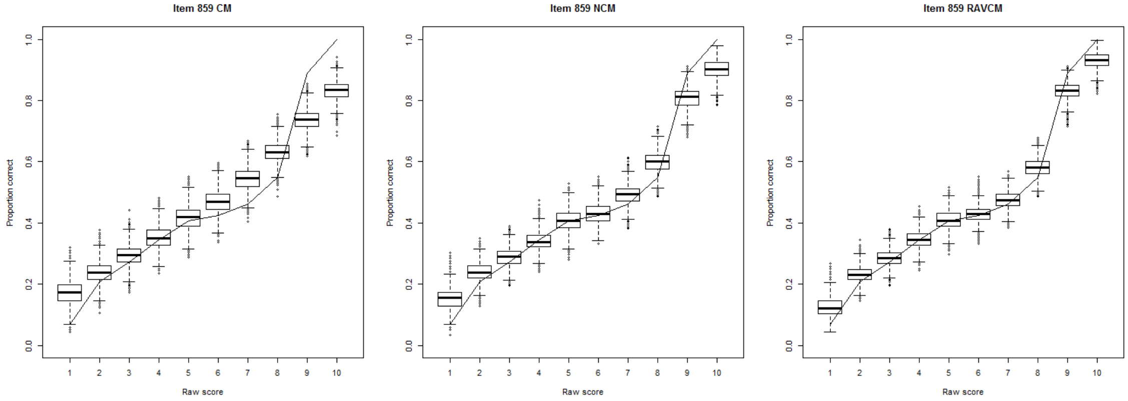

Figure 3.10 Item fit plot for item 859. The observed and replicated proportions correct are plotted for each raw-score group. The left panel (for the CM) shows sign of misfit while the middle and right panels (for the NCM and RAVCM, respectively) show better fitting.

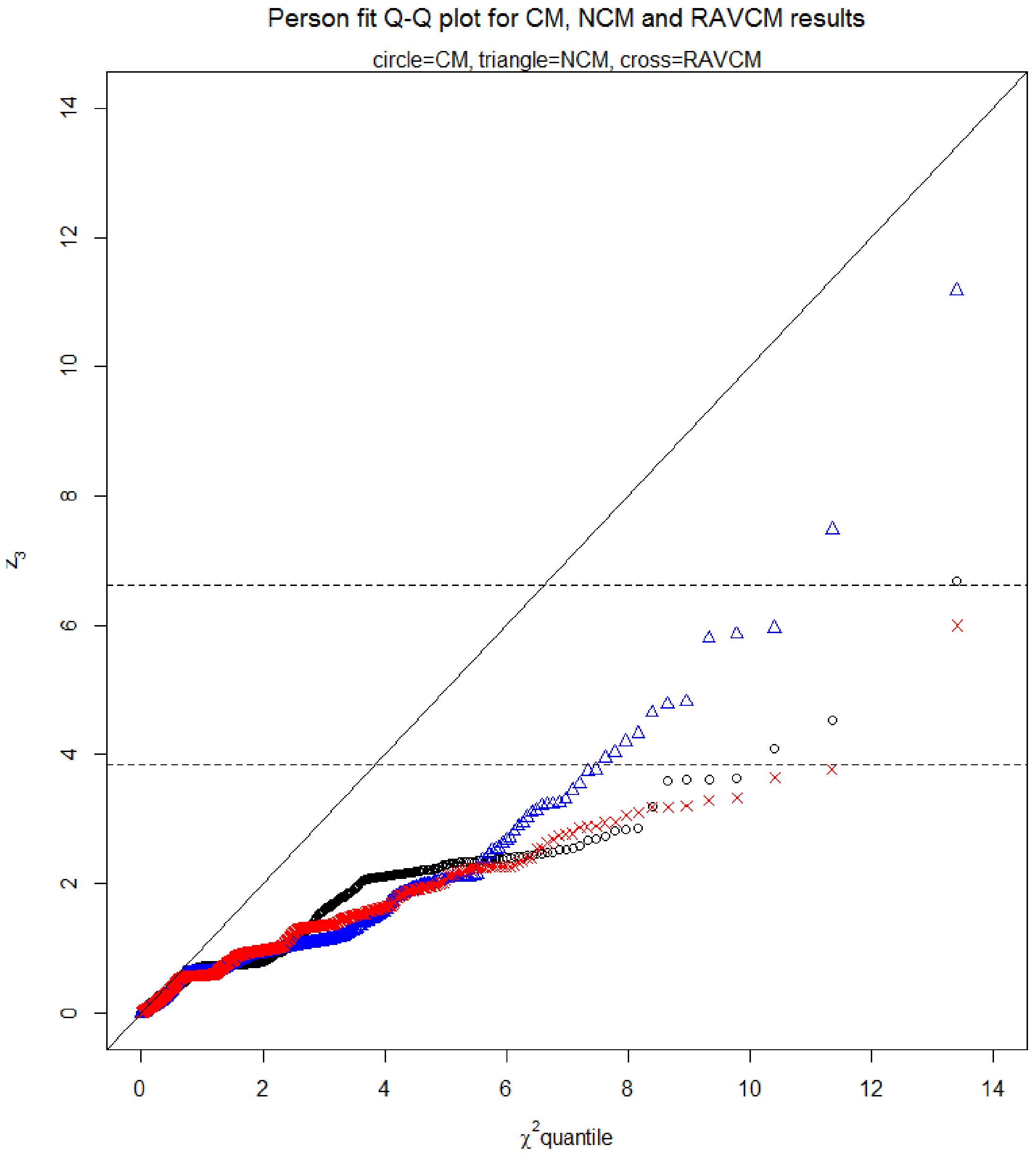

Figure 3.11 Q-Q plot of z_3^2 person fit statistics given by the CM, NCM and RAVCM. The lower and higher horizontal dashed line of reference marks the 5% and 1% cut-off of χ^2 distribution, respectively.

These result shows that the CM and NCM are not adequate to explain the responses of some items and examinees. Therefore, it is very possible that the true model is not simply the CM or NCM. In this case, the RAVCM provides better fit.

3.4.3 An Exploratory Simulation Study

It is doubtful that the item log-likelihood can really distinguish between the compensatory and non-compensatory items, especially when the number of items is small (see the next chapter). A method to explore the true model for each item is simulation. We can simulate data using different true models, and compare the pattern of log-likelihoods given by the estimated models with the log-likelihoods in the last section. When the models used to generate items are or similar to the true models, the likelihood pattern given by the fitted model with simulated data should be similar to the pattern in Table 3.7. The purpose of this exploratory simulation study is to find such models that can generate data similar to the real one. In this case we can claim that these models are close to the true models of the items.

Several data sets are simulated using the estimated item parameters and a standard multivariate normal distributed abilities with correlation 0. (With these generated abilities, the log-likelihoods given by the fitted models can be much larger than those given by the real data. Therefore, we compare the pattern but not the value of the log-likelihoods.) The CM, NCM, and RAVCM are then fit to the simulated data. In the first simulation, responses for most items are generated using the CM. Only the item 859 is generated using the RAVCM as it gives much larger log-likelihood than the CM in Table 3.7. The item likelihoods given by the three fitted models are shown in Table 3.9. The likelihoods are not similar to those in Table 3.7. For the generated data, the fitted CM has the largest likelihood for most the items. The test likelihood of the CM is also much larger than the NCM and the RAVCM. For the real data, the CM does not fit as well as this, which indicates that many items are not compensatory.

Table 3.9 Item log-likelihoods given by CM, NCM, and RAVCM fitted to the DGAM data. Item responses are generated using the RAVCM for item 859. The others are generated using the CM.

	851	855	859	839	843	847	863	867	871	879
CM	-612	-876	-361	-808	-816	-830	-1033	-922	-1019	-832
NCM	-1015	-830	-242	-828	-855	-844	-1021	-956	-1048	-853
RAVCM	-596	-901	-405	-825	-842	-852	-1039	-931	-1018	-841

	Total
CM	-8109
NCM	-8492
RAVCM	-8250

The following simulation shows that to obtain log-likelihood similar to those in the real data it is necessary to simulate more items using the NCM or RAVCM. Table 3.10 shows the log-likelihoods with a pattern more similar to the log-likelihoods for the real data. In this generated data set, only responses for 4 items are generated using the CM. 4 and 2 items are generated using the NCM and the RAVCM, respectively. To some extent, this study shows that some items are indeed not compensatory, and are better fit by a model that accounts for that.

Table 3.10 Item log-likelihood given by CM, NCM, and RAVCM fitted to another simulated data. The responses of item 839 and 859 are simulated using the RAVCM. The responses of item 871, 855, 863, 867 are simulated using the NCM. The responses of the other items are simulated using the CM.

	851	855	859	839	843	847	863	867	871	879
CM	-731	-1022	-612	-811	-820	-847	-839	-1004	-834	-855
NCM	-1045	-994	-362	-825	-858	-860	-782	-789	-809	-882
RAVCM	-1089	-997	-252	-816	-826	-894	-868	-732	-838	-866

	Total
CM	-8380
NCM	-8210
RAVCM	-8183

3.5 Discussion: Definition of Compensation and Model Interpretation

In literatures, the MIRT models are classified as compensatory and noncompensatory (e.g., Ackerman, 1989; Bolt & Lall, 2003; Way, Ansley, & Forsyth, 1988). Traditionally, compensation is defined as “High ability on one dimension can compensate for the low ability on the second dimension” (e.g., Ackerman, Gierl, & Walker, 2003). The compensation level is described by a parameter λ in the variable compensation model GMIRT (Ackerman & Bolt, 1995; Simpson, 2005). This work put the compensatory model and the non-compensatory model into the same framework, and proposes a new possible type of compensation: in the middle of compensatory and non-compensatory.

An issue with the traditional definition and measure of compensation is the definition of “high ability” and “low ability”. These two terms show that the compensation depends on the value of “compensated ability” and “compensator ability”. For example, for a non-compensatory item (Figure 3.12), if θ_1 is very small (e.g., -3), then the probability of correct response will not increase as θ_2 increase, for almost any θ_2 value. However, if θ_1 is “a little small” (e.g., 0), the probability can increase from 0 to about 0.5 as θ_2 increases from a small value. In the second case, θ_2 does compensate for θ_1 when θ_2 is small (e.g., -3), while θ_2 cannot compensate for θ_1 anymore when it is large enough (e.g., 2). Therefore, compensation does exist in the non-compensatory model. In fact, some researchers tend to use the term “partially compensatory” instead of “non-compensatory” as it is more accurate (e.g., DeMars, 2016).

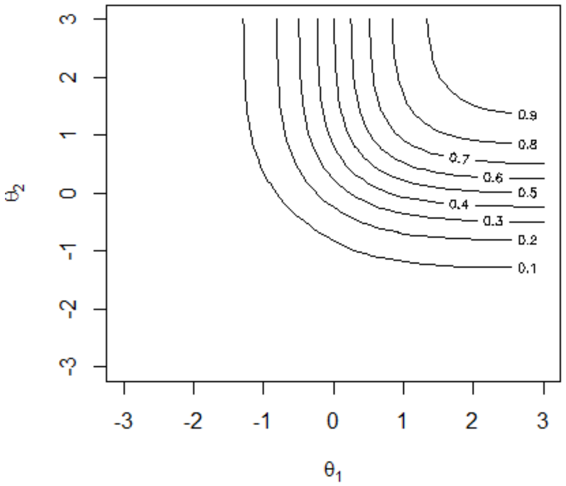

Figure 3.12 A non-compensatory example to explain the definition of compensation.

If the compensation depends on the values of abilities, then it might be difficult to give all information about compensation using a single parameter, for example, the compensation parameter λ in the VCM. For example, even if two items are both compensatory (with $\lambda = 0$), their compensations can be different, as stated in (Ackerman, Gierl, & Walker, 2003): “compensation is greatest when an item has equal discrimination parameters.” Figure 3.13 shows an example. The two items are both compensatory with linear contours. The first item has discrimination parameters $a_1 = a_2 = 1$; the second item has $a_1 = 1$, $a_2 = 0.2$. For the first item, θ_1 and θ_2 can compensate for each other, while for the second item, if θ_1 is low, the the probability of correct response will not increase a lot as θ_2 increases. From this examples, it can be concluded that the compensation is directed. Compensation levels can be different for different compensation directions. It is necessary to specify which ability is compensating (compensator) when talking about compensation.

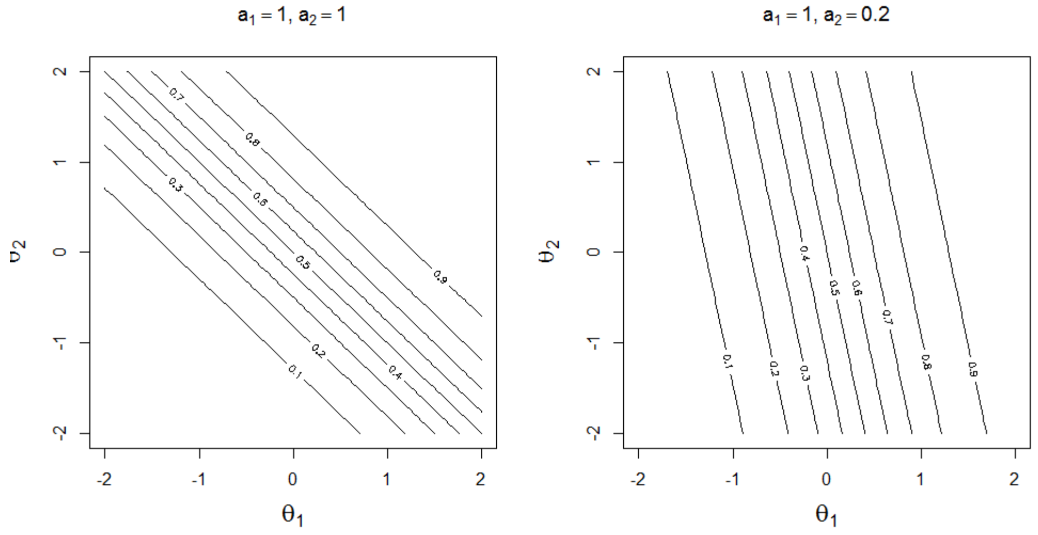

Figure 3.13 Two compensatory items to explain the definition of compensation.

Since compensation depends on the compensator and the values of abilities, it can be redefined by a question: "With fixed abilities on compensated dimensions, how fast does the probability of correct response increase as the compensator increases?" "How fast" can be mathematically described using derivatives, for example, $\frac{\partial P(\theta_1,\theta_2,\theta_3,\ldots)}{\partial \theta_1}$ can be considered as the function of compensation level with compensator θ_1. If the other abilities are fixed, it will be a function of θ_1. This definition is formal, however, and hard to interpret. One can never list the derivatives of the item response function for all possible compensators and values of abilities.

In practice, if the shape of the contour plot is known, then these derivatives are known. As such, compensation of an item can be considered as equivalent to the shape of the item's contour plot. If the contour plot is known, then given any ability values, the probability of correct response can be calculated. This simply means that we have full information about how latent abilities combine, which is the nature of compensation. When researchers choose to use a MIRT model with a certain compensation structure such as the CM or the NCM, the decision is based on their assumptions on how the abilities combine, or cognitive strategies (Ackerman & Bolt, 1995).

In reality, the shape of the contour plot almost surely differs at least somewhat from the linear contour plot for the CM and the curved contour plot for the NCM. For example, the contour could be asymmetric as shown in Figure 3.4. If a model is too restrictive, then it cannot fit every item well, which leads to inaccurate or invalid ability estimates. It is important to have a MIRT model flexible enough to approximate more types of contour plot. This makes the RAVCM potentially useful in that it can approximate more types of contour plot than the common existing MIRT models. As seen in some cases, it gives better ability estimates than the existing MIRT models.

The RAVCM also solves the problem that the compensation of an item can change via rotation. For the same item, the contour plots under different correlation structures can be highly different, as shown in Figure 3.6. If the model is rotatable, then the compensation of items can be compared even if they are under different space. For example, if item 1 and item 2 are non-compensatory when the correlations of abilities are 0.6 and 0.3 respectively. Under the framework of the NCM, their compensations are not comparable, while under the framework of the RAVCM, one can transform the items to the uncorrelated space to compare their compensation.

There exists a trade-off between model flexibility and interpretability, and the flexible MIRT models are more difficult to interpret due to their extra complication. The two-dimensional RAVCM can be interpreted as a two-step model. Although it is interpretable, the interpretation is not intuitive as the unidimensional models or the CM. (For the unidimensional model, the a parameter can be interpreted as a slope parameter in the logistic model. For the CM, the a parameters can be interpreted as factor loadings. Both are easy to understand.) In high dimensional cases, it is more difficult to interpret the parameters in the RAVCM. The advantage of the RAVCM is that it can give more accurate ability estimates when the CM or the NCM are not the true model for all items, or the compensation of some items are very complex. Therefore, we can conclude that when more accuracy is wanted, the RAVCM is favorable; when easy interpretation is needed, the CM works better. In many cases, interpretability might be more important than model accuracy. However, it does not mean that people should always use the CM. The CM can give ability estimates anyway, but these are not the abilities what the test maker wants to measure if the abilities do not combine in a compensatory way for the designed items. This type of misfit is difficult to detect using goodness of fit tests or statistics, which is shown in the next chapter. When the compensation (contour) of the items are not very clear, exploratory studies are needed to find them, and thus the RAVCM can be helpful.

Chapter 4

Evaluating Competing MIRT Models with Different Goodness of Fit Statistics

Model and item misfit often lead to biased estimates and inappropriate interpretations for parameters in item response theory. Traditionally, the reason of item misfit are considered coming from the violation of one of the three assumptions of IRT, or model misspecification (Levine & Rubin, 1979). Model mis-specification can include violation of item response functions (Orlando & Thissen, 2003) and aberrant response patterns (Drasgow, Levine, & McLaughlin, 1991). In the past decades, many goodness of fit test and statistics are developed to evaluate the IRT model fitting, such as Q_1 (Yen, 1981), Z_3 (Drasgow, Levine, & Williams, 1985), and $S - \chi^2$ (Orlando & Thissen, 2000). These statistics are designed for unidimensional IRT models. There are relatively less research for the goodness of fit test and statistics for MIRT models. One example is Zhang & Stone (2008), which examine the utility of the $S - \chi^2$ likelihood based statistic in evaluating item fit for the compensatory model. The item misfit in their research are caused by the violation of the monotonicity assumption and ignored guessing effect, which have been studied for unidimensional IRT models. However, in addition to these common threats to goodness of fit in unidimensional IRT, mis-specifying the compensation in MIRT can cause model or item misfit. As this type of item misfit is not studied in Zhang & Stone (2008), their conclusion that the $S - \chi^2$ statistic is capable of evaluating item fit in the CM might not be always suitable. Further investigation of the performance of this statistic is needed.

In practice, the compensatory model is widely applied. However, as showed in table 3.2, when the compensatory model is fit, but the true model is the non-compensatory model, the RMSEs of estimated parameters can be large. Also, using the compensatory model for non-compensatory items is considered as a cause of differential item functioning (DIF; Liaw, 2015). In this case, the CM does not measure the true latent traits that generate the data, which is what the test maker want to measure. In real data analysis, RMSEs is unavailable. It would be good if there was a powerful goodness of fit test that could reject the CM with large probability. So, when different MIRT models are fit, statistics for model selection are needed.

The purpose of this study is to evaluate the effectiveness of goodness of fit statistics at distinguishing between the compensatory model and the non-compensatory model at the test-level and the item-level via simulation. Since the RAVCM can approximate both the CM and the NCM, it is also important to check if it can outperform the statistics in distinguishing the compensatory and non-compensatory items.

4.1 Statistics in this Study

The statistics evaluated in this study are the M_2 (Maydeu-Olivares & Joe, 2005), $S - \chi^2$ (Orlando & Thissen, 2000), Z_3 (Drasgow, Levine, & Williams, 1985), log likelihood, AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), and the angle between item vectors in the Rotatable Asymmetric Variable Compensation Model (RAVCM). They are listed in table 4.1.

Table 4.1 Statistics studied in Chapter 4

Statistic	Type
M_2	Overall fit test statistic
Z_3	Item fit test statistic
$S - \chi^2$	Item fit test statistic
Log likelihood	Information criterion
AIC	Information criterion
BIC	Information criterion
DIC	Information criterion
Angle between item vectors	A measure of compensation

Maydeu-Olivares & Joe's M_2

The M_r family of overall goodness of fit statistics is proposed by Maydeu-Olivares & Joe (2005). Its power and error rate for IRT models are discussed in Maydeu-Olivares & Joe (2006). These statistics, which are the classes of quadratic form statistics based on the residuals of margins or multivariate moments up to order r, are designed for testing the goodness of fit of large 2^n contingency tables. In the goodness of fit test, the mull and alternative hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ for some $\boldsymbol{\theta}$, versus $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ are considered, where $\boldsymbol{\pi}(\boldsymbol{\theta})$ is a parameteric multivariate model with parameter vector $\boldsymbol{\theta}$. The M_r statistic is given by

$$M_r = M_r((\theta)) = N(\boldsymbol{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}}))'\hat{C}_r(\boldsymbol{p}_r - \boldsymbol{\pi}_r(\hat{\boldsymbol{\theta}})),$$

where r is the maximum order of residual of marginal tables used by the statistic (e.g., M_2 is the quadratic form in univariate and bivariate residuals.); $\boldsymbol{p}$ is a vector of cell proportions; C_r is a function of the partial derivative of π. In Maydeu-Olivares & Joe's studies, among the M_r family, only the empirical Type I errors of M_2 remained accurate throughout the different sparseness conditions considered in their study. Therefore, for most IRT applications M_2 is the statistic of choice.

M_2 is a test for multiple contingency tables, which make it a good tool to examining the local independence assumption of IRT. Few research examines whether mis-specifying compensation leads to extreme M_2, but it is possible that the items are noticeably not independent conditioning on θ's given by a wrong model.

Drasgow, Levine, & Williams' Z_3

Drasgow, Levine, & Williams proposed a standardized goodness of fit index z_3 to measure whether each individual examinee's response pattern fits the three parameter logistic IRT model (1985). This person fit statistic is used in the real data analysis in the last chapter. The statistic can be expressed as

$$z_3 = \frac{l_0 - E_3(\hat{\theta}_d)}{\sigma_3(\hat{\theta}_d)}.$$

In this formula,

$$l_0 = \sum_{i=1}^{n} u_i log\{P_i(\hat{\theta}_d)\} + (1 - u_i) log\{Q_i(\hat{\theta}_d)\}$$

is a linear function of the item scores, where u_i is the dichotomous item score of item i, which can be 1 or 0; $\hat{\theta}_d$ is the maximum likelihood estimates of abilities; n is the total number of items.

The E_3 and σ_3 is the mean and standard deviation of the random variable version of l_0

$$X_3(t) = \sum_{i=1}^{n} U_i log\{P_i(t)\} + (1 - U_i) log\{Q_i(t)\},$$

conditioning on $\theta = t$.

This statistic is a standardized summation of the model likelihood function for each item, given a person's ability estimates. It can be transfered to an item fit statistic by taking the summation of the likelihood function for each person's response to an single item, i.e., take

$$l_0 = \sum_{j=1}^{N} u_j log\{P_i(\hat{\theta}_j)\} + (1 - u_j) log\{Q_i(\hat{\theta_j})\},$$

where the j is an index of examinee; P_i is the item response function of the item i. Also, the 3PL item response function P can be replaced by any item response function for MIRT. Therefore, it is possible to use this statistic for evaluating item fit for MIRT models.

Orlando & Thissen's $S - \chi^2$

The $S - \chi^2$ statistic is proposed by Orlando & Thissen (2000) for examining item fit for IRT models. In earlier studies, the goodness of fit statistics usually depended on the estimated ability parameters, such as in z_3 and Yen's Q_1 (1981). This might lead to unclear degree of freedom of the statistic. A second problem with the previous indices is that they grouped examinees into equal-size groups and check goodness of fit for each group. This grouping method is highly sample dependent. The number of groups and the cut off points can affect the statistics. Unlike the previous indices, $S - \chi^2$ depends on observed and expected proportions of correct responses only, and it divides examinee into groups according to their number of correct responses, and thus can avoid the two problems. It has the form

$$S - \chi_i^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})},$$

where N_k is the number of examinees in the number-correct score group k; O_{ik} and E_{ik} are the observed and expected proportions for item i and number-correct score group k, respectively. The O_{ik} are computed from data, and the E_{ik} are computed using a recursive algorithm (Thissen, Pommerich, Billeaud, & Williams, 1995).

It is demonstrated in their study that the $S-\chi^2$ statistic is effective for evaluating item fit for unidimensional IRT models. Its sampling distribution can approximate chi-square distribution closely and the test is powerful enough to find non-fitting items. Zhang & Stone (2008) evaluated its effectiveness for MIRT models and concluded that it is viable for MIRT. As discussed at the beginning of this chapter, they only investigate two types of non-fitting items (not including misspecification of compensation), and they consider that the other types of mis-fitting is also worth studying.

Family of information criteria

The likelihood function of a statistical model with estimated parameter, denoted by $\hat{L}$, can be used to evaluate goodness of fit of the model. To reduce computational complexity, we usually take the logarithm of it, denoted by $log(\hat{L})$. When comparing models, the model with higher log likelihood fits the data better. This does not account for model complexity however.

The Akaike information criterion (AIC; Akaike, 1974) considers the trade-off between the goodness of fit of and simplicity of a model. The AIC value of a model is

$$AIC = 2k - 2log(\hat{L}),$$

where k is the number of estimated parameters in the model. The AIC assesses goodness of fit by log likelihood, and includes a penalty term $2k$ for complex models. When comparing models, the model with smallest AIC value is preferred. AIC is based on information theory. In theory, there is no "correct" model. When using a model to recover the process that generates the data, there will always be loss of information. AIC is a measure of the relative information lost by a given model.

The Bayesian information criterion (BIC; Schwarz, 1978) is an information criterion for model selection, which is similar to AIC but with a larger penalty term. The BIC value of a model is

$$BIC = log(n)k - 2log(\hat{L}),$$

where n is the sample size of the data. The other parameters are the same as in AIC. Similar to the AIC, the model with the smallest BIC value is preferred in model comparison. Different from the AIC, BIC tries to find the true model. Under a certain Bayesian setup, BIC is considered as a measure of the posterior probability of a given model being true. Besides the theoretical difference between AIC and BIC, the difference in practice is that the AIC might choose a model with more parameters.

The deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) is a generalization of the AIC, which is useful when the Bayesian posterior distribution of the model is estimated using a Markov Chain Monte Carlo (MCMC) method. Define the deviance as $D = -2log(\hat{L}) + C$, where C is a constant that cancels out in later calculation. The deviance information criterion is expressed as

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

where $p_D = \bar{D} - D(\bar{\theta})$. $\bar{\theta}$ is the expectation of the parameters. $\bar{D}$ is the posterior mean deviance, which is easy to calculate from the samples generated in MCMC estimation. Usually, $\bar{D}$ is smaller for model with more parameters. p_D act as a penalty term. DIC is valid when the posterior distribution is normal.

Angle between item vectors in the RAVCM

In section 3.1, the RAVCM and its properties are introduced. The two-dimensional RAVCM is written as

$$
\begin{aligned}
&P(X_{ij} = 1|a_{1i}, a_{2i}, a_{3i}, a_{4i}, b_{1i}, b_{2i}, \boldsymbol{\theta_j}) \\
=&\{\frac{1}{1 + exp[-1.7(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} - b_{1i})]}\}\{\frac{1}{1 + exp[-1.7(a_{3i}\theta_{1j} + a_{4i}\theta_{2j} - b_{2i})]}\}.
\end{aligned}
$$

It has two item vectors, $(a_{1i},\ a_{2i})$ and $(a_{3i},\ a_{4i})$, where i is the index of items. When $a_{2i} = a_{3i} = 0$, the angle between the two item vectors is 90 degree ($\pi/2$), and the IRF is the same as the non-compensatory model. When $a_1/a_2 = a_3/a_4$, the angle between the two item vectors is 0 degree (0), and the model is compensatory as it has a linear contour plot, which is similar to the compensatory model. Therefore, the angle between $(a_{1i},\ a_{2i})$ and $(a_{3i},\ a_{4i})$ can be considered as a measure of the compensation level of an item. To compare the RAVCM to the other statistics, in this study we only examine if RAVCM can distinguish between the CM and the NCM items. As an ad-hoc rule, the RAVCM identifies the true model as compensatory if the angle $\leqslant \pi/4$, non-compensatory if the angle $> \pi/4$, where $\pi/4$ is 45 degree, in the uncorrelated space.

Figure 4.1 shows an example. The RAVCM identifies the two items as compensatory and non-compensatory, respectively, according to the angle between the two item vectors for each item. For the first item, the angle is smaller than 45 degree; for the second item, the angle is greater than 45 degree.

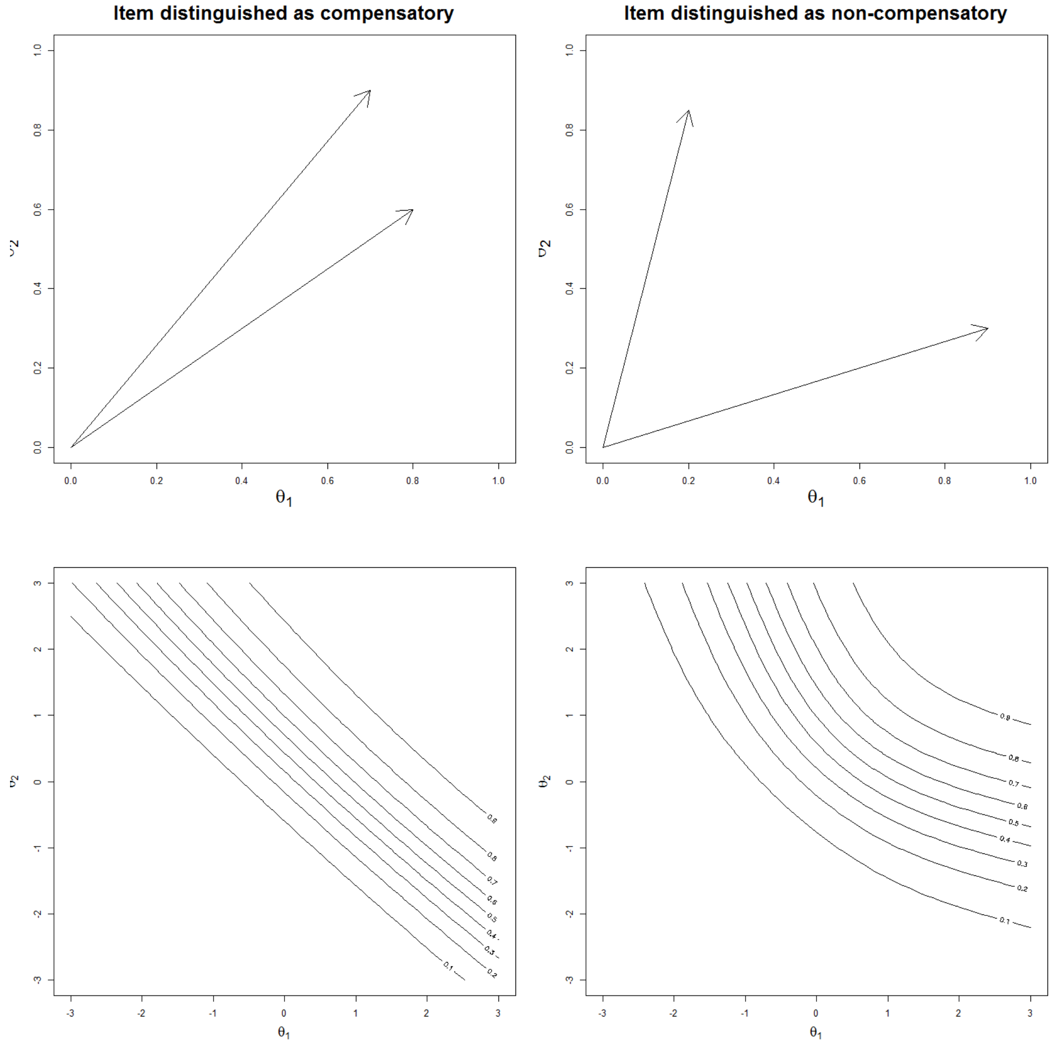

Figure 4.1 The item vectors and contour plots of two items distinguished as compensatory and non-compensatory, respectively, by the RAVCM.

4.2 Simulation Study

In this simulation study, the goodness of fit statistics are evaluated at the test-level and the item-level. (M_2 is only evaluated at the test-level as it is an overall test for contingency tables.)

For the test statistics M_2, Z_3, and $S - \chi^2$ the simulated data are generated using the CM and NCM. while the fitted model is always the CM. The purpose is to examine how often the statistics reject the CM (for M^2) or proportion of the rejected items (for Z_3 and $S - \chi^2$) when the CM is true or not true. The reason why only the CM is fit is that we assume the CM is more often misused compared to the NCM, since the CM is easier to estimate and interpret. When the compensation of the test items are not very obvious, one can still fit the CM and obtain estimated ability parameters. In this case, it is necessary to check if the CM is appropriate. On the other hand, the NCM is usually used when the design of the test items is obviously non-compensatory or multi-step. In this case people usually do not need to examine if the test is really non-compensatory.

For the other statistics, the simulated data are also generated using the CM and NCM. The log likelihood, AIC, BIC, and DIC for the fitted CM and NCM, and the angle of item vectors in RAVCM are calculated to identify the true model at the test-level and item-level. The proportion of correctly selected test and item are calculated.

There are four sample sizes considered in this study: 500 examinees, 15 items; 500 examinees, 30 items; 1000 examinees, 15 items; 1000 examinees, 30 items. For each simulated examinee and item, the ability parameters (θ) are generated from standard normal distribution. It means that θ_1 and θ_2 are uncorrelated. The difficulty parameter b are generated from the standard normal distribution for the CM, and normal distribution with variance 1 and mean -1 for the NCM. The discrimination parameters (a) are generated from truncated normal distribution $(0, 0.5)$ that only allows for positive value. The metropolis hasting within Gibbs sampler algorithm is used for model estimation. All settings and details are the same as in section 3.2.2.

4.2.1 Effectiveness of M_2

For each sample size, response data are generated using the CM and the NCM. The CM is fit. The M_2 statistic is calculated for the fitted model. When the p-value of the M_2 test is smaller than the threshold (0.05 or 0.10), the CM is rejected. The procedure is repeated 100 times.

Table 4.2 shows the rejection rates for different sample size and threshold. Generally, the rejection rate is higher when the CM is not the true model, compared to the case that the CM is the true model. However, the rejection rates are not very satisfying. The best case is 1000 examinees, 30 items with $\alpha = 0.10$. The rejection rate in this case is only 0.42. For the other sample size, the rejection rates are smaller. It is possible that the M_2 can work better for the sample sizes larger than those in this study, while for the sample sizes in this study, it seems not effective in rejecting the false compensatory tests.

4.2.2 Effectiveness of $S - \chi^2$ and z_3

In the first simulation, for each sample size, response data are generated using the CM and the NCM. The CM is fit. The $S - \chi^2$ and z_3 statistic are calculated for each item in the fitted model. When the p-value of the $S - \chi^2$ or z_3 test is smaller than 0.05, the fitted model CM is rejected at the item-level. The procedure is

Table 4.2 Rejection rate of M_2 under different true models in 100 repetition. The CM is always the fitted model.

Sample Size	True Model	Rejection Rate	Rejection Rate
		($\alpha = 0.05$)	($\alpha = 0.10$)
500 examinees	CM	.07	.09
15 items	NCM	.07	.21
500 examinees	CM	.07	.11
30 items	NCM	.18	.24
1000 examinees	CM	.05	.07
15 items	NCM	.13	.18
1000 examinees	CM	.09	.13
30 items	NCM	.30	.42

repeated 100 times. Table 4.3 shows the proportions of rejected items for different sample size. The $S - \chi^2$ statistic has low rejection rate for all sample sizes, while the z_3 has relatively large rejection rates. Especially, when the number of item is small (15), the z_3 has large rejection rate. These two statistics cannot distinguish between the CM and NCM as the rejection rates are similar for the two true models. For example, for 1000 examinees and 15 items, the rejection rates of $S - \chi^2$ are 0.09 and 0.11, when the CM is and is not the true model, respectively; the rejection rates of z_3 are 0.71 and 0.77, respectively. In this case, the $S - \chi^2$ tends to accept the CM while the z_3 tends to reject the CM, regardless what the true model is.

In the second simulation, for each sample size, half of the items are generated using the CM, and the other half are generated using the NCM. The CM is fit. The purpose is to check that whether the $S - \chi^2$ and z_3 can detect non-compensatory items, when the compensatory and non-compensatory items are mixed in a single test. When the p-value of the the $S - \chi^2$ or z_3 test is smaller than 0.05, the fitted model CM is rejected at the item-level. Table 4.4 shows the average error rates of $S - \chi^2$ and Z_3 in detecting non-compensatory items, showing the "incorrect rejection" rate (Type I

Table 4.3 Average proportion of items rejected by $S-\chi^2$ and Z_3 under different true models in 100 repetition. The CM is always the fitted model.

Sample Size	True Model	Proportion of items rejected by $S-\chi^2$ ($\alpha = 0.05$)	Proportion of items rejected by Z_3 ($\alpha = 0.05$)
500 examinees	CM	.06	.44
15 items	NCM	.08	.53
500 examinees	CM	.07	.15
30 items	NCM	.07	.23
1000 examinees	CM	.09	.71
15 items	NCM	.11	.77
1000 examinees	CM	.07	.34
30 items	NCM	.08	.46

error) and "incorrect acceptance" rate (Type II error). For both of the $S-\chi^2$ and Z_3, the incorrect acceptance rates are very large for all sample sizes. Therefore, the two statistics do not have enough statistical power to reject the non-compensatory items when the CM is the fitted model.

Table 4.4 Average error rate of $S-\chi^2$ and Z_3 when the true model is the CM for half of the items and is the NCM for the other half. The CM is always the fitted model. ($\alpha = 0.05$)

Sample Size	Incorrect rejection rate of $S-\chi^2$	Incorrect rejection rate of Z_3	Incorrect acceptance rate of $S-\chi^2$	Incorrect acceptance rate of Z_3
500 examinees 15 items	.07	.48	.91	.75
500 examinees 30 items	.07	.29	.95	.81
1000 examinees 15 items	.09	.43	.92	.87
1000 examinees 30 items	.08	.46	.94	.73

For the failure at detecting the non-compensatory tests and items, a hypothesized reason is the good-fit of the CM. As mentioned in the last chapter, when the true model is not the CM for some items, the likelihood of fitted CM can still be relatively large. This means that the CM can sometimes introduce "appropriate" latent abilities to make the overall model fit well, while the abilities might not be the ones that the test maker want to measure, if the items are not compensatory. In this case, as the overall fit is well, the test statistics $S-\chi^2$ and z_3 are not likely to be extreme.

4.2.3 Effectiveness of the other statistics

In the first simulation, the log-likelihood, AIC, BIC, DIC, and the angle between item vectors in the RAVCM are evaluated at the test-level. For each sample size, response data are generated using the CM and the NCM. The test-level log likelihood, AIC, BIC, and DIC are calculated for the fitted CM and NCM. The model with larger log likelihood, smaller AIC, BIC, or DIC are distinguished as the true model. The average angle between item vectors are given by the fitted RAVCM. If the average angle is smaller than or equal to 45 degree, the RAVCM will identify the test as compensatory, otherwise as non-compensatory.

The rates of successfully distinguished true model are shown in Table 4.5. When the true model is the CM, the log likelihood, AIC, BIC, and DIC are very likely to distinguish it: For all sample sizes, the rates of distinguished true model are at least 89%. The BIC and DIC can always distinguish the CM if it is the true model. The RAVCM can distinguish the CM as the true model in around 70% to 80% cases. When the true model is the NCM, the log likelihood and AIC performs well for tests with 30 items, but not very well for tests with 15 items. The BIC and DIC have very small rates of distinguishing the true model (NCM) for all sample sizes. The

RAVCM based method has the largest rates for all sample sizes for the NCM.

Table 4.5 Rate of true model distinguished by overall log likelihood, AIC, BIC, DIC, and the RAVCM in 100 repetition. The true model is selected using the overall information criterions given by the fitted CM and NCM, and the average angle between item vectors given by the RAVCM. The RAVCM identifies the true model as compensatory if the average angle $\leqslant \pi/4$, noncompensatory if the average angle $> \pi/4$.

Sample Size	True Model	log likelihood	AIC	BIC	DIC	RAVCM
500 examinees	CM	.94	.95	1	1	.75
15 items	NCM	.63	.62	.40	.18	.78
500 examinees	CM	.89	.91	1	1	.72
30 items	NCM	.94	.93	.20	.24	.95
1000 examinees	CM	.90	.91	1	1	.79
15 items	NCM	.60	.59	.23	.10	.91
1000 examinees	CM	.96	.98	1	1	.82
30 items	NCM	1	.99	.46	.09	1

In the second simulation, the log-likelihood, AIC, BIC, DIC, and the angle between item vectors in the RAVCM are evaluated at the item-level. For each sample size, half of the items are generated using the CM, and the other half are generated using the NCM. The item-level log likelihood, AIC, BIC, and DIC are calculated for the fitted CM and NCM. The model with larger item log likelihood, smaller AIC, BIC, or DIC are distinguished as the true model for that item. Each angle between item vectors is given by the fitted RAVCM. If the angle is smaller than or equal to 45 degree, the RAVCM will identify the item as compensatory, otherwise as noncompensatory.

For each repetition, the rates of items whose true model is successfully distinguished can be calculated for different statistics. Table 4.6 shows the average proportion in 100 repetition. The log likelihood, AIC and BIC do not work very well when the number of items is small (15), but work better when the number of items is large (30). The DIC does not work very well for all sample sizes. The RAVCM based method outperforms the other statistics for all sample sizes. For tests with 30 items, the RAVCM can find the true model for 85% and 87% of the items.

Table 4.6 Average proportion of items with their true model successfully distinguished by item log likelihood, AIC, DIC and the RAVCM in 100 repetition. Half of the items are generated using the CM. The other half are generated using the NCM. Each item's true model is determined using the information criterions given by the CM and the NCM at the item-level, as well as the angle between estimated item vectors in the RAVCM. The RAVCM identifies the true model as compensatory if the average angle $\leqslant \pi/4$, noncompensatory if the average angle $> \pi/4$.

Sample Size	loglik	AIC	BIC	DIC	RAVCM
500 examinees 15 items	.63	.57	.56	.51	.68
500 examinees 30 items	.73	.71	.63	.48	.85
1000 examinees 15 items	.58	.58	.51	.53	.79
1000 examinees 30 items	.80	.79	.70	.50	.87

Generally, the BIC and DIC do not perform well in distinguishing between the compensatory and non-compensatory tests and items. The reason might be that the BIC penalizes complex models to much, and the DIC require posterior distributions to be multivariate normal, which is not satisfied. The log likelihood and AIC works well for compensatory tests and items, while they sometimes fail to reject the CM when the true model is the NCM and the number of items is small, possibly due to the over-fit problem of the CM. The RAVCM generally works well, but when the true model is the CM, it sometimes incorrectly reject it. The reason might come from the estimation issues, such as the choice of prior distributions.

4.3 Conclusion and Discussion

At the test-level, the goodness of fit tests M_2, $S - \chi^2$, z_3, and the information criterions BIC and DIC do not have enough power to reject the CM when the NCM is true. In this case, there is a large probability that these common goodness of fit statistics fail to detect it. Therefore, even if the CM seems to fit well, it is possible that it is not the true model. For tests with a large number of items, the log likelihood and AIC outperform the other statistics in model selection. For tests with a small number of items, the RAVCM works better.

At the item-level, The goodness of fit tests M_2, $S - \chi^2$, z_3, and the information criterions BIC and DIC have large error rate for distinguishing the compensatory and non-compensatory items. The log likelihood and AIC did not perform well with small tests. The RAVCM has the smallest error rate under all of the 4 sample sizes in the study.

The RAVCM based method in this study takes $\pi/4$ as a threshold. If we take a larger angle as threshold, this method might give better result when the CM is the true model. In the future, it is worth examining how this threshold affect the RAVCM at distinguishing the CM and NCM tests and items.

Chapter 5

High-dimensional RAVCM

The performance and interpretation of the two-dimensional RAVCM are included in the previous chapter. In this chapter, the RAVCM with higher dimensions is discussed. For the CM and the NCM, their high-dimensional cases are easy to describe and for the CM, easy to estimate. For the RAVCM, the number of item parameters increases very fast as the number of dimensions increases. Also, it can be difficult to interpret the meaning of the item parameters for high-dimensional cases. To apply the RAVCM in practice, these problems must be solved.

The purpose of this chapter is to explore the high dimensional RAVCM and discuss the above problems. In section 5.1, two different forms of the high-dimensional RAVCM are introduced. They have linear and quadratic increases in item parameters, respectively, as the number of dimension increases. Section 5.2 shows the simulation results for the three-dimensional RAVCM. The estimability of the two forms and whether we need the more complex form, are discussed.

5.1 Two Model Forms for the High Dimensional RAVCM

It is very intuitive to imagine the high-dimensional forms of the CM and the NCM. For example, the three-dimensional CM can be written as

$$P(\boldsymbol{\theta}) = \frac{1}{1 + e^{-1.7\boldsymbol{a}(\boldsymbol{\theta} - b)}},$$

which looks the same as the two-dimensional CM. The only difference is that the $\boldsymbol{a}$ and $\boldsymbol{\theta}$ vector are three-dimensional. The three-dimensional NCM can be written as

$$P(\boldsymbol{\theta}) = \prod_{d=1}^{3} \frac{1}{1 + exp[-1.7(a_d\theta_d - b_d)]},$$

which includes one more unidimensional component. For both cases, the indices for item and examniee are omitted to simplify the representation.

These two models represent different strategies to include more parameters for the high-dimensional case. The CM simply expands the length of the item and ability vectors. The increment of parameters is linear. As the number of dimensions increases by 1, the CM only requires 1 more item parameter. The NCM, which is multiplicative, needs to include one more unidimensional component (so, two parameters). Although the increment is still linear, as one unidimensional component is related to only one ability, it makes the model much more complicated and hard to estimate. Following these two strategies, there are two ways to write the high-dimensional RAVCM. As the RAVCM is also multiplicative, a intuitive method is to include more components. The three-dimensional RAVCM with this strategy has the form:

$$\begin{aligned} P(\boldsymbol{\theta}) = & \frac{1}{1 + exp\{-1.7(a_1\theta_1 + a_2\theta_2 + a_3\theta_3 - b_1)\}} \\ & \times \frac{1}{1 + exp\{-1.7(a_4\theta_1 + a_5\theta_2 + a_6\theta_3 - b_2)\}} \\ & \times \frac{1}{1 + exp\{-1.7(a_7\theta_1 + a_8\theta_2 + a_9\theta_3 - b_3)\}}. \end{aligned}$$

The model is similar to a product of three three-dimensional CM component with the same abilities, which is very complicated. This method makes the increment of the parameters quadratic. The three-dimensional model requires $3 \times 3 = 9$ a parameters, and the four-dimensional model requires $4 \times 4 = 16$. This makes the model ugly in terms of estimation and interpretation. Also, if the item response function is a product of many compensatory components, the estimation can be very difficult and slow.

The other method is to include more dimensions in the item and ability vectors only, but not multiply more components. In this case, the three-dimensional RAVCM has the form:

$$\begin{aligned} P(\boldsymbol{\theta}) = & \frac{1}{1 + exp\{-1.7(a_1\theta_1 + a_2\theta_2 + a_3\theta_3 - b_1)\}} \\ & \times \frac{1}{1 + exp\{-1.7(a_4\theta_1 + a_5\theta_2 + a_6\theta_3 - b_2)\}}. \end{aligned}$$

For this form, the increment of the parameters is linear, which is the same as the CM and the NCM. The three-dimensional model only requires $2 \times 3 = 6$ a parameters, and the four-dimensional model requires $2 \times 4 = 8$.

Considering simplicity, the second (simple) form is much more acceptable than the first one (complex). For interpretation, the RAVCM can still be considered as a multi-step model. In the high-dimensional case, each step might depend on more abilities, but the item does not have to include more steps. As such, it is reasonable to use the simpler model form. However, as dimensionality increases, the complexity of compensation increases very fast. Compared to the complex model form, the simple model form can approximate less types of compensation. For example, the simple form cannot approximate the three-dimensional NCM, while the complex form can. For the simple form, we also examine how well it performs when the complex form is the true model.

5.2 Simulation Study

In this section, the purpose is to briefly introduce the estimation of the high-dimensional RAVCM, and to examine if the simple form of the three-dimensional RAVCM is estimable, and how it performs when the complex form is the true model via simulation.

The estimation can be implemented via the Metropolis-Hasting within Gibbs sampling, which is the same algorithm for the two-dimensional RAVCM. Most settings and details are the same as in section 3.2.2. The only additional problem for the three-dimensional model is identifiability. When the first item vector of the first item is fixed on the θ_1 axis, the first ability is identifiable, but the second and the third are not. The label of θ_2 and θ_3 and their related a parameter can switch without changing the probability of correct response. Without adding more unidimensional items, this problem always happens. Fortunately, the dimensional switch only happens once for the MCMC procedure (at the beginning), and it does not harm the stability of the MCMC chain. In simulation, one can check if the dimension switch happens by calculating the RMSEs of θ_2 and θ_3. The RMSEs are large when the label of θ_2 and θ_3 are switched. In this case, the true RMSEs for θ_2 and θ_3 should be calculated using the true θ_3 and θ_2, respectively. The item vectors of the three dimensional RAVCM can also be exchanged, similar to the two-dimensional RAVCM. To solve the problem, we assume that the first item vector has smaller angle with θ_1 axis (very similar to the two-dimensional RAVCM). In MCMC estimation, we can simply reject the generated values that do not satisfy the constraint. Or, more efficiently, if the angle between the second generated item vector and the θ_1 axis is larger, we can switch the label of the first item vector and the second to satisfy this constraint. In this study, the first method is used.

Three simulation studies are included in this section. The true models for the three studies are the three-dimensional CM, three-dimensional RAVCM (simple form), and three-dimensional RAVCM (complex form). The fitted models are the three-dimensional CM for the first study, and the three-dimensional RAVCM (simple) form for the second and third study. The RMSEs of estimated abilities and item response surface as well as log likelihood of the fitted model are recorded. Since the CM is considered as easy to estimate, the first study can give a standard for "good estimates" that can be compared with the estimates given by the RAVCM in terms of the size of the error. The purpose of the second simulation is to examine the estimability of the three-dimensional RAVCM. In the third part, we want to see whether the simple form of the RAVCM can still estimate the abilities well when the complex form is the true model.

In these simulations, the sample sizes are the same as in section 3.3 (500 or 1000 examinees, 15 or 30 items). The result of each simulation study is based on an average of 50 replications. The three ability parameters are generated from standard normal distributions, and are independent to each other. The a parameters are generated from truncated normal distribution, as in section 3.3. The b parameters are generated from normal distribution with variance 1.5 and mean 0 for the CM, -1 for the simple form of RAVCM, -1.5 for the complex form of the RAVCM, since the overall difficulty is larger when the number of components in the RAVCM is larger.

Table 5.1 shows the result of the first simulation study. When the CM is the true and fitted model, the RMSEs of the estimated abilities and item response surfaces are relatively small, especially when the number of item is large. The log likelihoods are even larger than the two-dimensional CM. The result shows that the three-dimensional CM can be well estimated using the MCMC method.

Table 5.1 Performance of the three-dimensional CM when it is the true model. ("R" is the abbreviation of "RMSE".)

Sample Size	R(θ_1)	R(θ_2)	R(θ_3)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees, 15 items	0.65	0.70	0.70	0.080	-1950
500 examinees, 30 items	0.54	0.59	0.61	0.077	-4531
1000 examinees, 15 items	0.66	0.69	0.71	0.081	-4448
1000 examinees, 30 items	0.55	0.61	0.62	0.073	-8037

Table 5.2 shows the RMSEs and log likelihoods for the estimated three-dimensional RAVCM. Compared to the RMSEs and log likelihoods in Table 5.1, the abilities and item response surfaces are not recovered as well as the CM. When sample size is large (1000 examinees, 30 items), the RMSEs for the abilities and item response surfaces are relatively similar to the RMSEs for the CM. The overall result shows that the simple form of the three-dimensional RAVCM is estimable, but it cannot be estimated as well as the CM. To better recover the abilities and item response surfaces, a large number of items is important.

Table 5.2 Performance of the three-dimensional RAVCM (simple form) when it is the true model. ("R" is the abbreviation of "RMSE".)

Sample Size	R(θ_1)	R(θ_2)	R(θ_3)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees, 15 items	0.64	0.75	0.78	0.117	-2486
500 examinees, 30 items	0.62	0.63	0.64	0.115	-5346
1000 examinees, 15 items	0.69	0.71	0.77	0.120	-5285
1000 examinees, 30 items	0.55	0.64	0.67	0.096	-9808

Table 5.3 shows that when the true model is the complex form of the RAVCM, the simple form can recover the abilities and item response surfaces to some extent. The RMSEs are generally larger than those in Table 5.2, while similar for the 1000

examinees, 30 items case. This means that the simple form can approximate the complex form better when sample size is large. For smaller sample sizes, the estimates are not very satisfying.

Table 5.3 Performance of the three-dimensional RAVCM (simple form) when the true model is the complex form. (“R” is the abbreviation of “RMSE”.)

Sample Size	R(θ_1)	R(θ_2)	R(θ_3)	R($\hat{P}(\boldsymbol{\theta})$)	Log like
500 examinees, 15 items	0.73	0.76	0.83	0.124	-2615
500 examinees, 30 items	0.68	0.69	0.72	0.130	-5012
1000 examinees, 15 items	0.72	0.77	0.84	0.130	-5441
1000 examinees, 30 items	0.63	0.69	0.70	0.097	-10500

5.3 Conclusion and Discussion

From the simulation results, we conclude that the simple form of the RAVCM is estimable via the Metropolis-Hastings within Gibbs Sampler algorithm, although it cannot be estimated well as the three-dimensional CM. When the true model is either the simple form, or the complex form of the RAVCM, a large sample, especially large number of items is needed for good estimates. Using the simple model to recover the more complex one raises issues similar to using the CM in cases where the true model is the NCM or RAVCM.

The interpretation for the high dimensional RAVCM is difficult, and there is no easy visualization tool, such as a contour plot, for the item response functions. It is hard to imagine the compensation in a high dimensional space. A good interpretation for the RAVCM is still to consider it as a multi-step model. The only difference from the two-dimensional case is that the items require more different abilities in each step for the high dimensional RAVCM.

Chapter 6

Conclusion and Future Directions

This dissertation introduced the RAVCM, a new MIRT model that allows for different compensation levels and rotation of the item vectors. Generally, compared to the other MIRT models, the RAVCM can better deal with potential structures that could be encountered in practice, such as correlated abilities, or different true structures for different items. For the NCM, the item parameters cannot be well recovered when the correlation is misspecified. In practice, this will always happen since the true correlation is latent. The RAVCM can be estimated under any specified correlation structure, and the estimated model can be transformed to any other correlation structures, keeping the probabilities in the model the same. The ability to fit such a model is important if MIRT is to be used in practice, and also for ensuring that simulation studies of other methods are carried out under realistic circumstances.

When it is known that the CM is true for all items, the CM is always the best model since it is easy to estimate and interpret and has smaller estimation error. However, in real tests, it is impossible that all items are completely compensatory. In some cases, the test maker might want to interpret the abilities as non-compensatory or not completely compensatory. In some cases, the problem is that we do not know how compensatory each item is. The RAVCM can help do these exploratory tasks since it can approximate both the CM and the NCM, and it allows the compensation levels of items in a test to be different. When some items are not compensatory, the RAVCM gives better overall estimation than the CM.

The definition of compensation is discussed. It is suggested that we consider compensation as the full information of how abilities combine, or the shape of contour plot, instead of a simpler specific model form or a parameter. With this definition, there are many types of compensation in addition to compensatory and non-compensatory. If the purpose is to approximate the actual item response surface to have more accurate ability estimates, a flexible model such as the RAVCM can be useful.

The effectiveness of several statistics, including the angle between item vectors of the RAVCM, are evaluated at distinguishing between the CM and the NCM at the test-level and item-level. The results show that when the misfit comes from the misspecification of compensation, the common goodness of fit tests such as $S - \chi^2$ do not work well in this case, while a method based on the the RAVCM can do a relatively good job. (If the true model is the CM for all items, the log likelihood and AIC work better.) This result validates the importance of the RAVCM in exploratory studies for finding the compensation for each item.

In Chapter 5, a complex and a simple version of the high dimensional RAVCM are proposed. The high dimensional RAVCM is difficult to estimate and interpret. The simulation results show that the simple one is estimable and can perform well when the sample size is large. The estimates generally have larger average error compared to the CM.

Possible future work includes improving the estimation of the RAVCM, especially the high dimensional RAVCM. In many cases, the RMSEs of ability estimates given by the RAVCM are only a little smaller than the CM even if the CM is not the true model for all of the items. For the high dimensional case, the RMSEs for the RAVCM are relatively large. These show that the RAVCM is not perfectly estimated. As the RAVCM is very flexible, its MCMC estimation needs more iterations to converge, and sometimes has more identifiability problems. A better algorithm is needed to handle this.

Another area of future work is to find a statistic that can evaluate the goodness of item response surface recovery in MIRT. In simulation, one can check the difference between the true and estimated surfaces at where the abilities exist. In practice, there is no such a statistic. If we treated the RAVCM as a tool to explore the compensation of the items, then a statistic to evaluate the fitted RAVCM is needed.

Finally, it is very important to have more real data examples. These examples could demonstrate the usefulness of the RAVCM, and provide recommendations for future research to improve the model.

Bibliography

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and non-compensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*(4), 255-278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20*, 311-329.

Ackerman, T. A., & Bolt, D. B. (1995). How different cognitive strategies produce differential item performance. In *Annual Meeting of the American Educational Research Association, San Francisco, CA.*

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

Ackerman, T. A., & Turner, R. C. (2003). Estimation and application of a generalized MIRT model: Assessing the degree of compensation between two latent abilities. National Council for Measurement in Education conference, Chicago, IL.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika, 60*(2), 255-265.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17*(3), 251-269.

Babcock, B. G. E. (2009). *Estimating a noncompensatory IRT model using a modified Metropolis Algorithm* (Doctoral dissertation, University of Minnesota).

Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement, 35*, 317-329.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*(2), 153-169.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika, 66*, 541-561.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*, 395-479.

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

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395-414.

Chalmers, R. P., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement, 38*(5), 339-358.

De La Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295-311.

Dean, D. (2000). Statistical Shape Analysis. *Journal of Human Evolution, 38*(3), 455-457.

DeMars, C. E. (2016). Partially-compensatory multi-dimensional IRT models: Two alternate model forms. *Educational and Psychological Measurement, 76*, 231-257.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*(2), 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67-86.

Fraser, C., & McDonald, R. P. (1988). *NOHARM II: A FORTRAN program for* *fitting unidimensional and multidimensional normal ogive models of latent trait theory.* The University of New England, Armidale, Australia.

Fu, Y. (2015). *Dimensionality Assessment and Estimation for the Variable Compensation Model* (Doctoral dissertation, University of South Carolina).

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics, 5*(599-608), 42.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications.* Springer Science & Business Media.

Hendrix, L. (2011). *Fast EM Based Posterior Approximation for IRT Item Parameters* (Doctoral dissertation, University of South Carolina).

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187-200.

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51.

Kirisci, L., & Hsu, T. C. (1995). The Robustness of BILOG to Violations of the Assumptions of Unidimensionality of Test Items and Normality of Ability Distribution.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*(4), 269-290.

Li, Y., Jiao, H., & Lissitz, R. 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*(2).

Liaw, Y. L. (2015). *When can Multidimensional Item Response Theory (MIRT) Models be a Solution for Differential Item Functioning (DIF)? A Monte Carlo Simulation Study* (Doctoral dissertation).

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: a unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75-106.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298.

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

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*(1), 49-57.

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

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

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

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151-1172.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Simpson, M. A. (2005). *Use of a variable compensation item response model to assess the effect of working-memory load on noncompensatory processing in an inductive reasoning task* (Doctoral dissertation, University of North Carolina at Greensboro).

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, 59(2), 429-449.

Smith, J. (2009). *Some Issues in Item Response Theory: Dimensionality Assessment and Models for Guessing* (Doctoral dissertation, University of South Carolina).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

Sympson, J. B. (1978). *A model for testing with multidimensional items.* In Proceedings of the 1977 computerized adaptive testing conference (No. 00014). Minneapolis, MN: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39-49.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional IRT model. *Applied Psychological Measurement, 39*(2), 119-134.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*(3), 239-252.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*(3), 339-352.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245-262.

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*(2), 181-196.

Appendix A

R Code for Estimating the Two-dimensional RAVCM

```
# Acknowlegement: This R code is developed based on Yin Fu
   's code for estimating the CM and the NCM (2015).

# To use the Rcpp packages, Rtools is needed to be
   installed and added to the system path.

# The parameters below should be carefully adjusted to
   receive good estimates.
# Number of items
R_nitems=30
# Number of examinees
R_nexamn=1000

# Number of iterations
iterin=30000

# Initial values
a1in=rep(1,R_nitems)
a2in=rep(0,R_nitems)
```

```
a3in=rep(0,R_nitems)
a4in=rep(1,R_nitems)
b1in=rep(-0.5,R_nitems)
b2in=rep(-0.5,R_nitems)
theta1in=rep(0,R_nexamn)
theta2in=rep(0,R_nexamn)

# Prior of rho
rin=0

# Shape and Scale parameter of Weibull prior
wshain=1.25
wsclin=1

# Input data (A R_nexamn x R_nitems matrix)
datain=U

# Standard deviations of proposal distributions of theta1,
    theta2, a, b
c1in=1.50
c2in=1.50
cain=0.35
cbin=0.48

# Standard error and Mean of prior of b
pcbin=1
pmbin=-0.5
```

```
library(inline)
library(RcppArmadillo)
library(devtools)

############### MCMC for estimating the RAVCM (C++ code)
   ######################

mcmcRAVCM <- '
using namespace Rcpp;
using namespace arma;
using namespace std;

// parameters from R;
int nitems = as<int>(R_nitems);
int nexamn = as<int>(R_nexamn);
int niter = as<int>(iterin);
vec a1vec=as<vec>(a1in);
vec a2vec=as<vec>(a2in);
vec a3vec=as<vec>(a3in);
vec a4vec=as<vec>(a4in);
// set a11=1
a1vec[1]=1;
a2vec[1]=0;

vec b1vec=as<vec>(b1in);
vec b2vec=as<vec>(b2in);
```

```
// set b11=b12=-1;
b1vec[1]=-0.5;
b2vec[1]=-0.5;
vec theta1vec=as<vec>(theta1in);
vec theta2vec=as<vec>(theta2in);
float rho = as<float>(rin);
// shape and scale of weibull distribution (a prior)
float wsha = as<float>(wshain);
float wscl = as<float>(wsclin);
mat simu = as<mat>(datain);


// the standard deviation of the proposal distribution
float c1 = as<float>(c1in);
float c2 = as<float>(c2in);
float ca = as<float>(cain);
float cb = as<float>(cbin);
// the mean and standard deviation of the prior
   distribution;
float theta1 = mean(theta1vec);
float theta2 = mean(theta2vec);
float pcb = as<float>(pcbin);
float pmb = as<float>(pmbin);

// define matrix/vector that stores the chain;
mat a1mat(nitems,niter);
mat a2mat(nitems,niter);
mat a3mat(nitems,niter);
mat a4mat(nitems,niter);
mat b1mat(nitems,niter);
mat b2mat(nitems,niter);
mat theta1mat(nexamn,niter);
mat theta2mat(nexamn,niter);
a1mat.fill(0);
a2mat.fill(0);
a3mat.fill(0);
a4mat.fill(0);
b1mat.fill(0);
b2mat.fill(0);
theta1mat.fill(0);
theta2mat.fill(0);
float num1=1;


// set initial values (first column = starting values);
for(int i=0; i<nitems; i++) {
a1mat(i,0) = a1vec[i];
a2mat(i,0) = a2vec[i];
a3mat(i,0) = a3vec[i];
a4mat(i,0) = a4vec[i];
b1mat(i,0) = b1vec[i];
b2mat(i,0) = b2vec[i];
}
for(int i=0; i<nexamn; i++) {
```

```
theta1mat(i,0) = theta1vec[i];
theta2mat(i,0) = theta2vec[i];
}


// to record the acceptance rates of MCMC M-H; for the
   parameters at each iteration;
vec acceptance_theta(niter-1);
vec acceptance_a(niter-1);
vec acceptance_b(niter-1);
acceptance_theta.fill(0);
acceptance_a.fill(0);
acceptance_b.fill(0);

// MCMC core code
for(int i=0; i<(niter-1); i++) {
// load values from the previous draw

vec a1_prv(nitems);
vec a2_prv(nitems);
vec a3_prv(nitems);
vec a4_prv(nitems);
vec b1_prv(nitems);
vec b2_prv(nitems);
vec theta1_prv(nexamn);
vec theta2_prv(nexamn);
```

```
for(int l=0;l<nitems;l++){
a1_prv[l]=a1mat(l,i);
a2_prv[l]=a2mat(l,i);
a3_prv[l]=a3mat(l,i);
a4_prv[l]=a4mat(l,i);
b1_prv[l]=b1mat(l,i);
b2_prv[l]=b2mat(l,i);
}

for(int l=0;l<nexamn;l++){
theta1_prv[l]=theta1mat(l,i);
theta2_prv[l]=theta2mat(l,i);
}

vec theta1_new(nexamn);
vec theta2_new(nexamn);
vec a1_new(nitems);
vec a2_new(nitems);
vec a3_new(nitems);
vec a4_new(nitems);
vec b1_new(nitems);
vec b2_new(nitems);

theta1_new=theta1_prv;
theta2_new=theta2_prv;
a1_new=a1_prv;
a2_new=a2_prv;
```

```
a3_new=a3_prv;
a4_new=a4_prv;
b1_new=b1_prv;
b2_new=b2_prv;

//update theta first ;
for(int j=0;j< nexamn;j++){
//int j=0;
/*a) theta_1j^* (theta1s), theta_2j^*(theta2s) from the
   proposal
distributions
the standard deviation of the proposal distributions */
// simulate theta1 and theta2 from multivariate normal
vec thetastar(2);
vec mustar(2);
mustar[0]= theta1_prv[j];
mustar[1]= theta2_prv[j];
mat sigstar(2,2);
sigstar(0,0) = pow(c1,2);
sigstar(1,1) = pow(c2,2);
sigstar(0,1) = rho * c1 * c2 ;
sigstar(1,0) = rho * c1 * c2 ;
thetastar= mvrnorm(mustar,sigstar);
float theta1_star = thetastar[0];
float theta2_star = thetastar[1];

/* b) calculate the acceptance rate for each j (examinee);
```

```
    P(Xj| theta
.star,beta.prv)
use v to replace item "i" to avoid confusion since i
   represent
iteration; */
float prob_jsp= 1;
for(int v=0;v<nitems;v++){
prob_jsp= prob_jsp*prob_xij(a1_prv[v], a2_prv[v], a3_prv[v
   ], a4_prv[v], b1_prv[v], b2_prv[v], theta1_star, theta2
   _star,simu(j,v));
}

float prob_jpp= 1;
for(int v=0;v<nitems;v++){
prob_jpp = prob_jpp* prob_xij(a1_prv[v], a2_prv[v], a3_prv
   [v], a4_prv[v], b1_prv[v], b2_prv[v], theta1_prv[j],
   theta2_prv[j],simu(j,v));
}

float alpha_accept1;
alpha_accept1= (prob_jsp* exp(-(pow(theta1_star,2)/1.5+pow
   (theta2_star,2)/1.5-pow(theta1_prv[j],2)/1.5-pow(theta2
   _prv[j],2)/1.5 -2*rho*(theta1_star*theta2_star - theta1
   _prv[j]*theta2_prv[j])/1.5)/(2*(1-pow(rho,2)))))/(prob_
   jpp);
alpha_accept1 = min( alpha_accept1 , num1);
```

```
// take the new value with acceptance rate rtheta,
   otherwise, keep the value from the previous status;
float ramdom_unif = ::Rf_runif(0,1);
if (alpha_accept1> ramdom_unif ) {
theta1_new[j] = theta1_star;
theta2_new[j] = theta2_star; }
// this is the end of updating thetas;
}



//check acceptance rate of theatas ;
float acpt_theta = 1;
for(int v=0;v<nexamn;v++) {
if(theta1_new[v]!=theta1_prv[v]){ acpt_theta=acpt_theta
   +1;}
}
acceptance_theta[i] = acpt_theta/nexamn;

//update chain

for(int v=0; v<nexamn; v++) {
theta1mat(v,i+1) = theta1_new[v] ;
theta2mat(v,i+1) = theta2_new[v] ;
}
```

```
// update a
for(int v=0;v< nitems;v++){
/*a) a and b from the proposal distributions the standard
   deviation of
the proposal distributions
should be changed such that the acceptance rate is around
   25% */
float a1_star;
float a2_star;
float a3_star;
float a4_star;
a1_star= ::Rf_rnorm(a1_prv[v], pow(ca,2));
a2_star= ::Rf_rnorm(a2_prv[v], pow(ca,2));
a3_star= ::Rf_rnorm(a3_prv[v], pow(ca,2));
a4_star= ::Rf_rnorm(a4_prv[v], pow(ca,2));
if(a1_star<=0){a1_star=0.001;}
if(a2_star<=0){a2_star=0;}
if(a3_star<=0){a3_star=0;}
if(a4_star<=0){a4_star=0.001;}

// Fix the first item vector;
if(v==0){
a1_star=1;
a2_star=0;
}
```

```
//b) calculate the acceptance rate for each v (item)
//P(Xj| theta.star,beta.prv)
// prob_ratio = prob_vsp/prob_vpp
float prob_ratio =1;
for(int j=0;j<nexamn;j++){
prob_ratio= prob_ratio* prob_xij(a1_star, a2_star, a3_star
   , a4_star, b1_prv[v], b2_prv[v], theta1_new[j], theta2_
   new[j],simu(j,v))/prob_xij(a1_prv[v], a2_prv[v], a3_prv
   [v], a4_prv[v], b1_prv[v], b2_prv[v], theta1_new[j],
   theta2_new[j],simu(j,v));
}
float alpha_accepta;


// (a1,a2) should have smaller angle from theta1 axis;

if (a1_star/(pow(a1_star,2)+pow(a2_star,2))<a3_star/(pow(
   a3_star,2)+pow(a4_star,2))){
alpha_accepta=0;
}
else {
alpha_accepta= prob_ratio*pow(a1_star*a4_star/(a1_prv[v]*
   a4_prv[v]),wsha-1)*exp((pow(a1_prv[v],wsha) + pow(a4_
   prv[v],wsha)- pow(a1_star,wsha) -pow(a4_star,wsha))/pow
   (wscl,wsha))*exp(2*(a2_prv[v]+a3_prv[v]-a2_star-a3_star
   )) ;
}
```

```
alpha_accepta = min( alpha_accepta , num1);

// take the new value with acceptance rate rtheta,
   otherwise, keep the value from the previous status;
float ramdom_unif = ::Rf_runif(0,1);
if (alpha_accepta> ramdom_unif ) {
a1_new[v] = a1_star;
a2_new[v] = a2_star;
a3_new[v] = a3_star;
a4_new[v] = a4_star;
}
//below is the end of updating a,b;
}

//update chains of a
for(int v=0; v<nitems; v++) {
a1mat(v,i+1) = a1_new[v];
a2mat(v,i+1) = a2_new[v];
a3mat(v,i+1) = a3_new[v];
a4mat(v,i+1) = a4_new[v];
}
// Fix the first item vector
a1mat(0,i+1) = 1;
a2mat(0,i+1) = 0;
```

```
//check acceptance rate of a
float acpt_a = 1;
for(int v=0;v<nitems;v++) {
if(a1_new[v]!=a1_prv[v]){ acpt_a=acpt_a +1;}
}
acceptance_a[i] = acpt_a/nitems;



// update b
for(int v=1;v< nitems;v++){
/*a) a and b from the proposal distributions the standard
   deviation of
the proposal distributions
should be changed such that the acceptance rate is around
   25% */
float a1_star;
float a2_star;
float a3_star;
float a4_star;
a1_star= a1_new[v];
a2_star= a2_new[v];
a3_star= a3_new[v];
a4_star= a4_new[v];
float b1_star;
float b2_star;
b1_star= ::Rf_rnorm(b1_prv[v],pow(cb,2));
```

```
b2_star= ::Rf_rnorm(b2_prv[v],pow(cb,2));



//b) calculate the acceptance rate for each v (item)
float prob_ratio =1;
for(int j=0;j<nexamn;j++){
prob_ratio= prob_ratio* prob_xij(a1_new[v], a2_new[v], a3_
    new[v],a4_new[v],b1_star,b2_star, theta1_new[j], theta2
    _new[j],simu(j,v))/prob_xij(a1_new[v], a2_new[v], a3_
    new[v],a4_new[v],b1_prv[v], b2_prv[v], theta1_new[j],
    theta2_new[j],simu(j,v));
}

float alpha_acceptb;
alpha_acceptb= prob_ratio*exp((pow(b1_prv[v]-pmb,2)-pow(b1
    _star-pmb,2)+pow(b2_prv[v]-pmb,2)-pow
(b2_star-pmb,2))/(2*pow(pcb,2)));
alpha_acceptb = min( alpha_acceptb , num1);

// take the new value with acceptance rate rtheta,
    otherwise, keep the value from the previous status;
float ramdom_unif = ::Rf_runif(0,1);
if (alpha_acceptb> ramdom_unif ) {
b1_new[v] = b1_star;
b2_new[v] = b2_star;}

//Reject small b
```

```
if (b1_star < -2) {
b1_new[v] = b1mat(v,i);}
if (b2_star < -2) {
b2_new[v] = b2mat(v,i);}

//below is the end of updating b;
}

//update chains of b
for(int v=0; v<nitems; v++) {
b1mat(v,i+1) = b1_new[v];
b2mat(v,i+1) = b2_new[v];
}
// Fix the difficulty of the first item
b1mat(0,i+1) = -0.5;
b2mat(0,i+1) = -0.5;

//check acceptance rate of b ;
float acpt_b = 1;
for(int v=0;v<nitems;v++) {
if(b1_new[v]!=b1_prv[v]){ acpt_b=acpt_b +1;}
}

acceptance_b[i] = acpt_b/nitems;
//below is the end of iterations;
}
```

```
return List::create(Named("a1")=a1mat,Named("a2")=a2mat,
   Named("a3")=a3mat,Named("a4")=a4mat,Named("b1")=b1mat,
   Named("b2")=b2mat,Named("theta1")=theta1mat,Named("
   theta2")=theta2mat,Named("Acceptance_theta")=acceptance
   _theta,Named("Acceptance_a")=acceptance_a,Named("
   Acceptance_b")=acceptance_b,Named("theta1value")=theta1
   );
'




# for the RAVCM, the functions included in the main code
   by C++ are defined here below in "incravcm"
incravcm <- '
float irf_ravcm (float a1,float a2,float a3,float a4,float
    b1,float b2,float theta1,float theta2){
float result;
result= 1/(1+exp(-1.7*(a1*theta1+a2*theta2-b1))+exp(-1.7*(
   a3*theta1+a4*theta2-b2))+exp(-1.7*(a1*theta1+a2*theta2+
   a3*theta1+a4*theta2-b1-b2)));
return (result);
}

float prob_xij (float a1,float a2,float a3,float a4,float
   b1,float b2,float theta1,float theta2,float xij){
float pij = irf_ravcm(a1,a2,a3,a4,b1,b2, theta1,theta2);
float result = pow(pij,xij) * pow((1-pij),(1-xij));
return (result);
}

arma::vec mvrnorm(arma::vec mu, arma::mat sigma){
int ncols = mu.size();
arma::vec Y(ncols);
for (int i=0; i<ncols; i++){
Y[i] = norm_rand();
}
arma::mat temp = ((arma::chol(sigma)).t())*Y;
arma::vec res = mu + temp.col(0);
return( res );
}
'


#compile the c++ code in R and get the R function to run
   estimation

estimate.RAVCM<- cxxfunction(signature(iterin="int", R_
   nitems="int",
R_nexamn="int", a1in="numeric", a2in="numeric",a3in="
   numeric",a4in="numeric",b1in="numeric",
b2in="numeric", theta1in="numeric", theta2in="numeric" ,
   rin="numberic",datain="numeric",c1in="numeric",c2in="
   numeric",cain="numeric",cbin="numeric",wshain="numeric"
   ,wsclin="numeric",pcbin="numeric",pmbin="numeric"),
   mcmcRAVCM, plugin = "RcppArmadillo",includes=incravcm)

################## The end of compiling
   ####################################

# Run MCMC code
testresult=estimate.RAVCM(iterin,R_nitems,R_nexamn,a1in,
   a2in,a3in,a4in,b1in,b2in,theta1in,theta2in,rin,datain,
   c1in,c2in,cain,cbin,wshain,wsclin,pcbin,pmbin)

# get the acceptance rates from the output
Acceptance.theta= testresult$Acceptance_theta
Acceptance.a = testresult$Acceptance_a
Acceptance.b = testresult$Acceptance_b
# calculate the acceptance rates of the examinee & item
   paramers
mean(Acceptance.theta)
mean(Acceptance.a)
mean(Acceptance.b)

# Get the values in the chains
a1.out = testresult$a1
a2.out = testresult$a2
a3.out = testresult$a3
a4.out = testresult$a4
b1.out = testresult$b1
```

```
b2.out = testresult$b2

theta1.out = testresult$theta1

theta2.out = testresult$theta2
```