Designing *p*-Optimal Item Pools for Multidimensional Computerized Adaptive Testing

By

Liyang Mao

A DISSERTATION

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

Measurement and Quantitative Methods - Doctor Of Philosophy

2014

ABSTRACT

DESIGNING *P*-OPTIMAL ITEM POOLS FOR
MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTS

By

Liyang Mao

The interest in multidimensional computerized adaptive testing (MCAT) has grown considerably over the last few years. While a significant amount of research has been conducted on item selection and ability estimation methods for MCAT, few studies specifically addressed the item pool design for MCAT. To ensure a proper functioning of MCAT, a well-designed item pool is imperative. A well-designed item pool should consist of a number of well-balanced items that achieve appropriate test precision, item usage, as well as lower the cost of item creation. One method to develop such an item pool is the *p*-optimality method, which is proposed by Reckase (2003 & 2007) for unidimensional CAT. This paper aims to develop *p*-optimal item pools for MCAT by extending the Reckase's method to a multidimensional context.

The extension includes the generation of a multidimensional optimal item based on the D-Optimality item selection creation, the definition of the MDIFF-bin to describe multidimensional item succinctly for item pool design, and the interpretation for the *p*-optimal item pool in a multidimensional context. In this paper, a total of 24 *p*-optimal item pools were designed and then developed for different test specification, with different correlation among dimensions, based on different bin size, and under the condition with or without item exposure control. The characteristics for the 24 *p*-optimal item pools are summarized. A simulation study was conducted to evaluate the performance of the *p*-optimal item pools against baseline pools existing in research literature.

Results show that *p*-optimal item pools achieve similar levels of measurement accuracy as baseline pools, but they consist of fewer items and perform better in terms of item pool usage and test security. The characteristics and the performance of the *p*-optimal item pools are affected by factors such as test specification, correlation among dimensions, bin size, and item exposure control. The results in this study can provide a general guideline for the item pool development for MCAT. More importantly, because the *p*-optimal item pool is specifically tailored to the MCAT programs, the *p*-optimal item pool design procedure described in this study can be adapted to other MCAT programs with different features and purposes. The end product of the *p*-optimal item pool design can be used as an instructive guide for item creation, item pool development, and item pool management.

ACKNOWLEDGEMENTS

This dissertation is not just a research study. It reflects how I have grown during my studies at Michigan State University. Without the guidance, support, encouragement, and care from many people, I could not accomplish this dissertation and graduate with a PhD degree.

I want to express my deepest appreciation to my advisor, Mark Reckase. I thank him for giving me the great opportunities to work with him on both coursework and research projects; I thank him for his profound knowledge and enthusiastic that encourages me to be a better scholar; I thank him for his patience and support when I exploring my research interests; and I also thank him for his generous help when I need him.

I also want to express my gratitude to Edward Roeber, a member of my dissertation committee and guidance committee, for his support, encouragement, and also warm care throughout my PhD study. My gratitude also extends to my dissertation committee members, Joseph Martineau and Richard Houang, who have provided me insightful suggestion and help for my dissertation.

I also would like to thank Bettie Menchik for providing me the opportunity to work on several education policy projects for the Michigan Department of Education. I would attribute most of my knowledge about education policy to her generous help and guidance. I truly enjoyed working with her and we have become very good friends. I also thank Neelam Kher, Michelle William, and Kimberly Maier for their support during my study.

I am sincere grateful for my friends Tingqiao Chen, Chang Chi, Emre Gonulates, Eun Hye Ham, Xin Luo, Bing Tong, Keyin Wang, Xuechun Zhou, and many others, who have enriched my life in graduate school.

I also want to thank my wonderful boyfriend, Jianxun Wang, for making me smile and happy in my life.

Finally, I want to express my appreciation to my parents, Yukun Hou and Yingjian Mao, for their unconditional love since I was born.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

Over the last few decades, computerized adaptive testing (CAT) has achieved great popularity in educational assessments. Different from a conventional paper-and-pencil test, CAT uses a computer to deliver test items that are selected by tailoring each item to the ability level of an examinee. Such delivery of tests has several advantages, such as increasing measurement precision, reducing testing time, faster score reporting, and flexible scheduling of examinees (Wainer, 2000). Starting in the 1990s, CAT has been successfully applied to many operational testing programs, including the Armed Services Vocational Aptitude Battery (ASVAB), the Computerized Adaptive Placement Assessment and Support Services (COMPASS), the Graduate Management Admission Test (GMAT), and the National Council Licensure Examination (NCLEX). Furthermore, in the 2014-15 school year, almost half states in the United State will replace their current K-12 assessments by the CAT-based assessment system developed by the Smarter Balanced Assessment Consortium (SBAC, 2013).

Most operational adaptive tests have been developed based on a unidimensional item response theory (UIRT) model. Nevertheless, the interest in CAT based on multidimensional item response theory (MIRT) models (refer to as multidimensional CAT, MCAT) has grown considerably as shown by the increasing number of articles in the literature (e.g., Segall, 2010; Seitz & Frey, 2013; Wang & Chang, 2011; Yao, 2013). One reason that MCAT has become very popular is that current educational assessments often cover multiple content standards, so that those assessments may not be strictly unidimensional (Reckase, 2009). In a mathematics test for Grade 4, for example, there is a concern about providing an adequate number of *algebra*, *geometry*, *number operation*, *data analysis*, and *measurement* items to each examinee, because these content areas are defined as separate components of mathematics proficiency by the

Common Core State Standards (CCSS, 2010). In this situation, it would be straightforward to apply MCAT for assessments with multidimensional features.

In addition, MCAT would be preferred when diagnostic information (i.e., subscores) is to be reported. In educational assessments, although the total score is useful for some decision making, subscores complement the total score by providing information about examinees' strengths and weaknesses on each content area. Therefore, test users usually ask for subscores for diagnostic purposes. Teachers also prefer subscores because subscores can help them design specific instruction for each student. In MCAT, subscores on all content areas can be estimated simultaneously using a MIRT model. In unidimensional CAT (UCAT), however, the UCAT needs to be carried out separately, one content area at a time, to estimate each subscore one by one. Therefore, in subscore estimation, MCAT often yields better measurement efficiency than UCAT (Luecht, 1996; Segall, 1996; Wang & Chen, 2004; Yao, 2012; Mao, Luo & Zhou, 2013).

Like a UCAT program, a MCAT program also consists of several components and procedures. It begins with an item pool that contains an adequate number of items calibrated using a MIRT model. Then, the MCAT methods usually follow an iterative process: (1) assign an initial ability level to an examinee, (2) select a test item from the item pool using an item selection method, (3) administer the selected test item to a examinee and collect the response, and (4) score the response and update the ability estimates. This process continues until a certain stopping criterion is met. Operational implementation of CAT often includes constraints such as content balancing and item exposure control to address the validity and security issue.

While a significant amount of research has been conducted on generalizing the item selection and ability estimation methods from UCAT to MCAT (Mulder & van der Linden, 2009; Segal, 1996; Wang & Chang, 2011; Yao, 2013), few studies can be found that specifically addressed

the item pool design issue for MCAT. In all of the existing studies about MCAT, the multidimensional item pools are either built from pure simulation (i.e., van der Linden, 1999) or created from operational UCAT programs or paper-and-pencil tests (i.e. Diao, 2009; Song, 2010; Yao, 2013). The quality of these item pools is unknown. Because a CAT program cannot function well without an item pool that contains sufficient number of appropriate items for all the examinees, item pool design is critical for MCAT programs. Therefore, in order to design quality item pools for MCAT, current item pool design methods for UCAT need to be generalized to MCAT.

For UCAT programs, there are two methods focusing on item pool design: one is the shadow test approach (Veldkamp & van der Linden, 2000); the other one is the *p*-optimality approach (Reckase, 2003 & 2010). According to Veldkamp and van der Linden (2000), before items are selected to administer, a shadow test is first assembled from a large item pool (usually called "master pool") using a linear integer programming model. Then a test item is selected from the shadow test, not directly from the item pool, to administer. The integer programming model guarantees that all constraints (i.e., content balancing and item exposure control) on test administration can be met. However, it is still unclear how to design a master pool and what are the desired features of a master pool. Without a multidimensional master pool, the shadow test approach cannot be implemented for MCAT programs.

Unlike the shadow test approach, the *p*-optimality approach developed by Reckase (2003 & 2010) directly addressed the item pool design issue. The definition of Reckase's *p*-optimal item pool is an item pool "that always has an item available for selection that *p*% matches the desired characteristics specified by the item selection routine for the CAT" (Reckase, 2007). To design such an item pool, an examinee is first randomly sampled from the target examinee population to

take the CAT. Each administered item is simulated to be optimal for this examinee. This procedure is then repeated for the subsequent examinees. Because items created for one examinee can be used for another, the *p*-optimal item pool is the union of the item sets that are administered to each examinee. After the simulation procedure is repeated for a large number of examinees, the number of item in the item pool will eventually approach an upper bound. Thus, the final product of the simulation is an item pool blueprint that tells the pool size and item distribution of the item pool. This blueprint can be directly used as the target for item creation and item pool development.

Therefore, this study aims to generalize the *p*-optimality method (Reckase, 2003 & 2007) to a multidimensional context. Although the generalization seems conceptually straightforward - just implement the simulation procedure based on a MIRT model, to practically implement this procedure, a number of technical challenges need to be solved. For example, the optimal item is unique for each examinee when the unidimensional Rasch model (Lord, 1980) is used. In the multidimensional context, however, the optimal item is not unique, because one optimal item can be found on each direction of measurement. How to select the most appropriate optimal item is the first challenge.

The MCAT in this study is based on the multidimensional Rasch model. The reason for selecting the multidimensional Rasch model is because the idea of *p*-optimal item pool was first proposed based on the unidimensional Rasch model. It is thus straightforward to choose the multidimensional Rasch model when this idea is extended to MCAT for the first time. In this study, a *p*-optimal item pool will be first generated based on the simplest two-dimensional model with simple structure. Then *p*-optimal items pools for MCAT with higher dimensions and with non-simple structure will be generated next.

Specifically, the research questions of this study are:

1. Can the *p*-optimality method be generalized to design item pools for MCAT based on the test design and the examinee population characteristics?
2. How does the performance of a MCAT using the *p*-optimality item pool design method compare with the performance using other item pool designs?
3. How do the characteristics of the *p*-optimal item pool change with exposure control and different test specifications (i.e. the number of dimension, correlation among dimensions, simple structure or not)?

Previous work has suggested that MCATs have great potential, but few studies investigate the item pool design for MCATs. By extending the idea of the *p*-optimal item pool to a multidimensional context, the results from this study could provide a general guideline about the desired characteristics of the *p*-optimal item pool for certain MCATs. More importantly, because the *p*-optimal item pools are specifically tailored to the MCAT programs, the simulation procedures described in this study can be adapted to other MCAT programs with different features and different test purposes. The end product of the *p*-optimal item pool design tells the characteristics of the optimal item pool. If the operational item pool is developed based on the *p*-optimal item pool design, the item pool is expected to ensure the proper functioning of MCATs and to produce reliable measurement outcomes.

Chapter 2 Unidimensional and Multidimensional CAT

This chapter first introduces the computerized adaptive testing (CAT) in Section 2.1. The unidimensional IRT model and the unidimensional CAT are briefly discussed in Section 2.2. The multidimensional IRT model and the multidimensional CAT are explained in Section 2.3.

2.1 Computer Adaptive Testing

CAT is a special form of a computer-delivered test that is adaptive to the examinee's ability level. The “adaptive” means test items are selected on the basis of the examinee's responses to the items previously administrated. One early use of adapting the difficulty of a test to each individual examinee is the Binet-Simon (1905) intelligence test. The items in this test were grouped according to mental age, and the selection of items is determined by the examinee’s mental age estimate, which is derived from the responses to the items administered earlier. From the 1970s, with the development of item response theory and the breakthrough in modern computer technology, the idea of adaptive testing was refined and developed into the current CAT procedures.

For a typical CAT program, the test begins with the first item selected based on an initial estimate of an examinee’s ability level. After each item is administered, a new ability level is estimated and the next item with optimal properties at the new estimate is selected to administer. This process is repeated until it meets certain stopping rules, such as, the precision of proficiency estimate is adequate, or a fixed number of items have been administrated. Therefore, a basic CAT application consists of four major components: an item pool, an item selection procedure, a scoring procedure, and a test stopping rule (Reckase, 1989). In practice, constraints such as

content balancing and exposure control are often imposed on the item selection procedure to ensure the test validity and test security.

2. 2 Unidimensional IRT and CAT

2.2.1 Unidimensional IRT Models

Item response theory (IRT) is a group of mathematical models that describes the relationship between examinee ability and the possibility of answering test items correct. Unidimensional item response theory (UIRT) models assume examinees' responses to test items depend on one single latent trait (Lord, 1980). The item response function (IRF) for the three-parameter logistic (3PL) model (Birnbaum, 1968) is defined by

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{\mathrm{e}^{a_i(\theta_j-b_i)}}{1+\mathrm{e}^{a_i(\theta_j-b_i)}}, \qquad (2.1)$$

where $\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\theta_j,a_i,b_i,c_i\right)$ is the probability of a correct response to item *i* by person *j*; u_{ij} is the response on item *i* by person *j* (1 is correct and 0 is incorrect); θ_j is person *j*'s continuous latent ability; b_i, the item difficulty parameter of item *i*, denotes the inflection point of the IRF; a_i, the discrimination parameter for item *i*, is proportional to the slope of the IRF at its inflection point; c_i, the lower asymptote of the IRF, is the guessing parameter for item *i*.

If the guessing parameter is set to 0 for all the items, the 3PL model becomes a two-parameter logistic (2PL) model specified by the following IRF:

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\theta_j,a_i,b_i\right)=\frac{\mathrm{e}^{a_i(\theta_j-b_i)}}{1+\mathrm{e}^{a_i(\theta_j-b_i)}}, \qquad (2.2)$$

and if the item discrimination parameter is further restricted to be 1 across all the items, the 2PL model results in a Rasch model, which is defined by

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\theta_j,a_i,b_i\right)=\frac{\mathrm{e}^{\theta_j-b_i}}{1+\mathrm{e}^{\theta_j-b_i}}. \qquad (2.3)$$

In IRT, the term "information," also called Fisher information, plays an important role in parameter estimation as it is a statistical indicator of the quality of the estimate of a parameter. The formula for item information can be derived in a number of different ways, but the one provided by Lord (1980) is the most well known. Let $P_i(\theta)$ denote the IRF for item *i*, and let $Q_i(\theta) = 1 - P_i(\theta)$. Then the Fisher information can be obtained by

$$I_i(\theta) = [\frac{\partial P_i(\theta)}{\partial \theta})]^2 / P_i(\theta) Q_i(\theta). \qquad (2.4)$$

When the 3PL model is used, (2.4) becomes

$$I_i(\theta) = a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} [\frac{P_i(\theta) - c_i}{1 - c_i}]^2. \qquad (2.5)$$

When c_i = 0, the information for the 2PL model is

$$I_i(\theta) = a_i^2 P_i(\theta) Q_i(\theta), \qquad (2.6)$$

and when c_i = 0 and a_i = 1, the information for the Rasch model can be simplified to

$$I_i(\theta) = P_i(\theta) Q_i(\theta). \qquad (2.7)$$

IRT models and Fisher information play a central role in CAT, from item calibration to item selection and ability estimation. In Section 2.2.2 and 2.2.3, item selection methods and ability estimation methods for unidimensional CAT (UCAT) will be briefly introduced. The practical constraints for UCAT will be introduced in Section 2.2.4.

2.2.2 Item Selection Methods for UCAT

Items in CAT are selected to be adaptive to the examinee's ability level estimate. The most widely used item selection procedures for UCAT are the maximum Fisher information method (Weiss, 1982), the maximum posterior precision method (Owen， 1975), and the maximum global information method (Chang & Ying, 1996).

The maximum Fisher information method selects the item that provides the maximum amount of Fisher information at examinee's current ability estimate, $\hat{\theta}$. Therefore, the unconstrained Fisher information-based item selection methods administers items that maximize (2.4) at $\theta = \hat{\theta}$.

The maximum posterior precision method is also known as the Owen's Bayesian method. It selects the next item maximizes the expected posterior precision of $\hat{\theta}$. In the early stage of a CAT, the Owen's Bayesian method may select different items from the Fisher information method, because of the effect of the prior. As the test length increases, the effect of the prior decreases and the results from the two methods become similar (Chang & Stout, 1993).

The maximum global information method selects items based on the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), which is a non-symmetric measure of the difference between two probability distributions. In the early stage of a CAT, when the estimated ability is away from the examinee's true ability, the global information method performs better than the Fisher information method with respect to the efficiency and precision of ability estimation (Chang & Ying, 1996; Chen, Ankenmann, & Chang, 2000). After several items are administered and the estimated ability become close to the true ability, the KL divergence effectively reduces to Fisher information.

2.2.3 Ability Estimation Methods for UCAT

In CAT, after each response, the examinee's ability estimate is updated, based on his or her responses to all previous items. The two commonly used estimation procedures are the maximum likelihood method and the Bayesian method (Bejar & Weiss, 1979).

Maximum likelihood estimation (MLE) method is to find an estimate that result in the highest likelihood for the observed string of item responses. The likelihood function is defined as:

$$L(u|\theta) = \prod_{i=1}^{n} P_i(\theta), \qquad (2.8)$$

where $P_i(\theta)$ is the IRF for item i. The highest point on the likelihood function can be located by taking the derivative of (2.8). Iterative numerical methods such as Newton-Raphson method (Wainer, 1990) are often used to solve the derivative equation. MLE ability estimates have desirable properties like asymptotical unbiasedness. However, problems can rise at the early stage of CAT, since MLE cannot provide finite estimates for responses to single items or for patterns of responses that are all correct or all incorrect. To solve the problem, we can either constrain $\hat{\theta}$ to a reasonable range (e.g., -4 to 4) or use alternative estimation methods such as Bayesian procedure.

In Bayesian estimation, by summing the prior distribution, the posterior distribution of θ can be specified based on the Bayes' theorem. The mean of the posterior distribution (refer to as EAP) or the mode of the posterior distribution (refer to as MAP) can be used as the examinee's ability estimate. EAP is more widely used in UCAT because of its stability (Bock & Mislevy, 1982).

2.2.4 Practical Constraints for UCAT

In practice, item selection depending solely on the item selection methods described above might bring concerns about test validity and security. For instance, if a content area requires more instructional time, more items measuring this content area should be administered. Also, if some test items are overexposed and examinees have seen them before taking the test, the validity of the test will be affected. To address these considerations, operational testing programs often impose constraints on item selection process. A brief summary of the content balancing constraint and item exposure control constraint is provided below.

Content balancing procedures ensure each examinee receives approximately the same proportion of items from each content area. The proportion can be determined based on the test specification. Several approaches have been proposed to ensure content balancing in UCAT, such as the weighted deviations model approach (Swanson & Stocking, 1993), the shadow-test approach (van der Linden & Reese, 1998), the modified multinomial model (Chen & Ankenmann, 2004), the maximum priority index method (Cheng & Chang, 2009), and so on. Several research studies (e.g., Cheng & Chang, 2009, Leung, Chang, and Hau, 2003, and van der Linden, 2005) have compared the performance of some of these methods. Generally speaking, the shadow-test approach and the maximum priority index are more flexible in dealing with several constraints, and the weighted deviations model is more widely used in operational testing programs (Buyske, 2005). Detailed descriptions of these content balancing methods can be found in He (2010) and van der Linden (2010).

Item exposure control procedures aim to preventing test items from overexposing to examinees. Numerous item exposure control procedures have been proposed in the last few decades. The most commonly used procedure is the Sympson-Hetter (SH) procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985). This procedure assigns an exposure control parameter to each item based on the frequency of item selections during an iterative CAT simulation. During the test operation, if the exposure control parameter is larger than a random number, the item is administered; otherwise another item is selected and goes through the SH procedure again. Another well known item exposure control procedure is the *a*-stratified procedure proposed by Chang and Ying (1999). This procedure mainly addresses the issues of overdrawing items with high discrimination from item pools. The *a*-stratified procedure first partitions the item pool into several levels according to the *a*-parameter of items. Items with

small *a*-parameter have high priority in the early stage of the test. Items with large *a*-values are saved for the later stage in a CAT administration. The maximum priority index procedure (Cheng & Chang, 2009) used for content balancing also can be used for item exposure control. This procedure adds a weight to item selection method. Items with higher exposure rates are weighted less. This weight index ensures no item is exposed more than a predetermined rate. A detailed summary of the item exposure control procedures described above can be found in Georgiadou, Triantafillou, and Economides (2007).

All the item selection methods, ability estimation methods, and operational constraints heretofore discussed derived to select the appropriate item to administer and pinpoint an examinee's true ability. They are all directly related to the item pool design as the desired item pool should always consist of an appropriate item for every item selection and ability estimation process. In Chapter 3, I will explain how item selection methods, ability estimation methods, and operational constraints determine the item pool design. However, before explaining reasons for inefficiencies, I first describe multidimensional IRT models and multidimensional CAT.

2.3 Multidimensional IRT and CAT

2.3.1 Multidimensional IRT Models

Most operational CAT programs use UIRT models. Nevertheless, the test items in educational and psychology assessments usually measure more than one latent trait so that many researchers have found that examinees often need to use multiple skills to answer test items (Childs & Oppler, 2000; Wu & Adams, 2006; Svetina, 2013). Similar to UIRT, Multidimensional IRT (MIRT) is also a collection of mathematical models that describe the interaction between persons and test items. The difference is that the MIRT models deal with situations when more than one ability are required for test performance (Reckase, 2009).

There are two major types of MIRT models: compensatory and partially-compensatory. The compensatory model is based on a linear combination of ability dimensions, and a high ability on one dimension can compensate for a low ability on another dimension. Sympson (1978), however, argued that the compensatory model is not realistic for certain types of items, because not all skills can compensate each other. Thus he developed a partially-compensatory model to address this issue. Although the partially-compensatory model is more theoretically sound than compensatory models, studies have found compensatory models actually fit real test data better (Ansliey, 1984; Bolt & Lall, 2003). In addition, estimation difficulty for the partially-compensatory model hinders its development and application. As a result, compensatory models are more prevalent in the current literature, and thus will be the only ones focused on in this study.

The compensatory form of the multidimensional three-parameter logistic (M3PL) model is given by Reckase (2009), which is

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i,c_i\right)=c_i+(1-c_i)\frac{\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}{1+\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}, \qquad (2.9)$$

where $\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)$ is the probability of a correct response to item *i* by person *j*; u_{ij} is the response on item *i* by person *j* (1 is correct and 0 is incorrect); $\boldsymbol{\theta_j}$ is a row vector of person *j*'s abilities in a m-dimensional space; $\boldsymbol{a_i}$ is a row vector of the discrimination for item *i*; d_i is a scalar that is related to item difficulty; and c_i is the guessing parameter for item *i*. From equation (2.9), the exponent of *e* is a linear function of θs plus the intercept term *d*, $\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i$. The addition of the θs implies the compensatory nature of the model. If c_i is assumed to be 0 for all the items, the M3PL model becomes the multidimensional two-parameter logistic (M2PL) model, which is defined as

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)=\frac{\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}{1+\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}. \qquad (2.10)$$

The multidimensional extension of the Rasch model was not simply the M2PL model with all the a-parameter set to 1.0, as the relationship between the Rasch model and the 2PL model for the UIRT case. The consequence of setting all the a-parameter to 1.0 is that the $\boldsymbol{a_i}\boldsymbol{\theta_j}' + d_i$ becomes $(\theta_{j1} + \theta_{j2} + \cdots + \theta_{jm}) + d_i$. Therefore, the M2PL model is reduced to a unidimensional Rasch model with $\theta = \theta_{j1} + \theta_{j2} + \cdots + \theta_{jm}$. The multidimensional Rasch model in current literature was proposed by Adams et al. (1997). The model they specified is for the general case that includes both dichotomously and polytomously scored test items. Reckase (2009) provide the dichotomous case of Adam's model, which is

$$\mathrm{P}\left(u_{\mathrm{ij}}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a_i},d_i\right)=\frac{\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}{1+\mathrm{e}^{\boldsymbol{a_i}\boldsymbol{\theta_j}'+d_i}}. \qquad (2.11)$$

Equation (2.10) and (2.11) appear to be identical. The only difference between the two is the way that the $\boldsymbol{a_i}$ vector is specified. In (2.10), $\boldsymbol{a_i}$ is a characteristic of item *i* that is estimated from the data. In (2.11), $\boldsymbol{a_i}$ is a characteristic of item *i* that is specified by the test developer. Adams et al. (1997) specified two variations for the model: between-item and within-item dimensionality. For between-item dimensionality, the $\boldsymbol{a_i}$-vector has elements that are all zeros except for one element. For the two-dimensional case, $\boldsymbol{a_i}$-vector of [1 0] or [0 1] would indicate between-item dimensionality. The vector [1 0] would specify that the item was only affected by ability level on dimension 1 and the vector [0 1] specifies that the item is only affected by ability level on dimension 2. For within-item dimensionality, the $\boldsymbol{a_i}$-vector has more than one nonzero element. A specification for within-item dimensionality might have a vector such as [1 1] indicating that the item is affected equally by both dimensions. In some literature (Reckase,

2009; Segall, 1996; Yao, 2013), the feature of between-item dimensionality is called simple structure, and the within-item dimensionality is called non-simple structure.

In a compensatory MIRT model, in order to make the $\boldsymbol{a_i}$- and d_i- parameter more meaningful, Reckase (1985) and Reckase and Mckinley (1991) developed two statistics to interpret the characteristics of the test items: multidimensional discrimination (MDISC) and multidimensional difficulty (MDIFF). They are defined as:

$$MDISC_i = \sqrt{\boldsymbol{a_i}' \boldsymbol{a_i}}, \qquad (2.12)$$

$$MDIFF_i = \frac{-d_i}{MIDSC_i}, \qquad (2.13)$$

where parameters are defined as before. $MDISC_{\mathrm{i}}$ is the slope of the item response surface at the steepest point, and indicates the discriminating power of the item. $MDIFF_{\mathrm{i}}$ is the distance from the origin to the point of the steepest slope. It represents the multidimensional difficulty of the item: high values indicate difficult items and low values indicate easy items. Thus, the MDISC and the MDIFF value for a MIRT model are analogous to the item discrimination and the item difficulty value for a UIRT model.

The concept of information that is used in UIRT also can be generalized to the multidimensional case. The definition of information for a MIRT model is the same as the definition for a UIRT model, except that information for MIRT is an *m***m* matrix, denoted by $\boldsymbol{I}(\boldsymbol{\theta})$. The {*r*-th, *s*-th} element of this matrix is denoted by $\boldsymbol{I_{rs}}(\boldsymbol{\theta})$. For the M3PL model, the diagonal elements of $\boldsymbol{I}(\boldsymbol{\theta})$ are (Segall, 1996)

$$\boldsymbol{I_{rr}}(\boldsymbol{\theta}) = \sum_{\mathrm{i}\in\mathrm{v}} \frac{a_{ri}^2 Q_{\boldsymbol{i}}(\boldsymbol{\theta})[P_{\boldsymbol{i}}(\boldsymbol{\theta}) - c_i][c_i P_{\boldsymbol{i}}(\boldsymbol{\theta}) - P_{\boldsymbol{i}}^{\boldsymbol{2}}(\boldsymbol{\theta})]}{P_{\boldsymbol{i}}^{\boldsymbol{2}}(\boldsymbol{\theta})(1 - c_i)^{\boldsymbol{2}}}, \qquad (2.14)$$

and the off-diagonal elements are

$$\boldsymbol{I_{rs}(\theta)} = \sum_{\mathrm{i}\in\mathrm{v}} \frac{a_{ri}a_{si}Q_i(\boldsymbol{\theta})[P_{\boldsymbol{i}}(\boldsymbol{\theta}) - c_i][c_iP_{\boldsymbol{i}}(\boldsymbol{\theta}) - P_{\boldsymbol{i}}^{\mathbf{2}}(\boldsymbol{\theta})]}{P_{\boldsymbol{i}}^{\mathbf{2}}(\boldsymbol{\theta})(1 - c_i)^{\mathbf{2}}}, \qquad (2.15)$$

where a_{ri} is the *r*-th element of the $\boldsymbol{a_i}$-vector for item *i*, a_{si} is the *s*-th element, and other symbols are used as previously defined. For the M2PL and the multidimensional Rasch model, the information matrix for item *i* can be simplified to

$$\boldsymbol{I_i(\theta)} = P_{\boldsymbol{i}}(\boldsymbol{\theta})Q_{\boldsymbol{i}}(\boldsymbol{\theta}) \begin{bmatrix} a_{i1}^2 & \cdots & a_{i1}a_{im} \\ \vdots & \ddots & \vdots \\ a_{i1}a_{im} & \cdots & a_{im}^2 \end{bmatrix}. \qquad (2.16)$$

For the multidimensional Rasch model, the $\boldsymbol{a_i a_i}'$ matrix in (2.16) only consists of 0's and 1's.

2.3.2 Generalization of UCAT to MCAT

The merging of MIRT and CAT has been an intriguing direction to explore. When unidimensional algorithms are generalized to multidimensional, we add a huge amount of complexity. Luecht (1996) pointed out that unlike a unidimensional CAT (UCAT), which is merely trying to locate examinees on a latent ability scale, a multidimensional CAT (MCAT) must locate examinees on a plane or a hyperplane and administers items that minimize the joint estimation errors for those ability estimates. Although a MCAT is much more complicated than an UCAT, researchers (e.g., Segall, 1996; Wang & Chen, 2004; Yao, 2012; Mao, Luo & Zhou, 2013) have demonstrated that MCAT is worth the added complications, as MCAT often yields better measurement efficiency than UCAT.

Therefore, to generalize UCAT to MCAT, Reckase (2009) suggested four basic components to be addressed: (1) item pool development, (2) item selection method implementation, (3) examinees' ability estimation, and (4) stopping rule determination. In practical, practical constraints (i.e., content balancing and item exposure control) are also important components for MCAT. Because the desired features of the item pool are dependent on the other four, the

procedures for item selection will be presented first in 2.3.3, followed by the ability estimation method in 2.3.4, stopping rules in 2.3.5, and practical constraints in 2.3.6. The development of multidimensional item pool is described in Chapter 3.

2.3.3 Item Selection Methods for Multidimensional CAT

Item selection is crucial for UCAT as well as for MCAT. If the selected items only provide littler information for ability estimation, the adaptive test will not function well. Like the unidimensional item selection methods, the multidimensional methods are also based on maximizing or minimizing some criterion values at the current ability estimates. The maximum determinant of the Fisher information matrix (D-Optimality) method, Bayesian D-Optimality method, and the maximize KL Information method will be introduced in this section.

The D-Optimality, proposed by Segall (1996) can be considered as the multidimensional extension of the maximum Fisher information method for UCAT. Suppose *k*-1 items have already been administered to an examinee and the *k*-th item is to be determined. The D-Optimality method selects the *k*-th item that maximizes the quantity

$$|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|\,, \qquad (2.17)$$

where $I_{s_{k-1}}(\hat{\boldsymbol{\theta}})$ is the summation of information for the previous *k*-1 items, and $I_{i_k}(\hat{\boldsymbol{\theta}})$ is the item information for the *k*-th item. The item information is defined in (2.14 – 2.16). Note that the || in (2.17) means the determinant of a matrix. The process for selecting the next item is to identify the item that has an item information matrix that, when added to the current test information matrix, will result in the largest value for the determinant of the sum (Reckase, 2009).

Yao (2012) pointed out that the D-Optimality method has an undesirable quality. Towards the beginning of the MCAT, the information matrix may not have full rank, resulting in the

quantity of (2.17) equals to 0. However, this issue can be remedied by applying the Bayesian version of the D-Optimality to the problem of item selection.

The Bayesian D-Optimality method (Segall, 1996) takes a prior distribution into account. It selects the *k*-th item that maximizes

$$|I_{s_{k-1}}(\widehat{\boldsymbol{\theta}}) + I_{i_k}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{\Phi}^{-1}|\,, \qquad (2.18)$$

where $\boldsymbol{\Phi}^{-1}$ is the prior distribution, which is the inverse of $\boldsymbol{\Phi}$, and $\boldsymbol{\Phi}$ is the variance-covariance matrix of the examinees' multidimensional ability. For the first few items, the Bayesian D-Optimality method is expected to select different items compared with the D-Optimality method, but as the test length increase, the two methods should become similar.

The maximum KL information method for MCAT was first presented by Veldkamp and van der Linden (2002). This method is an extension of the Chang and Ying (1996) for a UCAT to solve the issue of selecting proper items when ability is poorly estimated in the early stage of the UCAT. When only one item is considered, the KL information is given by

$$K_i(\boldsymbol{\theta}, \boldsymbol{\theta_0}) = ln\left[\frac{P_i(\boldsymbol{\theta_0})}{P_i(\boldsymbol{\theta})}\right] + \big(1 - P_i(\boldsymbol{\theta_0})\big) ln\left[\frac{P_i(\boldsymbol{\theta_0})}{P_i(\boldsymbol{\theta})}\right]. \qquad (2.19)$$

The item selection rule presented by Veldkamp and van der Linden (2002) is to select the item that maximizes

$$K_i^B\big(\widehat{\boldsymbol{\theta}}^{k-1}\big) = \int K_i\big(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^{k-1}\big)\, f(\boldsymbol{\theta}|u_1, \dots, u_{k-1}) \partial\boldsymbol{\theta}, \qquad (2.20)$$

where $K_i^B\big(\widehat{\boldsymbol{\theta}}^{k-1}\big)$ is the Bayesian posterior expected information after *k*-1 items, and $f(\boldsymbol{\theta}|u_1, \dots, u_{k-1})$ is the posterior density after *k*-1 items. The implementation of the maximum KL information method requires very long CPU time because of the calculation for two integrals. The first integral is the estimation of the $f(\boldsymbol{\theta}|u_1, \dots, u_{k-1})$, and the second one is shown in (2.20). For this reason, the item selection based on KL information is not considered in this study.

2.3.4 Ability Estimation Methods for Multidimensional CAT

The ultimate goal for most MCAT is to estimate the multidimensional ability for examinees. Assume an item has been selected using one of the item selection methods described in Section 2.3.3, and an examinee has provided a response for this item. An ability estimation method is then used to update the estimate of the examinee ability. The two general classes of ability estimation methods for MCAT are: maximum likelihood and Bayesian. These two methods are described in this section.

For the maximum likelihood estimation (MLE) method (Segall, 1996), MIRT ability is estimated by finding the mode $\widehat{\boldsymbol{\theta}}$ that maximize the likelihood function $L(\boldsymbol{u}|\boldsymbol{\theta})$, i.e.,

$$\frac{\partial}{\partial \boldsymbol{\theta}} ln\, L(\boldsymbol{u}|\boldsymbol{\theta}) = 0. \qquad (2.21)$$

Using Newton-Raphson method, suppose $\boldsymbol{\theta}$ is the approximation that maximize $ln\, L(\boldsymbol{u}|\boldsymbol{\theta})$, then

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{\delta}^{(j)}, \qquad (2.22)$$

where $\boldsymbol{\delta}^{(j)}$ is the m*1 vector defined as

$$\boldsymbol{\delta}^{(j)} = \left[H(\boldsymbol{\theta}^{(j)})\right]^{-1} * \frac{\partial}{\partial \boldsymbol{\theta}} ln\, L\left(\boldsymbol{u}\middle|\boldsymbol{\theta}^{(j)}\right). \qquad (2.23)$$

The $H(\boldsymbol{\theta}^{(j)})$ in (2. 23) is a *m*m* matrix of second derivatives evaluated at $\boldsymbol{\theta}^{(j)}$. The diagonal elements of $H(\boldsymbol{\theta})$ take the form

$$\boldsymbol{H_{rr}}(\boldsymbol{\theta}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri}^2 Q_{\boldsymbol{i}}(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^{\mathbf{2}}}, \qquad (2.24)$$

and the off-diagonal elements are of the form

$$\boldsymbol{H_{rs}}(\boldsymbol{\theta}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri} a_{si} Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1 - c_i)^{\mathbf{2}}}, \qquad (2.25)$$

The $\frac{\partial}{\partial \boldsymbol{\theta}} ln\, L\left(\boldsymbol{u}\middle|\boldsymbol{\theta}^{(j)}\right)$ in (2.23) is a *m**1 vector of partial derivatives of $ln\, L\left(\boldsymbol{u}\middle|\boldsymbol{\theta}^{(j)}\right)$ with the *r*-th element defined as

$$\frac{\partial}{\partial \theta_r} \ln L(\boldsymbol{u}|\boldsymbol{\theta}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri}[P_i(\boldsymbol{\theta}) - c_i][u_i - P_i(\boldsymbol{\theta})]}{(1 - c_i)P_{\boldsymbol{i}}(\boldsymbol{\theta})}, \qquad (2.26)$$

With all the terms in (2.23) defined, the Newton-Raphson method can be used to obtain $\boldsymbol{\theta}^{(j+1)}$ repeatedly until $\boldsymbol{\delta}^{(j)}$ becomes sufficiently small. Similar to the unidimensional MLE, the multidimensional MLE also has the issue of infinite estimates in the early stage of the MCAT (Diao, 2009; Reckase, 2009). Reckase (2009) also pointed out that the minimum number of test items needed to get finite estimates for a three dimensional MCAT is three, but the actual number in the MCAT can be larger than that. To overcome this problem, a Bayesian procedure can be considered.

The Bayesian method (Segall, 1996) is similar to the MLE method, except that the likelihood function is the product of the likelihood and the prior:

$$f(\boldsymbol{\theta}|\boldsymbol{u}) = L(\boldsymbol{u}|\boldsymbol{\theta})\frac{f(\boldsymbol{\theta})}{f(\boldsymbol{u})}, \qquad (2.27)$$

where the $L(\boldsymbol{u}|\boldsymbol{\theta})$ is the likelihood function, $f(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, wand $f(\boldsymbol{u})$ is the marginal probability of $\boldsymbol{u}$. Segall (1996) defined the $f(\boldsymbol{\theta})$ as a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Phi}$. Because Yao (2012) found that the mode of the posterior distribution (known as MAP) yields better precision and requires less computation time than does the expectation of the posterior (known as EAP), only the MAP procedure is described in this study. The mode of the posterior distribution can be obtained by maximizing the natural logarithm of the posterior distribution, i.e.,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}|\boldsymbol{u}) = \mathbf{0}, \qquad (2.28)$$

where the $f(\boldsymbol{\theta}|\boldsymbol{u})$ is defined in (2.27). Because the (2.28) formula has no explicit solution, an iterative numerical procedure such as the Newton-Raphson procedure must be used. Suppose $\boldsymbol{\theta}$ is the approximation that maximizes $ln\, f(\boldsymbol{\theta}|\boldsymbol{u})$, then

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \boldsymbol{\delta}^{(j)}, \qquad (2.29)$$

where $\boldsymbol{\delta}^{(j)}$ is the m*1 vector defined as

$$\boldsymbol{\delta}^{(j)} = \left[J(\boldsymbol{\theta}^{(j)})\right]^{-1} * \frac{\partial}{\partial \boldsymbol{\theta}} ln\, f\left(\boldsymbol{\theta}^{(j)}\middle|\boldsymbol{u}\right). \qquad (2.30)$$

The $J(\boldsymbol{\theta}^{(j)})$ in (2.30) is a *m***m* matrix of second derivatives evaluated at $\boldsymbol{\theta}^{(j)}$. The diagonal elements of $J(\boldsymbol{\theta})$ take the form

$$\boldsymbol{J_{rr}}(\boldsymbol{\theta}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri}^2 Q_{\boldsymbol{i}}(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^{\mathbf{2}}} - \phi^{rr}, \qquad (2.31)$$

where ϕ^{rr} is the *r*-th diagonal element of $\boldsymbol{\Phi^{-1}}$. The off-diagonal elements of $J(\boldsymbol{\theta})$ are of the form

$$\boldsymbol{J_{rs}}(\boldsymbol{\theta}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri} a_{si} Q_i(\boldsymbol{\theta})[P_i(\boldsymbol{\theta}) - c_i][c_i u_i - P_i^2(\boldsymbol{\theta})]}{P_i^2(\boldsymbol{\theta})(1-c_i)^{\mathbf{2}}} - \phi^{rs}, \qquad (2.32)$$

where ϕ^{rs} is the {*r*-th, *s*-th} element of $\boldsymbol{\Phi^{-1}}$. The $\frac{\partial}{\partial \boldsymbol{\theta}} ln\, f\left(\boldsymbol{\theta}^{(j)}\middle|\boldsymbol{u}\right)$ in (2. 30) is a *m**1 vector of partial derivatives of $ln\, f\left(\boldsymbol{\theta}^{(j)}\middle|\boldsymbol{u}\right)$ with the *r*-th element defined as

$$\frac{\partial}{\partial \theta_r} \ln f(\boldsymbol{\theta}|\boldsymbol{u}) = \sum\nolimits_{\mathrm{i}\in \mathrm{v}} \frac{a_{ri}[P_i(\boldsymbol{\theta}) - c_i][u_i - P_i(\boldsymbol{\theta})]}{(1-c_i)P_{\boldsymbol{i}}(\boldsymbol{\theta})} - [\frac{\partial}{\partial \theta_r}(\boldsymbol{\theta} - \boldsymbol{\mu})']\boldsymbol{\Phi^{-1}}(\boldsymbol{\theta} - \boldsymbol{\mu}). \qquad (2.33)$$

The $\frac{\partial}{\partial \theta_r}(\boldsymbol{\theta} - \boldsymbol{\mu})'$ in (2. 33) denotes as a 1**m* vector with the *r*-th element equal to 1 and all other elements equal to 0. With all the terms in (2.29) defined, the $\boldsymbol{\theta}^{(j+1)}$ can be obtained repeatedly until $\boldsymbol{\delta}^{(j)}$ become sufficiently small.

2.3.5 Stopping Rules for MCAT

The stopping rules for an UCAT fall in to two groups: fixed length and variable length. The fixed length stopping rule is very easy to adapt to a MCAT. For the fixed length rule, the total number of items to be administered to each examinee is pre-determined based on the purposes of the test and practical considerations. When the number of items is reached in the test administration, the CAT will stop and the final ability estimate is computed. Because the fixed stopping rule is easy to implement, most MCATs in the research literature use the fixed length to as their stopping rules (e.g., Diao, 2009; Segall, 1996; Wang & Chang, 2011; Yao, 2012).

Variable length CAT controls the test length using a statistical criterion. For example, in a UCAT, if the standard error of measurement for θ estimate is smaller than a critical value, the test stops and the final θ estimate is reported. Therefore, variable length CAT administers different number of items to different examinees. Yao (2013) proposed two stopping rules for MCAT, the standard error (SE) and predicted standard error (PSE). The results showed that the PSE yields slightly worse results than the SE, but with fewer items. The detailed description for these two methods can be obtained from Yao's paper.

2.3.6 Practical Constraints for MCAT

Content balancing and item exposure control are as important to MCAT as to UCAT. Among the numerous content balancing methods for UCAT, the shadow test approach is the first one that has been successfully implemented in MCAT by Veldkamp and van der Linden (2002). Because the shadow test approach requires an existed master pool, it is not applicable in this study. The Maximum Priority Index (MPI) method is another content balancing that has been implemented in MCAT by Frey, Cheng, & Seitz, (2011), and also been used in Yao (2012) and Yao (2013). According to Yao (2012), the MPI index for item *i* is defined by

$$PI_i = \prod_{l=1}^{D} f_{il}^{c_{il}}, \qquad (2.34)$$

where the constraint matrix $C = (c_{il})_{J*D}$, indicating the loading information for item i on dimension l. If item i loads on dimension l, $c_{il} = 1$; otherwise $c_{il} = 0$. Suppose the percentage of items in each content area is fixed. Then the f_{il} is defined by

$$f_{il} = \frac{(X_l - x_l)}{X_l}, \qquad (2.35)$$

where X_l is the number of items that should be administered from dimension l, and so far x_l such items have been selected. At the beginning, f_{il} is 1 when no item has been selected from dimension l, and it gets smaller as x_l increases. When $x_l = X_l$, $f_{il} = 0$; no more items will be selected from this dimension. The MPI is implemented by multiplying the PI_i to the item selection criteria. For example, for the D-Optimality method, item $i = k$ is selected if $|I_{s_{k-1}}(\widehat{\boldsymbol{\theta}}) + I_{i_k}(\widehat{\boldsymbol{\theta}})| * PI_k$ has a maximum value among all the items in the item pool.

The MPI method has also been used for item exposure control in MCAT in Yao (2012 & 2013). Suppose the maximum exposure rate of item i is fixed to R_i. For each selection step, let n_i be the number of examinees that have already selected item i. Then the index for the item exposure control is defined by

$$f_{il} = \frac{(R_i - n_i/N)}{R_i}, \qquad (2.36)$$

where N is the total number of examinees, and n_i/N is the actual exposure rate for item i. This index makes sure that no item is selected with exposure rate larger the predefined rate, R_i. To implement the MPI for item exposure control, Yao (2012 & 2013) multiplied the f_{il} in (2.36) to the item selection criterion. The results shows the MPI can effectively control the item exposure rate, and increase the item pool usage to 100% when several item selection methods are used,

including the D-Optimality method. Although the 100% item pool usage is desirable, the number is inflated due to the misuse of the MPI. For a two dimensional CAT, after 20 items has been administered, the $|I_{s_{20}}(\hat{\boldsymbol{\theta}}) + I_{i_{21}}(\hat{\boldsymbol{\theta}})|$ value for all items in the item pool ranges from 3.39 to 3.82. If the f_{il} in (2.36) is smaller than 0.88 ($f_{il} = .88$ implies $n_i/N = .12R_i$, which is much smaller than R_i), the value of the $\left|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})\right| * f_{il}$ for the item associated with the largest $|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|$ value will be smaller than 3.82. That is for say, if f_{il} is multiply to the selection criterion of the D-Optimality method, the best item available in the pool will not be selected, even though its actual exposure rate is much smaller than the maximum rate, R_i.

The reason why the MPI method functions properly in UCAT but not in MCAT is the difference in the item selection criterion. The minimum value of the Fisher information is close to 0, but the minimum value of the $|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|$ is not 0. This issue can be solved by rescaling the item selection criterion, and then multiplying f_{il} to the rescaled criterion, instead to the criterion itself.

In this study, a non-linear method is used to rescale the criterion of the D-Optimality method. First, a percentile rank is calculated for all the items available in the item pools, that is

$$|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|^{Re} = \text{Percentile}\left(\left|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})\right|\right), \qquad (2.37)$$

where the $|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|^{Re}$ denotes the rescaled criterion. Second, select the item with a maximum value of $|I_{s_{k-1}}(\hat{\boldsymbol{\theta}}) + I_{i_k}(\hat{\boldsymbol{\theta}})|^{Re} * f_{il}$. This item exposure control procedure is referred to as the "Modified MPI" in this study.

Yao (2012) also used the Sympson–Hetter (SH) method for item exposure control in a MCAT, but she didn't recommend it as it is very time consuming to create the "exposure-control table", and the computation time increases exponentially with the number of dimensions. Therefore, the SH method is not considered in this study for item exposure control.

Chapter 3 *p*-Optimality Method and the Extension to MCAT

This chapter first introduces the concept of a *p*-optimal item pool in Section 3.1. Section 3.2 then presents the *p*-optimality method for describing an item pool and its application to item pool design using the unidimensional Rasch Model. Finally, the extension of the method to the MCAT item pool design based on the multidimensional Rasch model is discussed in Section 3.3 in detail.

3.1 From Optimal Item Pool to *p*-Optimal Item Pool

Before introducing the details about the *p*-optimal item pool, it is important to define the optimal item pool first. Reckase (2010) defined the best possible, or optimal, item pool as that, whenever the CAT item selection algorithm is searching for a test item to administer, exactly the item that is desired is available in the item pool. If a desired item is always available for every item selection, than the item pool can be considered to be optimal.

Suppose that a fixed length UCAT is based on the unidimensional Rasch model, and uses maximum Fisher information for item selection. For this type of UCAT, the maximum Fisher information method selects items with the difficulty parameter, b_i, exactly equal to the current ability estimate $\hat{\theta}$. This is because the information function for the unidimensional Rasch model, which is $I_i(\theta) = P_i(\theta)Q_i(\theta)$, reaches its maximum value of 0.25 when $b_i = \theta_j$.

An optimal item pool for this CAT procedure is the one that always has an item in the pool with *b*-parameter exactly equal to $\hat{\theta}$ for every item selection process for every examinee. Because θ is a continuous variable that has infinite number of values on the θ scale, if items in the item pool exactly match all the $\hat{\theta}$, the item pool has to consist of infinite number of items.

To make the concept of the optimal item pool realistic for practical item pool design, a *p*-optimal item pool (Reckase, 2010) was introduced to approximate an optimal pool of smaller size with little loss of specified characteristics (i.e., item information). Reckase (2010) referred the *p*-optimal item pool design method as the *p*-optimality method. Reckase (2010) also defined the *p*-optimal item pool as an item pool "that always has an item available for selection that *p*% matches the desired characteristics specified by the item selection routine for the CAT." The implementation of the *p*-optimality method in UCAT based on the unidimensional Rasch model is described below.

3.2 *p*-Optimal Item pool Design for UCAT

For the UCAT described above, a *p*-optimal item pool will always has an available item that can provide at least *p*% of the maximum Fisher information at the current θ estimate. Figure 3.1 shows the Fisher information function for a test item based on the unidimensional Rasch model. The horizontal scale is $\theta - b$ so that the results can generalize to all the values of θ. The information reaches the maximum value when $\theta - b = 0$, that is $\theta = b$. Instead of requiring items with maximum information always available in the item pool, it might be acceptable to relax the criterion to at least 90% maximum information. That is, instead of needing items with $b = \hat{\theta}$, an item with b-parameter .65 unit away from $\hat{\theta}$ also meets the criterion (see Figure 3.1).

Therefore, if an item pool meets the criterion of always having an available item with b-parameter .65-unit away from $\hat{\theta}$, the item pool can be said to be .9-optimal, because the available item can provide at least 90% of the maximum possible information for ability estimation. This way of describing the design of an item pool is called *p*-optimal for proportion of maximum optimality (Reckase, 2010).

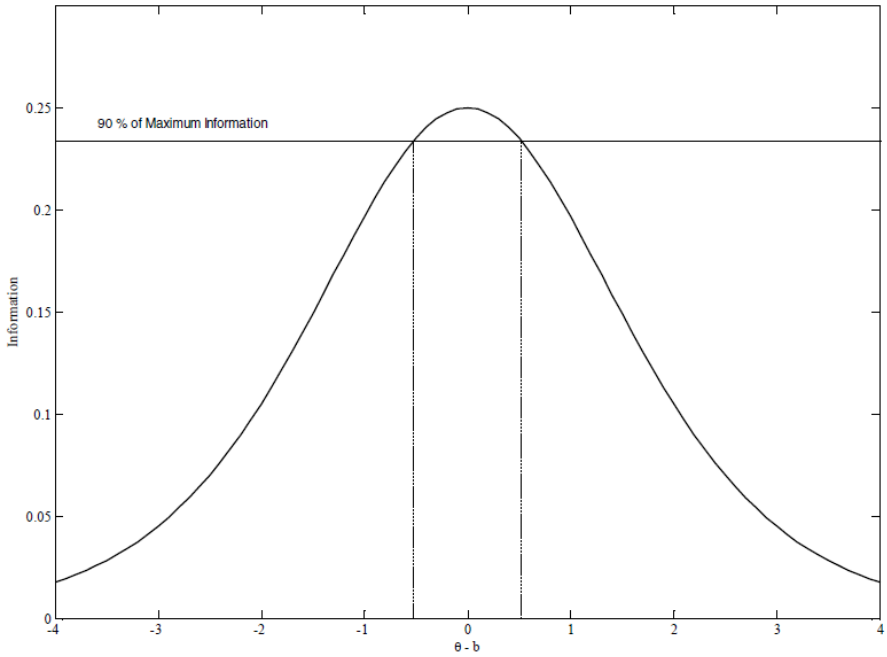


Figure 3.1: Information Function for a Test Item Fit by the Unidimensional Rasch Model

Such a *p*-optimal pool is designed by the following steps:

1) Specify the characteristics of a CAT program, such as the IRT model, test length, item selection method, ability estimation method, and stopping rule. In the example here, the UCAT is based on the unidimensional Rasch model, selects items using the maximum Fisher information method, estimates ability by the MLE, and with test length fixed at 30-item.
2) Randomly sample an examinee from the target examinee population and generate the first optimal item. The optimal item is an item with *b*-parameter equal to the initial value of $\hat{\theta}$ for this examinee.
3) Generate a response to this item based on this examinee's true θ. A random number is first generated from the Uniform(0,1) distribution. If the random number is greater than the probability of this examinee answering this item correct, a correct response is assigned to this examinee; otherwise, an incorrect response is assigned.

4) Update the $\hat{\theta}$ using the MLE method based on the response generated in step 3.
5) Generate the next optimal item with *b*-parameter equal to the updated $\hat{\theta}$.
6) Repeat the process of generating response, ability estimation, and optimal item generation until the stopping rule is satisfied.
7) Classify all the generated optimal items into "item bins". Item bins are defined as intervals on the *b*-parameter scale. For a .9-optimal pool, the criterion is that the *b*-parameter is within .65-unit distance away from $\hat{\theta}$. To meet this criterion, the width of the item bin should be set to .65. In this case, the first item bin is centered on the zero point and ranges from -.325 to .325. The rest of the item bins can be determined by stepping off in either direction.
8) Document the number of items in each item bin for this examinee.
9) Repeat steps 2 to 8 for another examinee. The union of the number of items in each bin forms the *p*-optimal pool for these two examinees (see Table 3.1). Union, instead of summation, is considered because the items used for the first examinee can be used for the second one.
10) Repeat this process for a large number of examinee. The union of items across all the examinees is the end product of the *p*-optimal pool design.

The end product of the *p*-optimal item pool design is a bin-count table, which tells the number of items in each item bin. This bin-count table can be used as the guidance for item creation. If items can be created to match the bin-count table, the item pool is deemed to be *p*-optimal. A more detailed description of this method can be found in Reckase (2010).

Table 3.1: The p-optimal pool for two examinees

Item bin	-3	-2.4	-1.8	-1.2	-0.6	0	0.6	1.2	1.8	2.4	3
Examinee 1	0	0	10	13	7	0	0	0	0	0	0
Examinee 2	0	0	0	9	15	6	0	0	0	0	0
Union	**0**	**0**	**10**	**13**	**15**	**6**	**0**	**0**	**0**	**0**	**0**

Note: the values on the first row represent the central point of each item bin;
the values on the second and third row represent the number of items in each item bin.

3.3 Extending the *p*-Optimality Method to MCAT

As discussed in Chapter 2, the desired features of an item pool depend on the item selection method, ability estimation method, stopping rule, as well as constraints such as content balancing and item exposure control. The *p*-optimality method for item pool design described above also depends on the selection of these methods. Therefore, the first step of extending the *p*-optimality method to MCAT is to determine the characteristics of the MCAT program. The psychometric model, item selection method, ability estimation method, stopping rule, and constraints for the MCAT considered in this study are defined below.

First, the multidimensional Rasch model defined by equation (2.11) is served as the psychometric model for the MCAT in this study. There are two reasons of choosing the multidimensional Rasch model. The first one is because the idea of *p*-optimal item pool design was proposed for a UCAT based on the unidimensional Rasch model. It is thus straightforward to choose the multidimensional Rasch model when this idea is extended to MCAT for the first time. The second reason is that the multidimensional Rasch model is relatively simple compared with the M2PL and the M3PL model defined by (2.09) and (2.10), respectively. Because the $\boldsymbol{a}$-parameter is fixed in the multidimensional Rasch model, the determination of the optimal item is much easier than the situation of unfixed $\boldsymbol{a}$-parameter (Gu, 2007). Given these two reasons, this

study only focuses on the *p*-optimal item pool based on the multidimensional Rasch model. Future studies can extend this method to other complex MIRT models.

Second, the D-optimality method (Segall, 1996) is used to select items in this study. The D-optimality can be considered as the multidimensional extension of the maximum Fisher information for UCAT; hence, the method of optimal item generation can be extended to the multidimensional context in a fairly straightforward fashion. Therefore, the *p*-optimal item pool design in this study is based on the D-optimality item selection method only.

Third, the Bayesian MAP (Segall, 1996) is the ability estimation method for the MCAT in this study. The Bayesian MAP is used here because Yao (2013) mentioned, in one of her unpublished manuscripts, the Bayesian MAP yields better precision than does the MLE and perform similarly or better than the Bayesian EAP. Also, because the Bayesian MAP solves the issues of infinite ability estimates in early MCAT, the Bayesian MAP method is adopted for the *p*-optimal item pool design in this study.

Fourth, the stopping rule in this study is the fixed length rule. The variable length stopping rule is not considered here for two reasons. First, Reckase (2010) has demonstrated the *p*-optimal item pool design for a fixed length CAT can be easily modified to be used in a variable length CAT. There is no need to describe both of them in this study. The second reason, again, is because the fixed length rule is relatively easy to be built into the *p*-optimal item pool design procedure.

Fifth, the content balancing constraint is not implemented in this study. The reason is that the D-optimality item selection method can balance the number of items administered from each dimension. In a two dimensional MCAT, for example, if more items are selected from Dimension 1, there will be more information on the direction of θ_1 and less information on the

direction of θ_2. Then the D-optimality method will select the next item from Dimension 2 until there is more information on the direction of θ_2. Therefore, when the test is completed, the number of items from each dimension is expected to be very similar, even though no content balancing is implemented. For some operational testing programs, the number of items for each content area is set to be different because some content area may require more instructional time. In this situation, the content balancing is necessary and can be built into the *p*-optimal item pool design procedure. This situation, however, is not considered in this study.

Sixth, the *p*-optimal item pool design with and without item exposure control is compared in this study, to answer the third research question for this study. The Modified Maximum Priority Index described in Chapter 2 is used for item exposure control.

In the following sections, the *p*-optimal item pool design for MCAT is first demonstrated on the simplest case: a MCAT measuring a two-dimensional ability,(θ_1, θ_2), using items fit by the two-dimensional Rasch model with simple structure, and without item exposure control. For this specific MCAT, there are only two clusters of items in the item pool. Items in Cluster 1 only measure θ_1 with $\boldsymbol{a}_i = (1,0)$. Items in Cluster 2 only measure θ_2 with $\boldsymbol{a}_i = (0,1)$. According to equation (2.11), the two-dimensional Rasch model can also be specified as:

$$P(\boldsymbol{\theta}) = \frac{\mathrm{e}^{a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i}}{1 + \mathrm{e}^{a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i}}. \qquad (3.1)$$

The feature of simple structure can simplify (3.1) into

$$P(\boldsymbol{\theta}) = \begin{cases} P_1(\boldsymbol{\theta}) = \dfrac{\mathrm{e}^{\theta_{1j} + d_i}}{1 + \mathrm{e}^{\theta_{1j} + d_i}}, \text{for items from Cluster 1} \\ P_2(\boldsymbol{\theta}) = \dfrac{\mathrm{e}^{\theta_{2j} + d_i}}{1 + \mathrm{e}^{\theta_{2j} + d_i}}, \text{for items from Cluster 2} \end{cases}. \qquad (3.2)$$

The method of optimal item generation, the extension of item bins, and the interpretation of *p*-optimal item pool for this specific MCAT are demonstrated in Section 3.3.1 to 3.3.3. An

example of the *p*-optimal item pool design for this MCAT is presented in Section 3.3.4. The *p*-optimal item pool design for MCAT with exposure control is introduced in Section 3.3.5.

3.3.1 Optimal Item Generation

For the UCAT, the optimal item is the one that maximizes the information function at the current $\hat{\theta}$. For the MCAT described above, according to equation (2.17), the *k*-th optimal item is the one that maximizes the quantity

$$|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|. \qquad (3.3)$$

where $I_{s_{k-1}}(\hat{\theta})$ is the summation of the information for the *k*-1 items that has been administered, denoted as

$$I_{s_{k-1}}(\hat{\theta}) = I_{i_1}(\hat{\theta}) + I_{i_2}(\hat{\theta}) + \cdots + I_{i_{k-1}}(\hat{\theta}); \qquad (3.4)$$

and $I_{i_k}(\hat{\theta})$ is the information function for *k*-th item that is going to be administered. According to equation (2.16), $I_{i_k}(\hat{\theta})$ for the two-dimensional multidimensional Rasch model with simple structure can be specified into

$$I_{i_k}(\hat{\theta}) = \begin{cases} P_1Q_1\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}P_1Q_1 & 0\\ 0 & 0\end{bmatrix}, \text{for items from Cluster 1} \\ P_2Q_2\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & P_2Q_2\end{bmatrix}, \text{for items from Cluster 2} \end{cases}, \qquad (3.5)$$

where P_1 and P_2 are defined in equation (3.2) and $Q_1 = 1 - P_1, Q_2 = 1 - P_2$.

Suppose among the *k*-1 administered items, k_1 of them are from Cluster 1 and k_2 of them are from Cluster 2, where $k_1 + k_2 = k - 1$. Substituting (3.5) in to (3.4), we obtain

$$I_{s_{k-1}}(\hat{\theta}) = \begin{bmatrix}\sum_{i=1}^{k_1} P_{1i}Q_{1i} & 0\\ 0 & \sum_{i=1}^{k_2} P_{2i}Q_{2i}\end{bmatrix}. \qquad (3.6)$$

Again, substituting (3.5) and (3.6) into (3.3), we obtain

$$|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$$

$$= \begin{cases} \left| \begin{bmatrix} \sum_{i=1}^{k_1} P_{1i}Q_{1i} & 0 \\ 0 & \sum_{i=1}^{k_2} P_{2i}Q_{2i} \end{bmatrix} + \begin{bmatrix} P_{1i_k}Q_{1i_k} & 0 \\ 0 & 0 \end{bmatrix} \right|, \text{if the kth item is from Cluster 1} \\ \left| \begin{bmatrix} \sum_{i=1}^{k_1} P_{1i}Q_{1i} & 0 \\ 0 & \sum_{i=1}^{k_2} P_{2i}Q_{2i} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & P_{2i_k}Q_{2i_k} \end{bmatrix} \right|, \text{if the kth item is from Cluster 2} \end{cases} . \quad (3.7)$$

By solving the determent, (3.7) becomes

$$|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$$

$$= \begin{cases} \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) + \left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)P_{1i_k}Q_{1i_k}, \text{if the kth item is from Cluster 1} \\ \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) + \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)P_{2i_k}Q_{2i_k}, \text{if the kth item from in Cluster 2} \end{cases} .$$

(3.8)

Because $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)$ and $\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)$ are constant across all the potential *k*-th item, maximizing $|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$ in is equivalent to maximizing $P_{1i_k}Q_{1i_k}$ or $P_{2i_k}Q_{2i_k}$. Based on the two-dimensional Rasch model with simple structure defined in equation (3.2), $P_{1i_k}Q_{1i_k}$ will be maximized when $P_{1i_k} = 0.5$, or $-d_i = \theta_{1j}$. Similarly, $P_{2i_k}Q_{2i_k}$ will be maximized when $-d_i = \theta_{2j}$.

Therefore, the optimal item for the *k*-th item is either the one from Cluster 1 with $-d_i = \theta_{1j}$ or the one from Cluster 2 with $-d_i = \theta_{2j}$. To determine which one is the true optimal, it only needs to compare $\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)$ with $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)$, because the first term in equation (3.8) is the same. If the following inequality holds

$$\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) > \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right), \quad (3.9)$$

the optimal item is from Cluster 1 with $\boldsymbol{a}_i = (1,0)$ and $-d_i = \theta_{1j}$. If this inequality holds

$$\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) < \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right), \qquad (3.10)$$

the optimal item is from Cluster 2 with $\boldsymbol{a}_i = (0,1)$ and $-d_i = \theta_{2j}$. If the two terms are the same, the optimal item is randomly picked.

In other words, after *k*-1 items are administered, if the test information on the direction of Dimension 1 is smaller, the *k*-th optimal item is an item measuring Dimension 1 with $-d_i = \hat{\theta}_1$. If the test information on the direction of Dimension 2 is smaller, the *k*-th optimal item should measure Dimension 2 with $-d_i = \hat{\theta}_2$. The information from previous administered item determines which cluster the optimal item is from, and the current $\boldsymbol{\theta}$ estimates determines the *d*-parameter for the optimal item.

3.3.2 Interpretation for the "p-Optimal"

For a unidimensional .*9*-optimal item pool, items that can provide at least 90% of the maximum possible Fisher information are always available for selection. For the unidimensional Rasch model, because Fisher information is *P*Q*, the "*.9*-optimal" means the selected item yield at least 90% of the maximum possible value of *P*Q*.

For the MCAT in this study, the item selection method is the D-optimality. Suppose the D-optimality method selects the *k*-th optimal item from Cluster 1. This item should have a maximum value of $|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})| = \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) + \left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)P_{1i_k}Q_{1i_k}$, compared with other items. By bringing the $\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)$ to the front, $|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$ becomes

$$|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})| = \left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)\left[\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right) + P_{1i_k}Q_{1i_k}\right]. \qquad (3.11)$$

Here, because the maximum determinant is not simply the *P*Q*, the 90% of the maximum value of $P_{1i_k}Q_{1i_k}$ is no longer equivalent to 90% of the maximum determinant. Therefore, the "*.9*-optimal" item pool no longer implies items that are at least 90% of the maximum determinant of the $I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})$ are always available.

In fact, the $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)$ in (3.11) is the sum of the information for all the administered items from Cluster 1 on the direction of θ_1, and the $P_{1i_k}Q_{1i_k}$ is the information for the *k*-th item on the direction of θ_1. The same interpretation can be made for items in Cluster 2. As mentioned above, the D-optimality method selects the item that adds the maximum information on the current test information on the direction of minimum information. If the item pool is *.9*-optimal, the selected item would be the one that adds at least 90% of maximum possible information on the current test information on the direction of minimum information. Therefore, the interpretation for the "*p*-optimal" item pool in MCAT is that items that can add at least *p*-proportion of maximum possible information on the current test information on the direction of minimum information are always available in the item pool.

3.3.3 Extending the "bin" concept

In UCAT, item bins are created by dividing the scale of the *b*-parameter into several intervals. These item bins are referred to as "*b*-bin," since they are defined on the *b*-parameter scale. As mentioned in Chapter 2, the *d*-parameter in an MIRT model is an intercept term that is related to both item difficulty and item discrimination. The item difficulty in MIRT models is the *MDIFF*. The value of *MDIFF* has the same interpretation as the *b*-parameter for UIRT models. Therefore, the "*MDIFF*-bin", instead of "*d*-bin," is used for the optimal item pool design.

For the two dimensional Rasch model with simple structure defined in (3.2), the item response function (IRF) for items from Cluster 1 is the same as the IRF for the unidimensional

Rasch model with $\theta = \theta_1$. Similarly, the IRF for items in Cluster 2 is the same as the IRF for the unidimensional Rasch model with $\theta = \theta_2$. Therefore, Figure 3.1 also can be used here to determine the size for the *MDIFF*-bin. For *.9*-optimal item pool, if an item from Cluster 1 is selected, the *d*-parameter of this item should be within .65-unit distance away from the current estimate of $-\theta_1$. If an item from Cluster 2 is selected, the *d*-parameter of this item should be within .65-unit distance away from $-\theta_2$. Therefore, the width of the interval on the *d*-parameter scale is .65. Because *MDIFF* is equal to $-d_i$ for the two dimensional Rasch model with simple structure, the size for the *MDIFF*-bin is also .65. In the case of *.86*-optimal, the interval on the *d*-parameter scale is 0.8 so that the size for the *MDIFF*-bin is 0.8. In the case of *.96*-optimal, the size for the *MDIFF*-bin is 0.4.

3.3.4 An example of the p-optimal item pool design for MCAT

For the MCAT described above, items are fitted by the two-dimensional Rasch model with simple structure, and the test length is fixed at 30. Suppose two examinees have taken this MCAT and their true abilities are (0.7, 1.5) and (-1.1, -1.0), respectively. For each examinee, the first item randomly chosen from either Cluster 1 with $\boldsymbol{a}_i = (1,0)$ and $-d_i$ exactly equal to the starting value of θ_1, or from Cluster 2 with $\boldsymbol{a}_i = (0,1)$ and $-d_i$ exactly equal to the starting value of θ_2. Then a response to this item is generated using the two-dimensional Rasch model and the $\widehat{\boldsymbol{\theta}}$ is updated with the Bayesian MAP method. The process of selecting the next item is:

1) Select two items first: one from Cluster 1 with $\boldsymbol{a}_i = (1,0)$ and $-d_i$ exactly equal to $\hat{\theta}_1$; and another from Cluster 2 with $\boldsymbol{a}_i = (0,1)$ and $-d_i$ exactly equal to $\hat{\theta}_2$.

2) Compute the value of $|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$ for the two items.

3) The optimal item is the one associated with a larger value of the $|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$.

The simulation continues as the test length reaches 30 items. The distributions of the MDIFF value of the administered items for these two examinees are shown in Figure 3.2 and Figure 3.3, respectively. These distributions used a *MDIFF*-bin width of 0.6 on the MDIFF scale to tally the number of items required in each bin. For both examinees, 15 items are from Cluster 1 and 15 from Cluster 2.

The comparison between the two distributions shows that the items selected for these two examinees have some in common. This means the second examinee can use the items that have been administered for the first examinee. Therefore, rather than needing 30+30 = 60 items, the *p*-optimal item pool for these two examinees requires only 56 items. This number is the count of the items in the union of the two sets. Figure 3.4 displays the distribution for the 56 items. Among the 56 items, half of them are from Cluster 1 and another half are from Cluster2.

When a third examinee is taking the test, the set of items required for that examinee can be determined. Then, the size and distribution of the *p*-optimal item pool can be determined by taking the union of items for the three examinees. This process can be continued until the number of items no longer increases. Figure 3.5 illustrates how the required item pool increases in size as the number of examinees increases. For the example given here, the item pool size reaches an asymptote at 340 items after 3,000 examinees.

Similar to the UCAT, the end product of the *p*-optimal item pool design for MCAT is a bin-count table, which tells the number of item in each MDIFF-bin for each dimension. The real *p*-optimal item pool used for test operation can be created based on this bin-count table.

3.3.5 p-Optimal Item Pool Design for MCAT with Exposure Control

If no item exposure control is implemented, the union of the optimal items for a large number of examinees is the blueprint for the operational *p*-optimal item pool development. If item

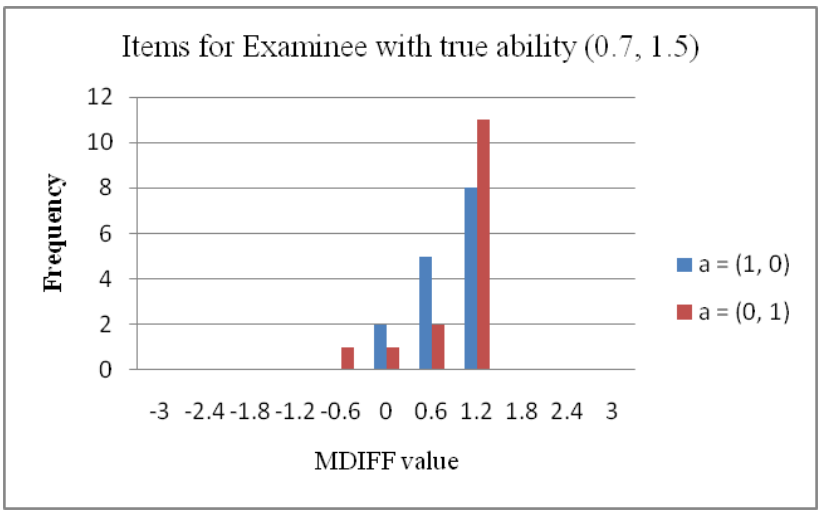


Figure 3.2: Item distributions for examinee with true ability (0.7, 1.5)

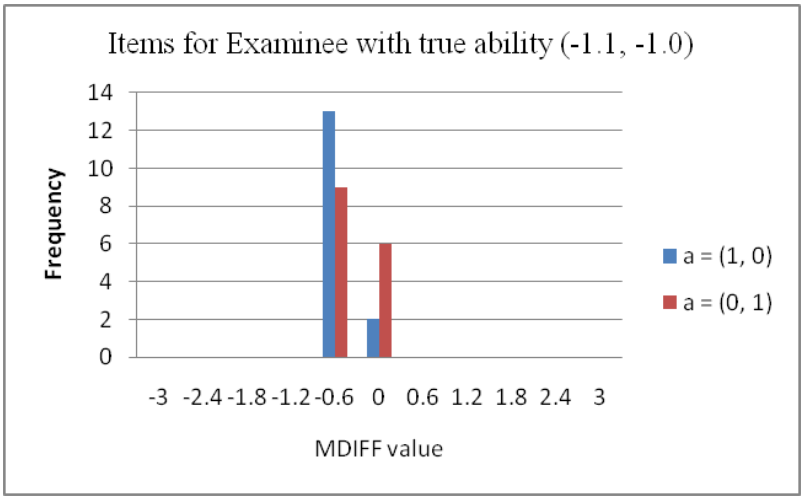


Figure 3.3: Item distributions for examinee with true ability (-1.1, -1.0)

exposure control is implemented in the adaptive test, a post-simulation adjustment (Gu, 2007) is used after the *p*-optimal item pool design process to make sure there are sufficient items in each bin where items are more often selected.

This study set a maximum item exposure rate, *R*, for all the items in the item pool. The item exposure rate is the number of times an item is administered divided by the total number of

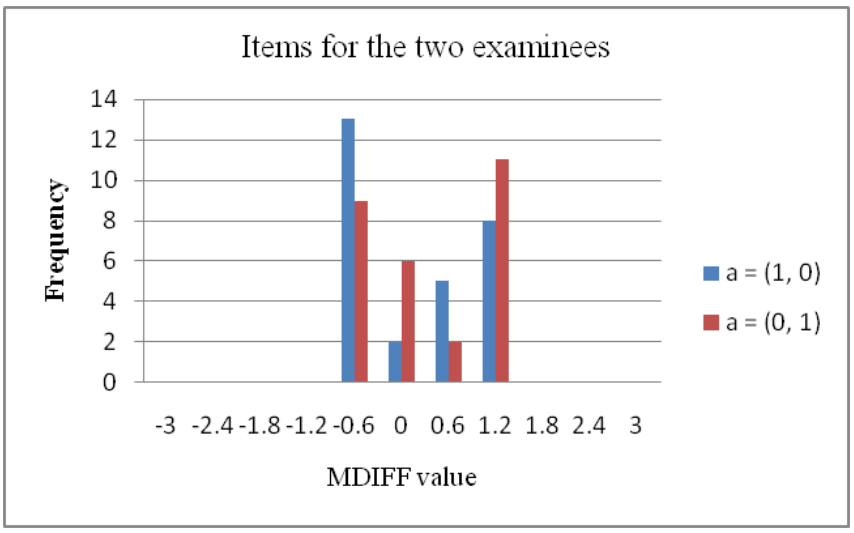


Figure 3.4: Item distributions for the two examinees

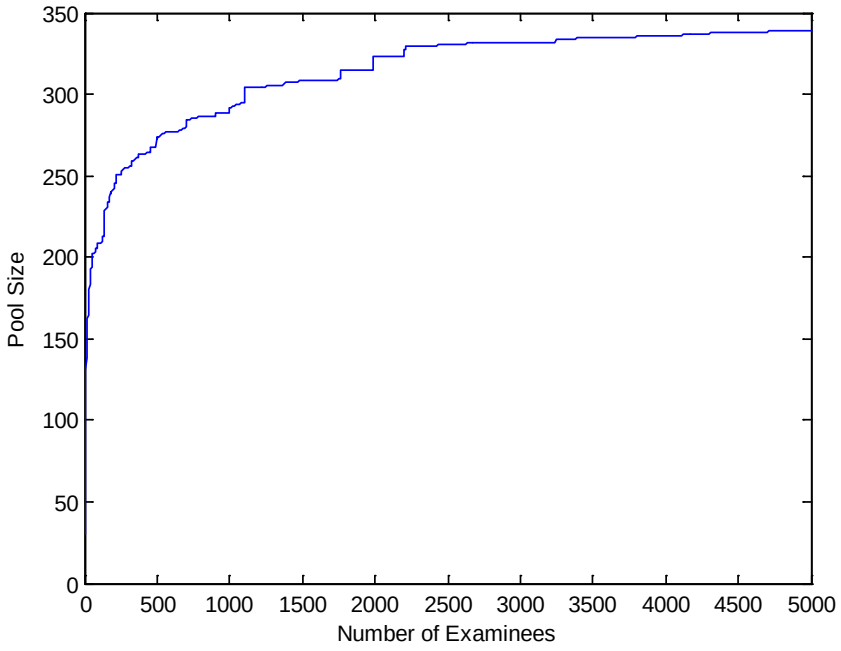


Figure 3.5: Increase in required pool size as number of examinees increases

examinees. During the *p*-optimal item pool design process, the actual item exposure rate for each item is not available, but the number of items from each MDIFF-bin that are administered can be documented. Suppose N is the total number of examinees used for the *p*-optimal item pool design process, m_j is the number of item in the j-th MDIFF-bin, and s_j is number of times

of an item from the *j*-th MDIFF-bin being administered. The expected item exposure rate, $\bar{r}_j$, for each item in the *j*-th MDIFF-bin can be obtained by

$$\bar{r}_j = \frac{s_j / m_j}{N}. \qquad (3.12)$$

Compare $\bar{r}_j$ with R for j = 1, 2, …, J, where J is the total number of MDIFF-bin's. If $\bar{r}_j$ is smaller than R, it implies that the number of items the *j*-th MDIFF-bin is sufficient so that the no post-simulation adjustment is not necessary. If $\bar{r}_j$ is larger than R, the number of items the *j*-th MDIFF-bin is insufficient and the adjustment is needed.

To ensure $\bar{r}_j \leq R$, the predicted number of item in the *j*-th MDIFF-bin, $\tilde{m}_j$, can be calculated by

$$\tilde{m}_j = \frac{s_j}{RN'}, \qquad (3.13)$$

where N' is the total number of examinees that is going to take the MCAT. The post-simulation adjustment is implemented by replacing m_j with $\tilde{m}_j$ for all the MDIFF-bin's with $\bar{r}_j > R$. In other words, the post-simulation adjustment sets the number of items the *j*-th MDIFF-bin to M_j, where M_j is defined by

$$M_j = \max\{m_j, \tilde{m}_j\}. \qquad (3.14)$$

If M_j is not an integer, it will be rounded up to the next integer.

3.3.6 p-Optimal Item Pool Design for MCAT with Non-Simple Structure

Suppose a third cluster of items that equally measures θ_1 and θ_2 with $\boldsymbol{a}_i = (1,1)$ is considered to the MCAT described above. This MCAT would be with the feature of non-simple structure. The two-dimensional Rasch model in this case can be written as

$$P(\boldsymbol{\theta}) = \begin{cases} P_1(\boldsymbol{\theta}) = \dfrac{e^{\theta_{1j}+d_i}}{1+e^{\theta_{1j}+d_i}}, \text{for items from Cluster 1} \\ P_2(\boldsymbol{\theta}) = \dfrac{e^{\theta_{2j}+d_i}}{1+e^{\theta_{2j}+d_i}}, \text{for items from Cluster 2} \\ P_3(\boldsymbol{\theta}) = \dfrac{e^{\theta_{1j}+\theta_{2j}+d_i}}{1+e^{\theta_{1j}+\theta_{2j}+d_i}}, \text{for items from Cluster 3} \end{cases}. \quad (3.15)$$

And the $I_{i_k}(\hat{\theta})$ can be specified as

$$I_{i_k}(\hat{\theta}) = \begin{cases} P_1Q_1\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}P_1Q_1 & 0\\ 0 & 0\end{bmatrix}, \text{for items from Cluster 1} \\ P_2Q_2\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & P_2Q_2\end{bmatrix}, \text{for items from Cluster 2} \\ P_3Q_3\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} = \begin{bmatrix}P_3Q_3 & P_3Q_3\\ P_3Q_3 & P_3Q_3\end{bmatrix}, \text{for items from Cluster 3} \end{cases}, \quad (3.16)$$

Suppose among the k-1 administered items, k_1 of them from Cluster 1, k_2 of them from Cluster 2, and k_3 of them from Cluster 3, where $k_1 + k_2 + k_3 = k - 1$. Substituting (3.16) into (3.6), we obtain

$$I_{s_{k-1}}(\hat{\theta}) = \begin{bmatrix} \sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i} & \sum_{i=1}^{k_3} P_{3i}Q_{3i} \\ \sum_{i=1}^{k_3} P_{3i}Q_{3i} & \sum_{i=1}^{k_2} P_{2i}Q_{2i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i} \end{bmatrix}. \quad (3.17)$$

Note the off-diagonal elements of the $I_{s_{k-1}}(\hat{\theta})$ are no longer zero because of adding items from Cluster 3. By substituting (3.16) and (3.17) into (3.9), we obtain

$$|I_{s_{k-1}}(\hat{\theta}) + I_{i_k}(\hat{\theta})|$$

$$= \begin{cases} \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)\left(\sum_{i=1}^{k_3} P_{3i}Q_{3i}\right) + \left(\sum_{i=1}^{k_2} P_{2i}Q_{2i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i}\right) P_{1i_k}Q_{1i_k} \\ \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)\left(\sum_{i=1}^{k_3} P_{3i}Q_{3i}\right) + \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i}\right) P_{2i_k}Q_{2i_k}, \\ \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i}\right)\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)\left(\sum_{i=1}^{k_3} P_{3i}Q_{3i}\right) + \left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_2} P_{2i}Q_{2i}\right) P_{3i_k}Q_{3i_k} \end{cases}$$

(3.18)

To determine the optimal item in this case, the amount of information on three directions needs to be compared: 1) $\sum_{i=1}^{k_1} P_{1i}Q_{1i}$ represents the amount of information on the direction of θ_1, 2) $\sum_{i=1}^{k_2} P_{2i}Q_{2i}$ is the amount of information on the direction of θ_2, and 3) $\sum_{i=1}^{k_3} P_{3i}Q_{3i}$ is the amount of information on the direction of 45 degree line (See Figure 3.6). The Direction 1, 2, and 3 shown in Figure 3.6 is the direction best measured by items from Cluster 1, 2, and 3, respectively. That is, the direction with the maximum discrimination power.

If the amount of information on the direction of θ_1 is the smallest (i.e., $\left(\sum_{i=1}^{k_2} P_{2i}Q_{2i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i}\right)$ is larger than $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i}\right)$ and $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)$), the optimal item is from Cluster 1 with $\boldsymbol{a}_i = (1,0)$ and $-d_i = \theta_{1j}$. If the amount of information on the direction of θ_2 is the smallest (i.e., $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_3} P_{3i}Q_{3i}\right)$ is larger than the other two), the optimal item is from Cluster 2 with $\boldsymbol{a}_i = (0,1)$ and $-d_i = \theta_{2j}$. If the amount of information on the direction of the 45 degree line is the smallest (i.e., $\left(\sum_{i=1}^{k_1} P_{1i}Q_{1i} + \sum_{i=1}^{k_2} P_{2i}Q_{2i}\right)$ is the largest), the optimal item is from Cluster 3 with $\boldsymbol{a}_i = (1,1)$ and $-d_i = \theta_{1j} + \theta_{2j}$. If the three terms are the same, the optimal item is randomly picked.

Because the *d*-parameter for optimal items from Cluster 3 is equal to $-(\theta_{1j} + \theta_{2j})$, the scale of the *d*-parameter for items from Cluster 3 is different from the scale of the *d*-parameter for items from Cluster 1 and 2. That is, two-unit distance on the *d*-parameter for items from Cluster 3 is corresponding to one-unit distance on the *d*-parameter for items from Cluster 1 and 2. Therefore, to meet the criterion of the *p*-optimal item pool, the width of the *d*-bin for items from Cluster 3 should be twice of the width for items from Cluster 1 and 2. Because this study adopts the MDIFF-bin instead of *d*-bin, and $MDIFF = -d_i/\sqrt{2}$ for items from Cluster 3, the width of the MDIFF-bin for items from Cluster 3 is $\sqrt{2}$ times larger than width of the MDIFF-bin for

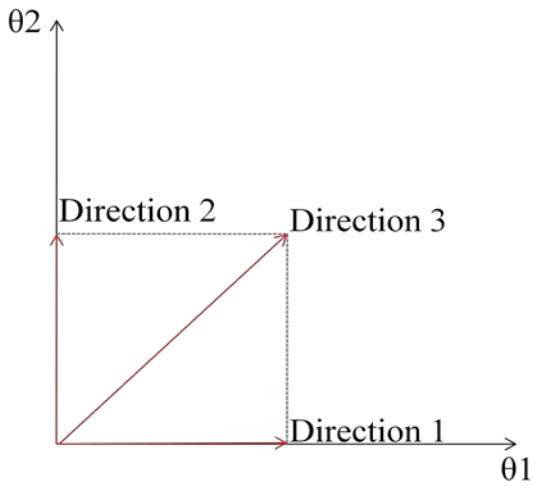


Figure 3.6: The test information on three directions

items from Cluster 1 and 2. For the .9-optimal item pool, the width of the MDIFF-bin for items from Cluster 1 and 2 is 0.65, and for items from Cluster 3 is $\sqrt{2} * 0.65$, which is 0.92. For the MCAT with higher order dimension, the width of the MDIFF-bin for items measuring more than one dimension can be determined in a similar way.

Chapter 4 Study Design and Procedures

In this chapter, the algorithms for the multidimensional computerized adaptive testing (MCAT) are first defined in Section 4.1. Section 4.2 then describes a simulation study that was used to compare the *p*-optimal item pools with other item pools existed in literature. The criteria for item pool comparison are introduced in Section 4.3.

4.1 MCAT Algorithms

The MCAT in this study is based on the multidimensional Rasch model defined by (2.11). Three test specifications are considered:

- Test specification 1: two-dimension simple structure. In this case, the item pool consists of two clusters of items: items from Cluster 1 with $\boldsymbol{a}_i = (1,0)$ only measure θ_1 and items from Cluster 2 with $\boldsymbol{a}_i = (0,1)$ only measure θ_2.
- Test specification 2: three-dimension simple structure. In this case, the item pool consists of three clusters of items: items from Cluster 1 with $\boldsymbol{a}_i = (1,0,0)$ only measure θ_1, items from Cluster 2 with $\boldsymbol{a}_i = (0,1,0)$ only measure θ_2, and items from Cluster 3 with $\boldsymbol{a}_i = (0,0,1)$ only measure θ_3.
- Test specification 3: three-dimension non-simple structure. In this case, the item pool consists of four clusters of items: items from Cluster 1 with $\boldsymbol{a}_i = (1,0,0)$ only measure θ_1, items from Cluster 2 with $\boldsymbol{a}_i = (0,1,0)$ only measure θ_2, items from in Cluster 3 with $\boldsymbol{a}_i = (0,0,1)$ only measure θ_3, and items from Cluster 4 with $\boldsymbol{a}_i = (1,1,1)$ equally measure θ_1, θ_2 , and θ_3.

For the MCAT simulation in this study, items are selected by the D-optimality method, $\boldsymbol{\theta}$ is estimated using the Bayesian MAP method, and test length is fixed at 30 items. The prior for the

Bayesian MAP is the multivariate normal distribution of the true $\boldsymbol{\theta}$ in this study (Segall, 1996). The MCAT with and without item exposure control are considered in this study. For the MCAT with exposure control, the maximum item exposure rate is fixed at 0.2, and the Modified MPI method is used to make sure the exposure rate for all items in the item pool are less than 0.2. A detailed description of the D-optimality, the Bayesian MAP, and the Modified MPI methods can be found in Chapter 2 and 3.

4.2 Simulation Procedure

This study is carried out in four major phases. In the first phase, a *p*-optimal item pool based on each test specification is designed and the bin-count table is created. In the second phase, the actual *p*-optimal item pools are developed based on the bin-count table created from the previous phase. In the third phase, a baseline pool for each test specification is developed for comparison purposes. In the fourth phase, a simulation study is carried out to evaluate the performance of the MCAT using a *p*-optimal item pool against the MCAT using a baseline pool.

Phase I. P-optimal Item Pool Design

Based on the test specifications and adaptive algorithms described in the section 4.1, *p*-optimal item pools are designed to guarantee that every item that was requested by the item selection rule is available for administration. As described in Chapter 3, the design for the *p*-optimal item pools should also based on characteristics of the target examinee population. In this study, the examinee population for the CAT-ASVAB in Segall (1996) is adopted to design the *p*-optimal item pools. The CAT-ASVAB measures nine content areas, and each content area is treated as one dimension. The correlation among the nine dimensions ranges from 0.2 to 0.9. The MCAT in this study only measures a two- or three-dimensional ability; thus, two or three

content areas from the CAT-ASVAB are selected to use in this study. To investigate how the correlation among dimensions affects the *p*-optimal item pool design, both moderately correlated content areas and highly correlated content areas are selected. The low correlation condition is not considered in this study as it is rare in educational assessments.

For the moderate correlation condition, the three content areas are the Arithmetic Reasoning (AR), Word Knowledge (WK), and Electronics Information (EI). For the high correlation condition, the three dimensions are the General Science (GS), Word Knowledge (WK), and Paragraph Comprehension (PC). The mean and the variance-covariance matrix for ability that requires for these content areas are shown in Table 4.1.

To design the *p*-optimal item pool for each condition, 3,000 examinees were randomly sampled from the multivariate normal distribution with mean vector and variance-covariance matrix described in Table 4.1. The number of 3,000 is used here because, as shown in Figure 3.5, the size of the *p*-optimal item pool reaches the asymptote after 3,000 examinees. For each examinee, all items administered in each cluster were allocated to the MDIFF-bins. Two sets of bin sizes, .4 and .8, corresponding to a .96- and .86-optimal pool respectively, were considered in this study.

In total, 24 *p*-optimal item pools (i.e., 3 Test Specifications * 2 correlations * 2 bin sizes * with or without exposure control) are designed in this study. To eliminate potential sampling errors, 100 replications were conducted. The final bin-count table for each *p*-optimal item pool is the average of its 100 replications. Table 4.2 shows a bin-count table for the *.96*-optimal item pool for the MCAT with test specification of three-dimension non-simple structure, moderate correlation among dimensions, and without exposure control. Table 4.3 is a bin-count table for

Table 4.1: Mean and covariance matrix for the two examinee populations

	Moderate Correlation		**High Correlation**	
	2-dimension	3-dimension	2-dimension	3-dimension
Dimension	AR and WK	AR, WK, and EI	GS and WK	GS, WK, and PC
Mean Vector	[0,0]	[0,0,0]	[0,0]	[0,0,0]
Variance-Covariance Matrix	$\begin{bmatrix} 1 & .61 \\ .61 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & .61 & .64 \\ .61 & 1 & .72 \\ .64 & .72 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & .91 \\ .91 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & .91 & .81 \\ .91 & 1 & .88 \\ .81 & .88 & 1 \end{bmatrix}$

Table 4.2: Bin count for a *.96*-optimal item pool

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0, 0)	1	3	4	5	6	7	7	7	8	7	7	7	6	5	4	3	1
a = (0, 1, 0)	1	3	4	5	6	7	7	7	7	7	7	7	6	5	4	3	1
a = (0, 0, 1)	2	3	4	5	6	7	7	7	8	7	7	7	6	6	5	3	1
MDIFF	**-5.6**	**-4.9**	**-4.2**	**-3.5**	**-2.8**	**-2.1**	**-1.4**	**-0.7**	**0**	**0.7**	**1.4**	**2.1**	**2.8**	**3.5**	**4.2**	**4.9**	**5.6**
a = (1, 1, 1)	1	3	5	6	7	7	8	8	8	8	7	7	7	6	5	3	1

Note: the values on the first row represent the central point of each item bin;
the values on the second and third row represent the number of items in each item bin.

Table 4.3: Bin count for a *.86*-optimal item pool

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0, 0)	2	5	7	8	8	8	7	5	2
a = (0, 1, 0)	2	5	7	8	8	8	7	5	2
a = (0, 0, 1)	2	5	7	8	8	8	7	5	3
MDIFF	**-5.6**	**-4.2**	**-2.8**	**-1.4**	**0**	**1.4**	**2.8**	**4.2**	**5.6**
a = (1, 1, 1)	2	6	8	9	9	9	8	6	2

Note: this table can be interpreted in the same way as Table 4.2.

the *.86*-optimal item pool for the same MCAT. There are 17 MDIFF-bins for the *.96*-optimal item pool and 9 MDIFF-bins for the *.86*-optimal item pool.

Phase II. P-Optimal item pool development

With the bin-count table for the 24 *p*-optimal item pools, the *p*-optimal item pool can be developed accordingly. In practice, real items should be created to match the bin-count table. In

this study, items are simulated. Items within each MDIFF-bin are set to be equally distributed. For example, if three are 8 items in the central MDIFF-bin for items with $\boldsymbol{a}_i = (1,0,0)$, 8 items with MDIFF value equally distributed from -0.2 to 0.2. The MDIFF value is then converted to *d*-parameter according to equation (2.9) and (2.10). Therefore, 24 *p*-optimal item pools can be developed by simulation based on bin-count tables created in the previous phase.

Phase III. Baseline Pool Development

To evaluate the 24 *p*-optimal item pools, baseline pools should be created as the bases for comparison. Previous studies (e.g., Gu, 2007; He, 2010; Reckase, 2010) use existing operational item pools as the bases. However, there is no existing operational MCAT program so far and herein the operational multidimensional item pool is not available. Therefore, the item pools used for MCAT in research articles are adopted in this study as the baseline pools. Some of the multidimensional item pools in current literature are modified from its correspondent unidimensional operational item pool. For instance, Segall (1996) and Yao (2012, 2013) created the multidimensional item pool based on the operational item pool for CAT-ASVAB. Other multidimensional item pools in the literature are created by pure simulation, such as the item pool used in van der Linden (1996, 1999).

In this study, because the target examinee population and content areas are based on the CAT-ASVAB, it is straightforward to develop the baseline pools based on the CAT-ASVAB as well. There are three test specifications for the 24 *p*-optimal item pools in this study. Item pools with different test specifications cannot be compared. Therefore, three baseline pools, one for each test specification, are created based on the CAT-ASVAB. Yao (2013) provided a detailed description of the multidimensional item pool for the CAT-ASVAB, including the pool size and

item distribution. Based on Yao's description, the development for three baseline pools is described below.

For Test Specification 1(two-dimensional simple structure), the baseline pool consists of 480 items with 240 items from each of the two clusters. In this study, the MCAT based on this test specification gives 15 items from each cluster to each examinee. In the CAT-ASVAB, 15 AR items and 15 WK items are administered, and the number of AR or WK item in the item pool is around 240. This is the reason for setting the size of the baseline pool to 2*240 = 480 for Test Specification 1. For Test Specification 2 (three-dimensional simple structure), the baseline pool consists of 480 items with 160 items from each of the three clusters. For Test Specification 3 (three-dimensional non-simple structure), the baseline pool is consisted of 560 items with 140 items from each of the four clusters. Similar reasons are used to determine the pool size for Test Specification 2 and 3. The mean and standard deviation (SD) of the MDIFF value for the items in the three baseline pools are presented in Table 4.4.

Phase IV. Simulation Study Conduct

A simulation study is conducted to compare the performance of the MCAT using *p*-optimal item pools against MCAT using baseline pools. The algorithm for the MCAT is described in Section 4.1. Two types of examinee distribution were used for the simulation.

First, to evaluate the MCAT performance in general, 5,000 examinees are randomly sampled from the multivariate normal distribution with mean vector and variance-covariance matrix specified in Table 4.1.

Second, to evaluate the MCAT performance at each $\boldsymbol{\theta}$ point, 100 examinees are generated at several $\boldsymbol{\theta}$ points. The 29 $\boldsymbol{\theta}$ points for the two dimensional case are displayed in Figure 4.1. No point on the upper left and lower right is selected. This is because, given θ_1 and θ_2 are highly or

Table 4.4: Item Statistics for the Three Baseline pools

	2-dimension simple structure			3-dimension simple structure			3-dimension non-simple structure		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Cluster 1	240	-0.76	2.55	160	-0.76	2.55	140	-0.76	2.55
Cluster 2	240	-0.35	3.07	160	-0.35	3.07	140	-0.35	3.07
Cluster 3				160	-0.17	2.12	140	-0.17	2.12
Cluster 4							140	0.10	2.58

Table 4.5: The 37 $\boldsymbol{\theta}$ Points for the Three Dimensional MCAT

No.	θ_1	θ_2	θ_3	No.	θ_1	θ_2	θ_3	No.	θ_1	θ_2	θ_3
1	-3	-3	-3	13	-1	-2	-1	25	1	2	1
2	-3	-3	-2	14	-1	0	-1	26	1	2	2
3	-3	-2	-3	15	-1	0	0	27	2	1	1
4	-3	-2	-2	16	0	-1	-1	28	2	1	2
5	-2	-3	-3	17	0	-1	0	29	2	2	1
6	-2	-3	-2	18	0	0	-1	30	2	2	2
7	-2	-2	-3	19	0	0	0	31	2	2	3
8	-2	-2	-2	20	0	0	1	32	2	3	2
9	-2	-2	-1	21	0	1	0	33	2	3	3
10	-2	-1	-2	22	0	1	1	34	3	2	2
11	-2	-1	-1	23	1	0	0	35	3	2	3
12	-1	-2	-2	24	1	0	1	36	3	3	2
								37	3	3	3

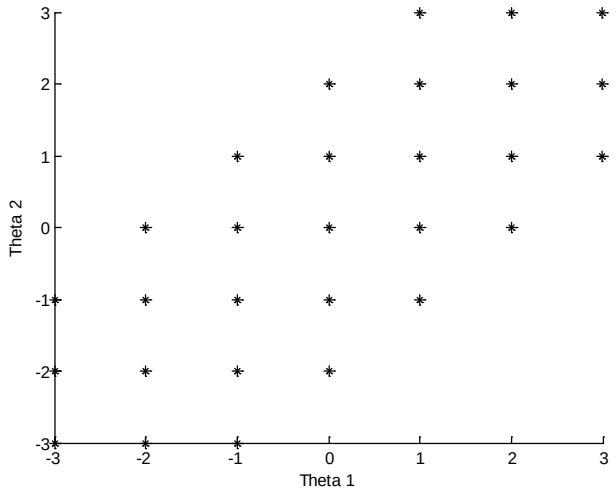


Figure 4.1: The 29 $\boldsymbol{\theta}$ Points for the Two Dimensional MCAT

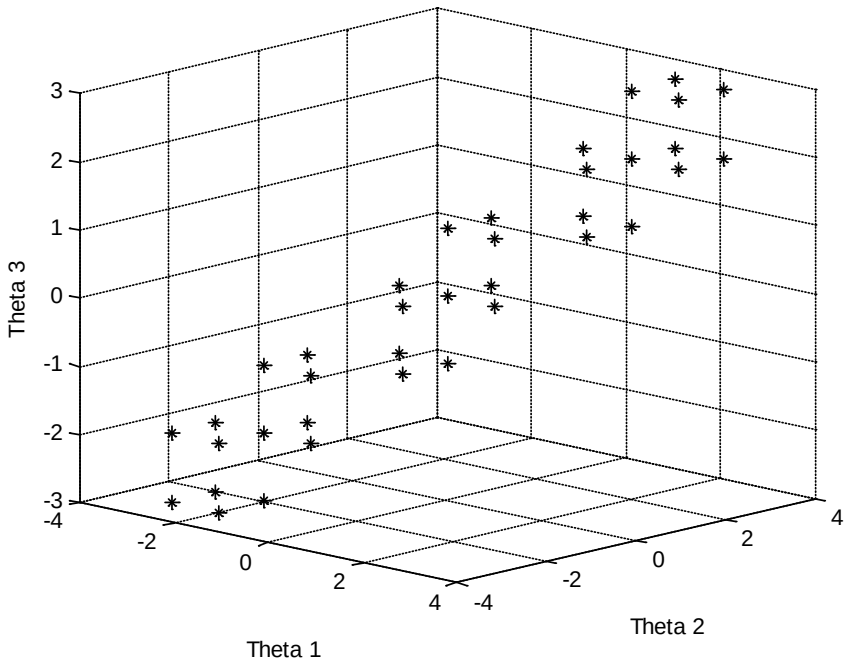


Figure 4.2: The 37 $\boldsymbol{\theta}$ Points for the Three Dimensional MCAT

moderately correlated, examinees are very unlikely to have a very high value in θ_1 and very low value in θ_2, or vice versa. The 37 $\boldsymbol{\theta}$ points for the three dimensional case are displayed in Table 4.5 and Figure 4.2. Again, because θ_1, θ_2, and θ_3 are correlated, only a limited number of points on the three dimensional space are selected.

4. 3 Evaluation Criteria

The performance of MCAT is evaluated based on precision of the ability estimation and the item pool utilization. The evaluation criteria for precision of the ability estimation include Pearson product-moment correlation between the true $\boldsymbol{\theta}$ and estimated $\boldsymbol{\theta}$, bias, and root mean squared error (RMSE). The bias and RMSE are denoted as:

$$Bias = \sum_{i=1}^{n} \frac{\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{\theta}_{\boldsymbol{i}}}{n}, \qquad (4.1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} - \boldsymbol{\theta}_{\boldsymbol{i}}\right)^2}{n}}, \qquad (4.2)$$

where n is the sample size.

For item pool utilization, the evaluation criteria are the overall pool usage, the test overlap rate, and the percentage of items with varying exposure rate. As Chang and Ying (1999) proposed, the efficiency of overall item pool usage can be measured by the discrepancy between the observed and expected item exposure rate. It follows χ^2 distribution and is denoted as

$$\chi^2 = \sum_{j=1}^{N} \frac{\left(r_j - L/N\right)^2}{L/N}, \qquad (4.3)$$

where r_j is the observed exposure rate for item *j*, *L* is the test length, *N* is the number of items in the item pool. A low χ^2 value implies that most of the items are fully used.

Test overlap describes item exposure as well, and it has been used as item pool security index. Overlap rate is defined as the average proportion of items that two randomly selected examinees have in common (Way, 1998):

$$R = \frac{T/C_n^2}{\sum_{i=1}^{n} L_i/n}, \qquad (4.4)$$

where T is the total number of item shared by C_n^2 pair of n examinees in the test and $\sum_{i=1}^{n} L_i/n$ is the total number of the items administered for n examinees. In practice, the overlap rate less than 15% is desired.

Item exposure rate is the ratio of the number of item administrations to the total number of examinees. In this study, the percentage of items over- and under-exposed for each item pool is

also reported. A rate higher than 0.2 is regarded as overexposed (Segall, Moreno, & Hetter, 1997), and a rate lower than 0.02 is regarded as underexposed (Gu, 2007).

Chapter 5 Simulation Results

The simulation results are summarized in two parts. The first part presents the general characteristics of the 24 *p*-optimal item pools, and how the characteristics are affected by test specification, exposure control, correlation among dimensions, and bin size. The second part describes the performance of the MCAT using each *p*-optimal item pool, and how their performance compared with the MCAT using baseline pools.

5.1 Item Pool Characteristics

Because the primary purpose of this study is to design and develop *p*-optimal item pools for MCAT, the results of the item pool develop are presented first in this chapter. The general characteristics for the *p*-optimal item pools and the baseline pools are summarized and compared in Section 5.1.1. The item distribution for the 24 *p*-optimal item pools is then described in 5.1.2.

5.1.1 Summary for Item Pool Characteristics

The summary characteristics, including pool size and the mean and standard deviation (SD) of the item difficulty, for the .*96*-optimal item pools and .*86*-optimal item pools are presented in Table 5.1 and 5.2, respectively. The twelve .*96*-optimal item pools are based on the bin-size of 0.4, and the twelve.*86*-optimal item pools are based on the bin-size of 0.8. The characteristics for the three baseline pools are also presented in the two tables.

All the .*96*-optimal item pools, as shown in Table 5.1, have smaller pool sizes than the baseline pools. For the 2-dimension simple structure and 3-dimension simple structure cases, the pool size for the .*96*-optimal item pools is about 110 to 150 items less than the baseline pools. For the 3-dimension non-simple structure case, the pool size for the .*96*-optimal item pools is about 150 to 190 items less than the baseline pools. The average difficulty level (i.e., the mean

Table 5.1: Summary for the *.96*-optimal item pools and baseline pools

Test specification	Statistics	High Correlation		Moderate Correlation		Baseline pool
		No Exposure Control	Exposure Control	No Exposure Control	Exposure Control	
2-dimension Simple Structure	Pool size	369	371	328	333	480
	Mean of Difficulty	0.01	-0.01	0.02	-0.01	-0.52
	SD of Difficulty	1.65	1.63	1.57	1.55	2.75
3-dimension Simple Structure	Pool size	366	370	322	330	480
	Mean of Difficulty	0.02	0.00	0.00	0.01	-0.32
	SD of Difficulty	1.61	1.58	1.51	1.48	2.72
3-dimension Non-simple Structure	Pool size	407	407	363	369	560
	Mean of Difficulty	-0.01	-0.01	-0.01	0.00	-0.32
	SD of Difficulty	2.09	2.09	1.95	1.94	2.68

Table 5.2: Summary for the *.86*-optimal item pools and baseline pools

Test specification	Statistics	High Correlation		Moderate Correlation		Baseline pool
		No Exposure Control	Exposure Control	No Exposure Control	Exposure Control	
2-dimension Simple Structure	Pool size	206	252	192	246	480
	Mean of Difficulty	0.00	0.00	0.00	0.00	-0.52
	SD of Difficulty	1.82	1.66	1.70	1.55	2.75
3-dimension Simple Structure	Pool size	207	251	190	236	480
	Mean of Difficulty	-0.02	0.00	-0.03	0.00	-0.32
	SD of Difficulty	1.76	1.60	1.63	1.47	2.72
3-dimension Non-simple Structure	Pool size	233	272	216	253	560
	Mean of Difficulty	0.00	0.01	0.01	0.00	-0.32
	SD of Difficulty	2.28	2.14	2.11	1.95	2.68

of MDIFF) for all the *.96*-optimal item pools is around zero. This is as expected because the mean ability of the target examinee population is zero and the *p*-optimal item pools are developed based on this examinee population. The mean difficulty level for all the baseline pools is slightly below zero, suggesting the items in the baseline pools are easier than the items

in the *p*-optimal item pools on average. The comparison between the SD's for the .*96*-optimal item pools and the baseline pools suggests that items in baseline pools are more widely distributed.

All the .*86*-optimal item pools, as shown in Table 5.2, also have smaller pool sizes than the baseline pools. The pool size for the .*86*-optimal item pools is about half or less than half of the baseline pools. The mean difficulty level for all the .*86*-optimal item pools is around zero. The SD of item difficulty for the .*86*-optimal item pools is also smaller than the baseline pools.

The comparison between the .*96*- and the .*86*-optimal item pools tells the effect of bin size on the *p*-optimal item pools design. First, the pool size of the .*96*-optimal item pools is much larger than the .*86*-optimal item pools. For conditions without item exposure control, the pool size of the .*96*-optimal item pools is about twice as much as the .*86*-optimal item pools. Therefore, a larger bin size results in a larger item pool. Similar results can be found for the UCAT in Reckase (2003). Second, the SD of item difficulty for the .*96*-optimal item pools is slightly smaller than the .*86*-optimal item pools. Although the range of the item difficulty for both the .*96*- and .*86*-optimal item pools is similar, the proportion of the difficult or easy items is slightly higher for the .*86*-optimal item pools, and thus the SD value is larger. For example, there are 6% items with MDIFF larger than 2.8 in the .*86*-optimal item pool for condition of 2-dimension simple structure, high correlation, and no exposure control; while there are only 4% for the .*96*-optimal item pool for the same MCAT.

In addition to the bin size, test specifications also affect the *p*-optimal item pools design. The pool size for the all the *p*-optimal item pools based on 2- and 3-dimension simple structure is very similar, except there is a 5-item difference between the two .*96*-optimal item pools with moderate correlation and no exposure control. The *p*-optimal item pools with test specification

of 3-dimension non-simple structure consist of about 10-12% more items than the rest of the two test specifications. Therefore, if the test length is the same, adding one more clusters of items that measure a different content area does not require a larger item pool (e.g., from 2-dimension simple structure to 3-dimension simple structure); however, if the added items measure more than one content area (e.g., from 3-dimension simple structure to 3-dimension non-simple structure), the pool size needs to be increased. In addition to the pool size, the SD of item difficulty is also affected by test specifications. The SD in the 2-dimension simple structure condition is slightly larger than the SD in the 3-dimension simple structure condition. The items in the *p*-optimal item pools based on both types of test specification have the same difficulty range, but the proportion of difficult item and easy item is slightly larger for the 2-dimension simple structure condition. The SD for the 3-dimension non-simple structure is much larger compared with the SD for the other two test specifications. This is because the item difficulty for items measuring all the three content areas (with $\boldsymbol{a}_i = (1,1,1)$) is more spread.

This study also examines how the correlation among dimensions affects the design for the *p*-optimal item pool. Table 5.1 and 5.2 show that if dimensions are highly correlated, the pool size and the SD of item difficulty will be larger than the condition that dimensions are moderately correlated. This is because, when dimensions are highly correlated, a slightly larger number of examinees will have very high ability in all dimensions, and thus more difficult items are needed in the item pool. For the similar reason, more easy items are also need in the item pool when dimensions are highly correlated.

If item exposure control is implemented in the MCAT, a larger item pool is necessary. Similar results can be found in Gu (2007), He (2012), and Zhou (2013) for UCAT. For the *.96*-optimal item pools, given the pool size is already over 350 items, adding item exposure control

only increases the pool size by less than 10 items. For the *.86*-optimal item pools, about 40-50 more items are added to the item pool in order to minimize item exposure rate and meanwhile to provide precise ability estimation. Because items with difficulty level around zero have a higher possibility to be selected (as more examinees are located in the middle), those additional items are all added to the MDIFF-bins in the middle, and therefore, the SD values for the *p*-optimal item pools with item exposure control decrease.

In summary, the characteristics of the *p*-optimal item pools change with different bin sizes, test specification, correlation among dimensions, as well as whether item exposure control is implemented. A larger item pool is necessary if the bin size decreases, the test becomes non-simple structure, dimensions are highly correlated, or item exposure control is considered.

5.1.2 Item distribution for p-optimal item pools

Each of the *p*-optimal item pool consists of items from more than one cluster. The number of items in each cluster for the *.96*- and *.86*-optimal item pools is presented Table 5.3 and 5.4, respectively. For the 2-dimension simple structure case, half items are from Cluster 1, and the other half are from Cluster 2. For the 3-dimension simple structure case, one third of items are from each cluster. For the 3-dimension non-simple structure case, there is same number of items from Cluster 1 – 3, and slightly more items from Cluster 4.

For the 2- and 3-dimension simple structure cases, the reason of items equally distributed between each cluster with simple structure is because the D-Optimality method selects the same number of items from each cluster. Based on equation (3.8), when an item from Cluster 1 is administered, the test information on the direction of dimension 1 will be slightly larger than that of dimension 2, and the next item from Cluster 2 will be selected next. After this item is administered and the ability estimate is updated, the test information on the direction of

Table 5.3: Item distribution for the *.96*-optimal item pools

Test Specification	Number of Items	High Correlation		Moderate Correlation	
		No Exposure Control	Exposure Control	No Exposure Control	Exposure Control
2-dimension Simple Structure	Item with $\boldsymbol{a}$ = (1,0)	184	185	164	167
	Item with $\boldsymbol{a}$ = (0,1)	185	186	164	166
	Total	369	371	328	333
3-dimension Simple Structure	Item with $\boldsymbol{a}$ = (1,0,0)	122	124	106	105
	Item with $\boldsymbol{a}$ = (0,1,0)	123	124	111	107
	Item with $\boldsymbol{a}$ = (0,0,1)	121	122	113	110
	Total	366	370	330	322
3-dimension Non-simple Structure	Item with $\boldsymbol{a}$ = (1,0,0)	100	100	88	89
	Item with $\boldsymbol{a}$ = (0,1,0)	102	101	87	89
	Item with $\boldsymbol{a}$ = (0,0,1)	100	100	91	92
	Item with $\boldsymbol{a}$ = (1,1,1)	105	106	97	99
	Total	407	407	363	369

Table 5.4: Item distribution for the *.86*-optimal item pools

Test Specification	Number of Items	High Correlation		Moderate Correlation	
		No Exposure Control	Exposure Control	No Exposure Control	Exposure Control
2-dimension Simple Structure	Item with $\boldsymbol{a}$ = (1,0)	103	126	96	123
	Item with $\boldsymbol{a}$ = (0,1)	103	126	96	123
	Total	206	252	192	246
3-dimension Simple Structure	Item with $\boldsymbol{a}$ = (1,0,0)	68	83	60	76
	Item with $\boldsymbol{a}$ = (0,1,0)	71	85	64	80
	Item with $\boldsymbol{a}$ = (0,0,1)	68	83	66	80
	Total	207	251	190	236
3-dimension Non-simple Structure	Item with $\boldsymbol{a}$ = (1,0,0)	56	66	52	61
	Item with $\boldsymbol{a}$ = (0,1,0)	56	66	52	61
	Item with $\boldsymbol{a}$ = (0,0,1)	56	65	53	61
	Item with $\boldsymbol{a}$ = (1,1,1)	65	75	59	70
	Total	233	272	216	253

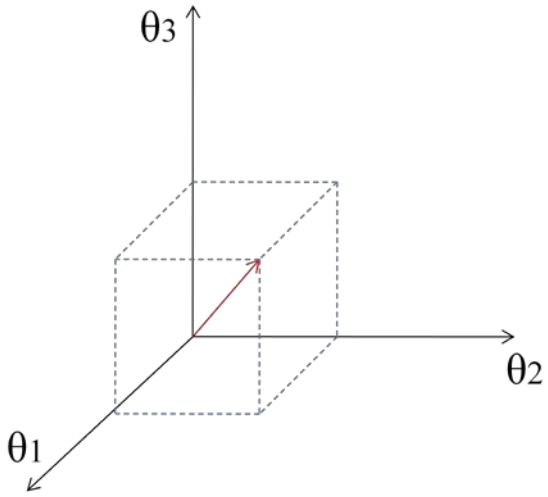


Figure 5.1: The direction of the information for items with ***a*** = (1,1,1)

dimension 2 will be larger than that of dimension 1, and an item from Cluster 1 will be selected. Therefore, items from each cluster take turns to be selected next. Occasionally, two items from the same clusters are administered successively. If this happen, two items from another cluster will be selected to balance test information between the two directions.

For the 3-dimension non-simple structure case, items measuring all the three dimensions are included in the item pool as the 4th cluster. Items from Cluster 4 provide 1 unit of information on the direction of the θ_1, θ_2,and θ_3 composite (see Figure 5.1), and also $\frac{\sqrt{3}}{3}$ unit of information on the direction of θ_1, θ_2,and θ_3. Items from Cluster 1 – 3 provide 1 unit of information on the direction of θ_1, θ_2,or θ_3, and also a small amount of information on the direction of the composite. Suppose three items, one from Cluster 1, one from Cluster 2, and one from Cluster 3, has been administered. At this point, the information on the direction of the diagonal is the smallest; thus, the fourth item is chosen from Cluster 4. Then, items from Cluster 1 – 3 are selected next. Most of the time, items from the four cluster take turns being selected. Because the amount of information that items from Cluster 4 provide on the direction of θ_1, θ_2,or θ_3 is

more than the amount of information that items from Cluster 1 – 3 provide on the direction of the composite, Cluster 1 – 3 may be skipped sometimes in each rotation. Therefore, in this study, about 8 to 9 items from Cluster 4, and about 7 to 8 items from each of the Cluster 1, 2, and 3, are given to each examinee. Because more items from Cluster 4 are administered, more items should be available in the item pool.

The distribution for the *.96*- and *.86*-optimal item pool without exposure control (two-dimension simple structure, high correlation) is presented in Figure 5.2 and 5.3, respectively. Each bar in the figure represents the number of item in each MDIFF-bin. For both item pool, the distribution for item difficulty is flatter than a normal distribution. Half of the items are from Cluster 1 and the other half are from Cluster 2. Figure 5.4 and 5.5 present the distribution for the.*96*- and *.86*-optimal item pool with exposure control (two-dimension simple structure, high correlation), respectively. For both item pools, items are distributed from -3.2 to 3.2, with many more items located in the middle bins. Figure 5.2 and 5.4 is only different in the central MDIFF-bin: 15 items in 5.2 and 17 in 5.4. They look different because the scale of y-axis is different. For the *.86*-optimal item pool, the difference between Figure 5.3 and 5.5 is in the three bins in the middle. Because of the item exposure control, the number of item in the central MDIFF-bin is double in Figure 5.5. The distribution for *p*-optimal item pools in other condition is in a similar shape, and therefore they are not represented here. The number of items in each item MDIFF-bin for all the 24 *p*-optimal item pool can be found in the Appendix.

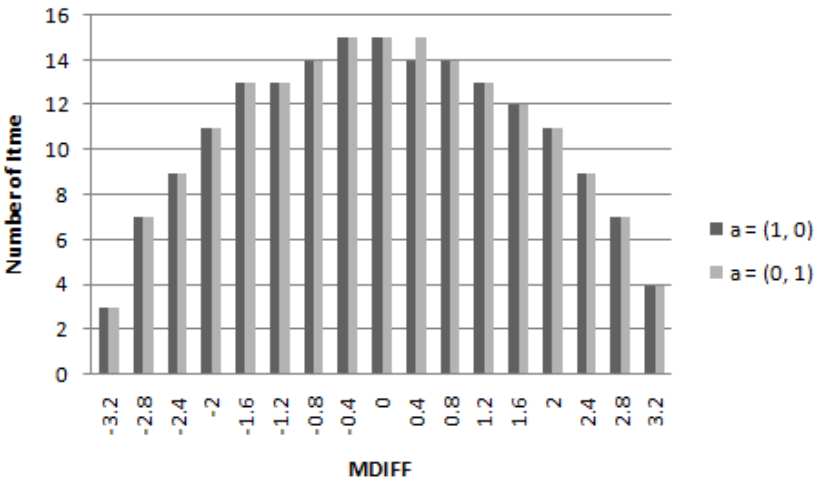


Figure 5.2: Item distribution for the .*96*-optimal item pool without exposure control (Two-dimension simple structure, high correlation)

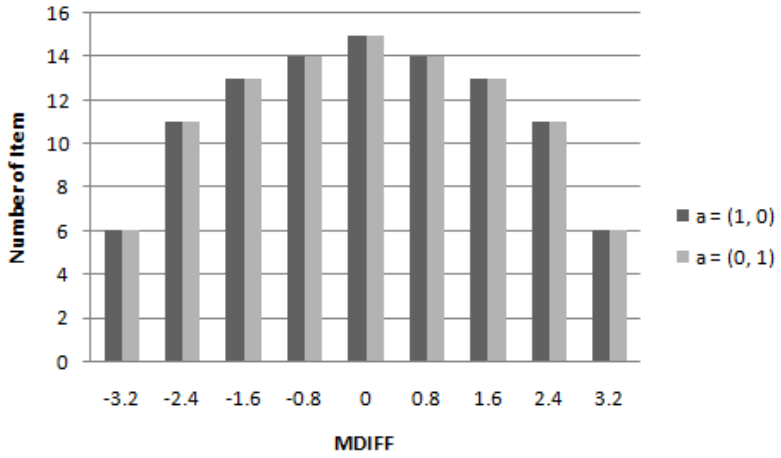


Figure 5.3: Item distribution for the .*86*-optimal item pool without exposure control (Two-dimension simple structure, high correlation)

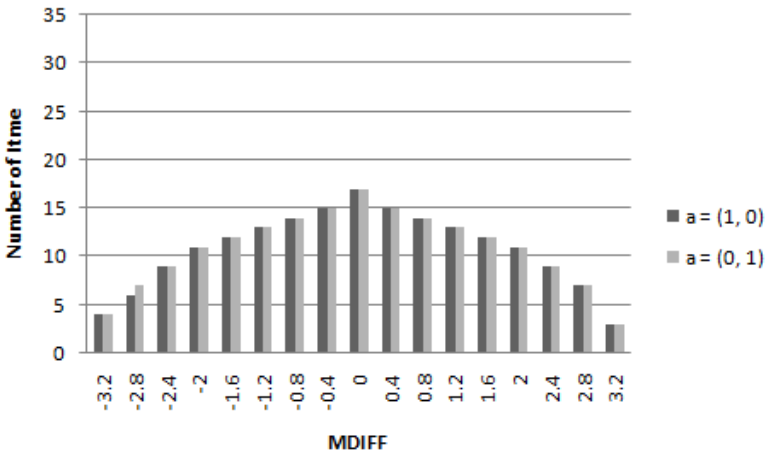


Figure 5.4: Item distribution for the .*96*-optimal item pool with exposure control (Two-dimension simple structure, high correlation)

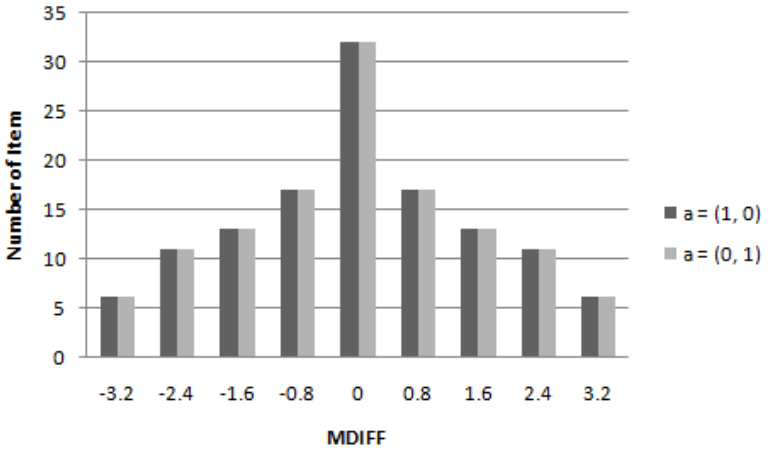


Figure 5.5: Item distribution for the .*86*-optimal item pool with exposure control (Two-dimension simple structure, high correlation)

5.2 Performance of the *p*-Optimal Item Pools

The previous section described the characteristics of the *p*-optimal item pools and how the characteristics change with the MCAT design (including bin size, test specification, correlation, and exposure control). In this section, the performance of the MCAT using the *p*-optimal item pools is evaluated based on the simulation results. Two questions are addressed: (1) how does the performance of the MCAT using *p*-optimal item pools compared with the MCAT using

baseline pools? and (2) how does the MCAT designs influence the performance of MCAT using the *p*-optimal item pools? The simulation results for Test Specification 1 (two-dimension simple structure) are first presented in 5.2.1 (for high correlation condition) and 5.2.2 (for moderate correlation condition), followed by Test Specification 2 (three-dimension simple structure) in 5.2.3 (for high correlation condition) and 5.2.4 (for moderate correlation condition), and Test Specification 3 (three-dimension non-simple structure) in 5.2.5 (for high correlation condition) and 5.2.6 (for moderate correlation condition).

5.2.1 Performance for item pools based on Test Specification 1 (high correlation)

Table 5.5 and 5.6 presents the results of the ability estimation and item pool utilization for the *.96*-optimal item pool, the *.86*-optimal item pool, and the baseline pool based on the condition of two-dimension simple structure test specification with θ_1 and θ_2 are highly correlated. The results in Table 5.5 are under the condition without item exposure control; and Table 5.6 is with item exposure control. In both tables, there are two values for bias, RMSE and correlation, representing the results for (θ_1, θ_2).

Under the condition without item exposure control (see Table 5.5), the two *p*-optimal item pools and the baseline pool show no bias on the $\boldsymbol{\theta}$ estimates. Also, the RMSE are all at 0.40, and the correlations between estimated $\boldsymbol{\theta}$ and true $\boldsymbol{\theta}$ are around 0.91. The average test information is also very similar among the three item pools. The amount of information on the direction of θ_1 and θ_2 is around 3.59. This value is very high for the MCAT in this study, because 15 items from each cluster are administered and the maximum amount of information an item can provide is 0.25. Because of the feature of simple structure, the off-diagonal values of the information matrix are zero. In general, the results suggest that the *.96*- and *.86*-optimal item pool can provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as the baseline pool.

Table 5.5: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control

(2-dimension simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
RMSE	(0.40, 0.40)	(0.40, 0.40)	(0.40, 0.40)
Correlation	(0.91, 0.91)	(0.91, 0.91)	(0.92, 0.91)
Average test information	$\begin{bmatrix} 3.59 & 0 \\ 0 & 3.60 \end{bmatrix}$	$\begin{bmatrix} 3.58 & 0 \\ 0 & 3.59 \end{bmatrix}$	$\begin{bmatrix} 3.59 & 0 \\ 0 & 3.60 \end{bmatrix}$
Overall Pool Usage	29.03	32.31	60.92
Overlap rate	0.16	0.30	0.19
% of overexposed item (r > 0.2)	11%	34%	9%
% of underexposed item (r < 0.02)	35%	33%	54%

Table 5.6: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control

(2-dimension simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.00, 0.00)	(0.00, 0.00)	(0.00, 0.00)
RMSE	(0.41, 0.41)	(0.41, 0.41)	(0.41, 0.42)
Correlation	(0.91, 0.91)	(0.91, 0.91)	(0.91, 0.91)
Average test information	$\begin{bmatrix} 3.34 & 0 \\ 0 & 3.35 \end{bmatrix}$	$\begin{bmatrix} 3.28 & 0 \\ 0 & 3.28 \end{bmatrix}$	$\begin{bmatrix} 3.30 & 0 \\ 0 & 3.10 \end{bmatrix}$
Overall Pool Usage	5.02	2.19	13.38
Overlap rate	0.09	0.13	0.09
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r < 0.02)	6%	0%	26%

Table 5.5 also presents the results about item pool usage. The overall pool usage index for the *.96*-optimal item pool is slightly smaller than that of the *.86*-optimal item pool, and the index for the baseline pool is about twice as much as the *.96*- and *.86*-optimal item pool. Because a small overall pool usage index implies more items in the item pool are fully used, the results

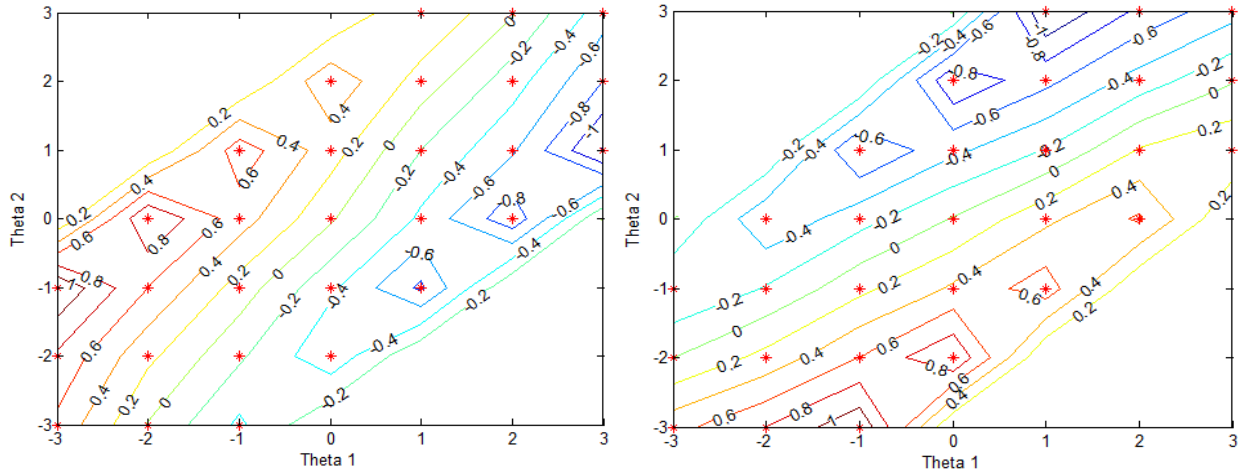
suggest that the .*96*-optimal item pool has slightly better usage than the .*86*-optimal item pool, and the two *p*-optimal item pools have much better usage than the baseline pool. More specifically, for the .*96*-optimal item pool, the overlap rate is 0.16, indicating that two randomly selected examinees will receive about 16% of items in common; and the percentage of overexposed and underexposed item are 11% and 35%, respectively. For the .*86*-optimal item pool, the results are: 30% of items overlap, 34% overexposed, and 33 % under exposed. Because more items from the .*86*-optimal item pool are overlapped and overexposed, the .*86*-optimal item pool is less secure than the .*96*-optimal item pool. This finding is reasonable because the size of the .*86*-optimal item pool is only 206 items, but the .*96*-optimal item pool has 369 items. The overlap rate for the baseline pool is 0.19, which is slightly higher than the .*96*-optimal item pool and lower than the .*86*-optimal item pool. Although a smaller number of items (9%) from the baseline pool are overexposed, more than half of the items (54%) are rarely used. It implies many items in the baseline pool are wasted. In brief, based on these pool usage results, the item pool usage for the .*96*- and .*86*-optimal item pool is much better than the baseline pool.

When item exposure control is implemented (see Table 5.6), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimation as the baseline pool, and yield better item pool usage than the baseline pool. Compared with the condition without item exposure control, item exposure control only results in a 0.01 to 0.02 increase for the RMSE, and about 0.3 decrease for the average test information. The reason why item exposure control rarely affects the ability estimation is because the *p*-optimal item pool design takes the item exposure rate into account and makes sure there is adequate number of items for selection. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed and overlapped items is also decreased. The .*86*-optimal item pool

has been fully used with no item underexposed in this condition. The overall pool usage index for the *.96*- and *.86*-optimal item pool and the baseline pool are 5.02, 2.19, and 13.38, respectively. The value is much smaller than the condition without item exposure control. Thus, item exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

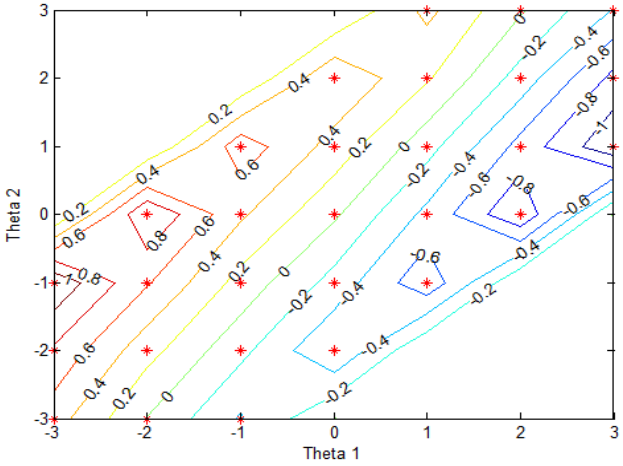
In addition to the overall performance, the conditional bias and RMSE at the 29 (θ_1, θ_2) points are also calculated in this study to evaluate the ability estimation at each $\boldsymbol{\theta}$ point. The results are presented by the contour plots. Each contour curve in the plot connects points with the same bias or RMSE value. The conditional bias for each $\boldsymbol{\theta}$ point is plotted in Figure 5.6 and 5.7, for the MCAT without and with item exposure control, respectively. In each Figure, the two plots (subplot a and b) at the top present the conditional bias for θ_1 and θ_2 for the .96-optimal item pool; the two plots (subplot c and d) in the middle present the conditional bias for the .86-optimal item pool; and the subplot e and f at the bottom present the conditional bias for the baseline pool. The conditional RMSE is plotted in Figure 5.8 and 5.9 in the same manner. The red points in the contour plot represent the 29 (θ_1, θ_2) points.

Under the condition without item exposure control (see Figure 5.6 for bias and 5.8 for RMSE), it is obvious that the plot for the .96-, .86-optimal item pool, and the baseline pool are very similar. This finding supports the results of the overall bias and RMSE, and also suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. In general, larger bias and RMSE occurs when θ_1 and θ_2 are very large or very small, which is the upper right corner and lower left corner in the contour plot. In addition to the value of the $\boldsymbol{\theta}$, the difference between θ_1 and θ_2 also affects the estimation accuracy. More specifically, when θ_1 is within (-1, 1) and θ_2 is near θ_1, the bias for θ_1 is close to 0 and the

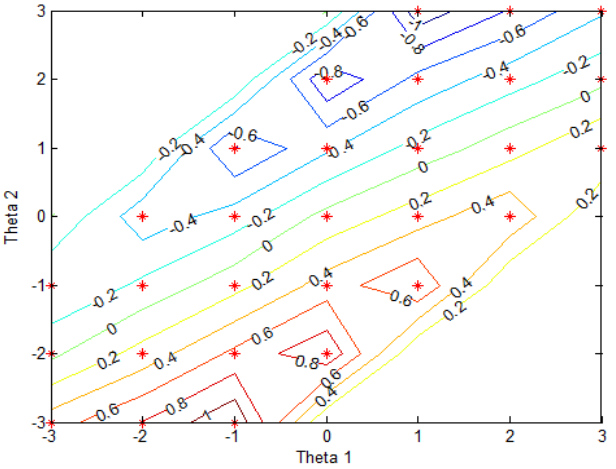


(a) Bias for θ_1 for the *.96*- optimal item pool

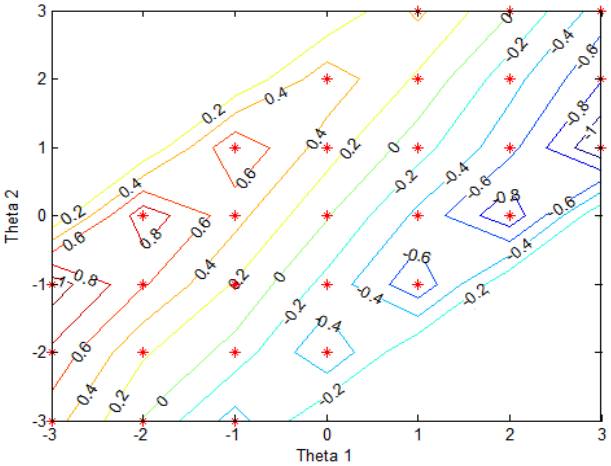
(b) Bias for θ_2 for the *.96*- optimal item pool



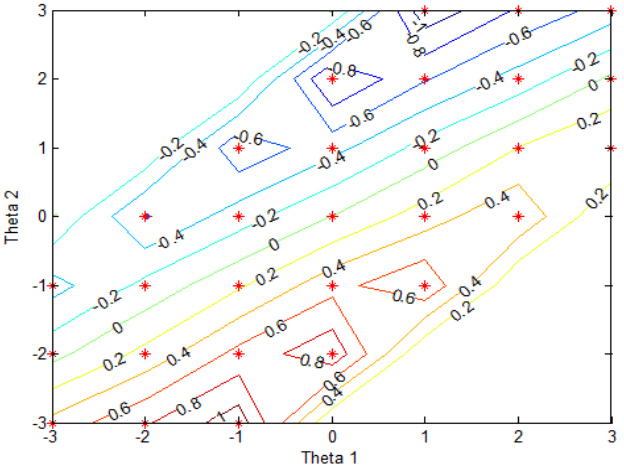
(c) Bias for θ_1 for the *.86*- optimal item pool



(d) Bias for θ_2 for the *.86*- optimal item pool

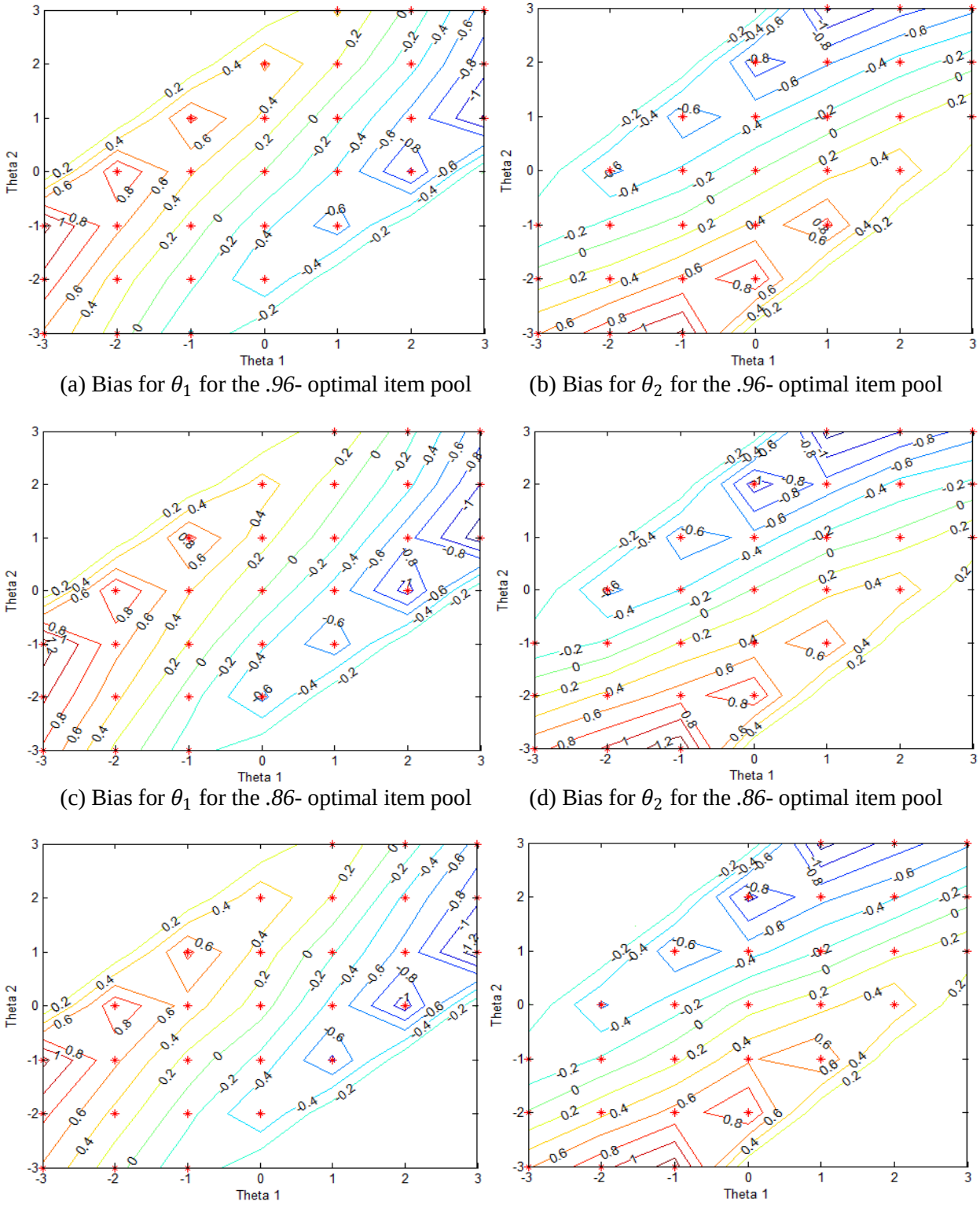


(e) Bias for θ_1 for the comparison pool



(f) Bias for θ_2 for the comparison pool

Figure 5.6: Conditional bias for the $\boldsymbol{\theta}$ estimates without exposure control (2-dimension simple structure, high correlation)



(a) Bias for θ_1 for the .*96*- optimal item pool

(b) Bias for θ_2 for the .*96*- optimal item pool

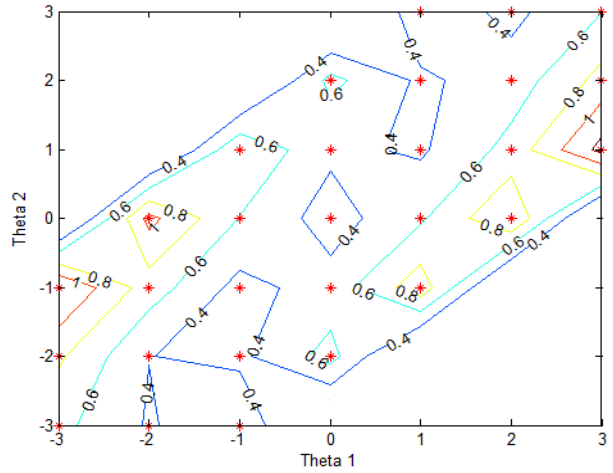
(c) Bias for θ_1 for the .*86*- optimal item pool

(d) Bias for θ_2 for the .*86*- optimal item pool

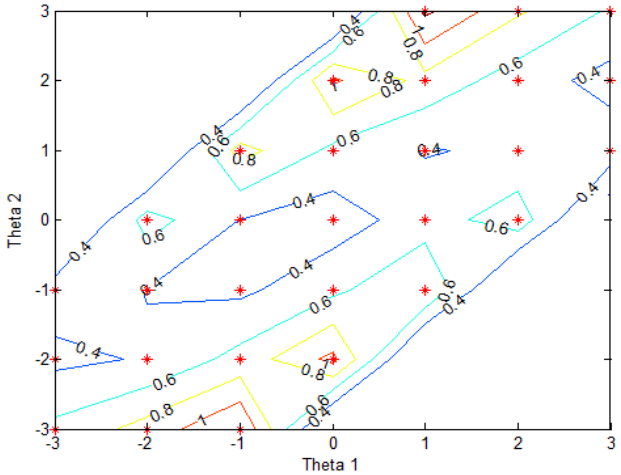
(e) Bias for θ_1 for the comparison pool

(f) Bias for θ_2 for the comparison pool

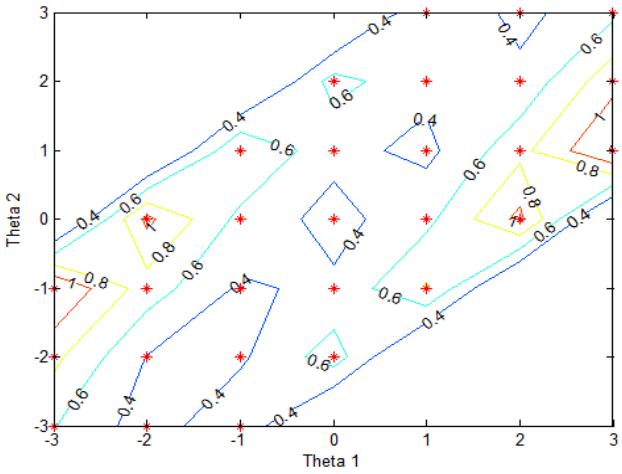
Figure 5.7: Conditional bias for the $\boldsymbol{\theta}$ estimates with exposure control
(2-dimension simple structure, high correlation)



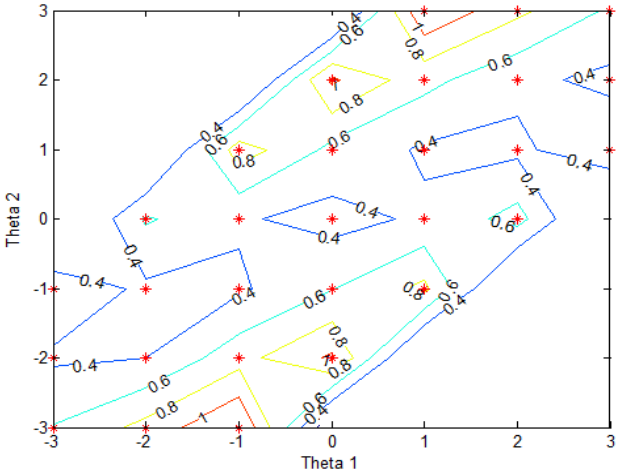
(a) RMSE for θ_1 for the *.96*- optimal item pool



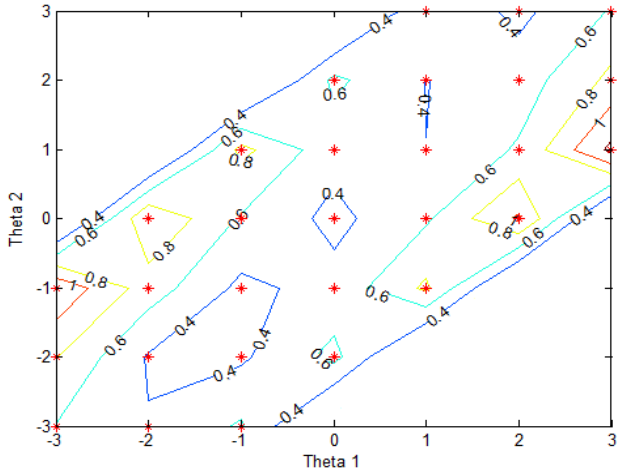
(b) RMSE for θ_2 for the *.96*- optimal item pool



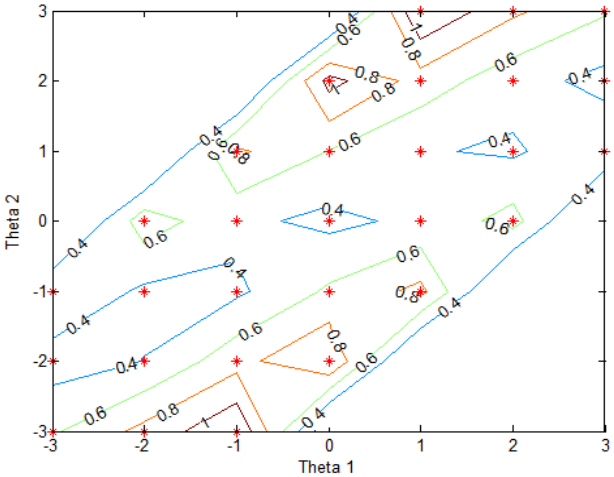
(c) RMSE for θ_1 for the *.86*- optimal item pool



(d) RMSE for θ_2 for the *.86*- optimal item pool

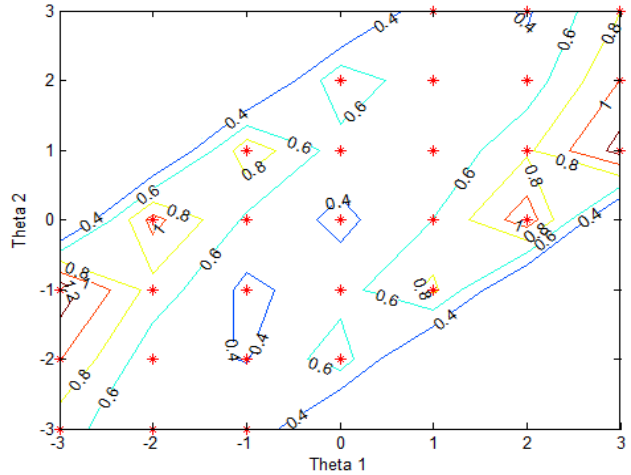


(e) RMSE for θ_1 for the comparison pool

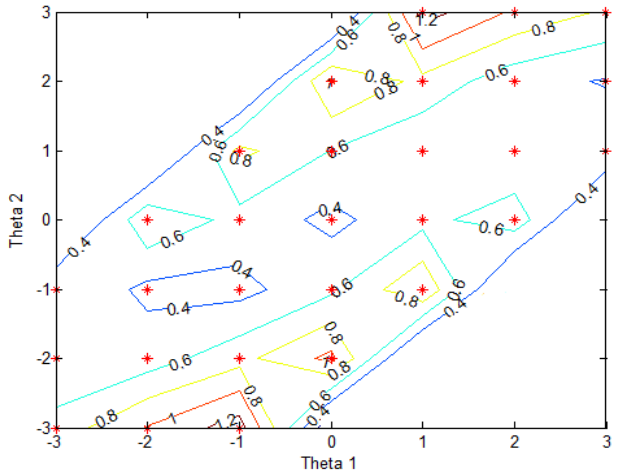


(f) RMSE for θ_2 for the comparison pool

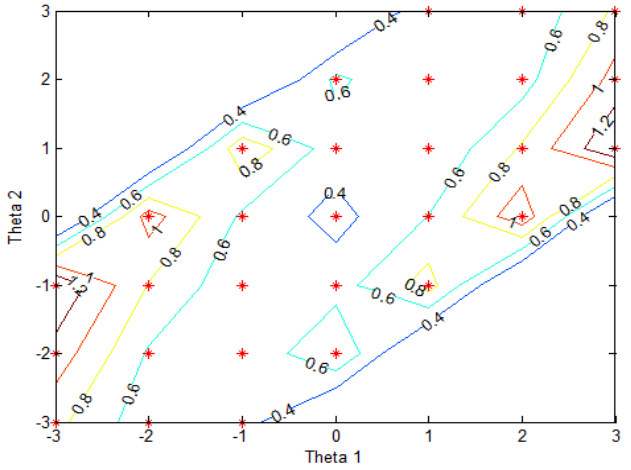
Figure 5.8: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (2-dimension simple structure, high correlation)



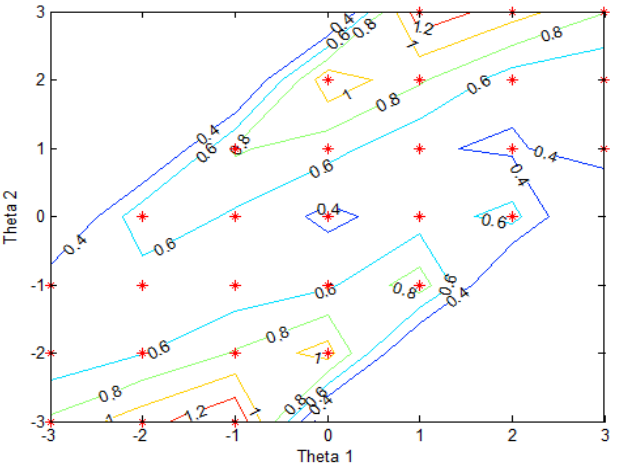
(a) RMSE for θ_1 for the *.96*- optimal item pool



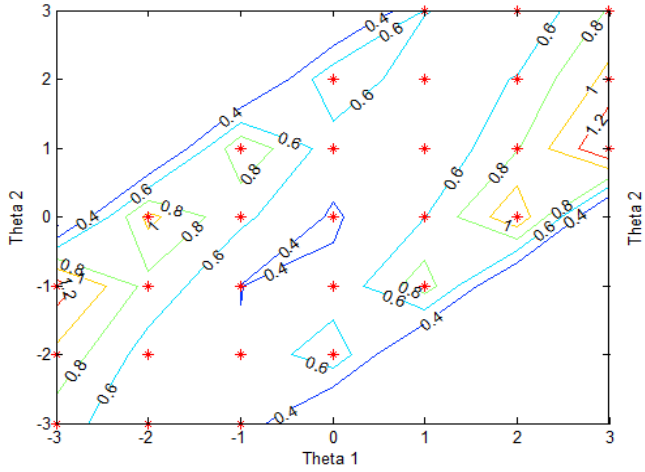
(b) RMSE for θ_2 for the *.96*- optimal item pool



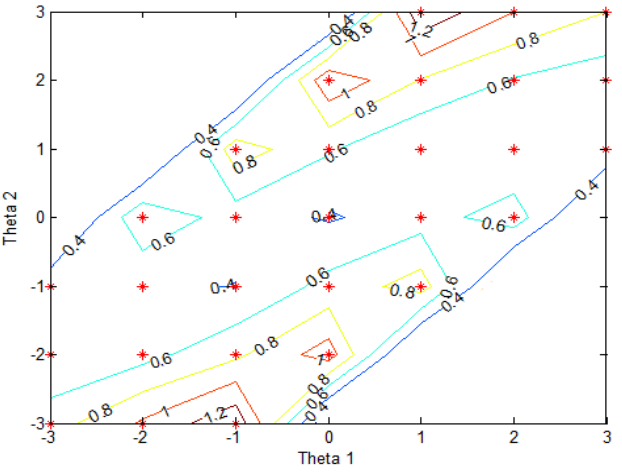
(c) RMSE for θ_1 for the *.86*- optimal item pool



(d) RMSE for θ_2 for the *.86*- optimal item pool



(e) RMSE for θ_1 for the comparison pool



(f) RMSE for θ_2 for the comparison pool

Figure 5.9: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (2-dimension simple structure, high correlation)

RMSE is less than 0.4. Negative bias and large RMSE appear when the value of θ_1 increases and the difference between θ_1 and θ_2 increases. For example, at point (3, 1) and (3, 2) in the plot, the bias for θ_1 is about -1.0 and RMSE for θ_1 is about 1.0. Meanwhile, positive bias and large RMSE appear when the value of θ_1 decreases and the difference between θ_1 and θ_2 increases. At point (-3, -1) and (-3, -2), the bias for θ_1 is about 1.0 and RMSE for θ_1 is about 1.0. Similar results for θ_2 can be observed from the right panel of Figure 5.6 and 5.8. When θ_2 is within (-1, 1) and θ_1 is near θ_2, the bias and RMSE for θ_2 is very small. When the value of θ_2 becomes more extreme and θ_1 is away from θ_2, large bias and RMSE values appear. This finding is probably due to the Bayesian MAP estimation method. As described in Chapter 2, the Bayesian method set the distribution of the true $\boldsymbol{\theta}$ as the prior. In this condition, the true $\boldsymbol{\theta}$ has a mean vector of (0, 0) and a high correlation between θ_1 and θ_2. The prior will shrink the ability estimation into the middle and reduce the difference between θ_1 and θ_2. In this study, the overall test length is 30 so that about 15 items are selected from each cluster. The effect of the likelihood function is probably not strong enough to overcome the effect of the prior. If the test length further increases, the effect of the likelihood function will dominate the effect of the prior eventually, and therefore reduce the bias and RMSE in those extreme cases.

When item exposure control is implemented, similar findings can be observed from Figure 5.7 and 5.9. Again, there is nearly no difference between the two *p*-optimal item pools, and between the *p*-optimal item pools and the baseline pool. The results support the finding based on the overall bias and RMSE, and further suggest the MCAT using the three item pools perform similarly in terms of the ability estimation on the 29 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1 and θ_2 are very large or very small, and when θ_1 and θ_2 are away from each other. A comparison between the condition with and without item exposure control shows,

when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger. The increase of estimation error is due to the item exposure control. Because the item exposure control prevents the most informative item from being frequently selected, the information available for ability estimation reduces slightly. When information reduces, the prior plays a more important role in the ability estimation. Thus, the measurement error at extreme $\boldsymbol{\theta}$ points becomes larger if item exposure control is added into the item selection process.

In summary, this section presents the results for the MCAT with the test specification of two-dimension simple structure and with high correlation between θ_1 and θ_2. In general, the *p*-optimal item pools perform similar as the baseline pool in terms of both overall and conditional accuracy of ability estimation, but the *p*-optimal item pools can save over 100 items and have a better item pool usage. When item exposure control is implemented, the item exposure rate and item overlap rate can be controlled very well. The *p*-optimal item pools still can provide reliable ability estimation with a relatively small pool size.

5.2.2 Performance for item pools based on Test Specification 1 (moderate correlation)

The results for the MCAT with the same test specification, but with θ_1 and θ_2 are moderately correlated, are presented in Table 5.7 and 5.8. The results in Table 5.7 are under the condition without item exposure control; and Table 5.8 is with item exposure control. In both tables, there are two values for bias, RMSE and correlation, representing the results for (θ_1, θ_2).

Under the condition without item exposure control (see Table 5.7), the *p*-optimal item pools and the baseline pool show nearly no bias on the $\boldsymbol{\theta}$ estimation. Also, the RMSE are all at 0.46, and correlations between estimated $\boldsymbol{\theta}$ and true $\boldsymbol{\theta}$ are around 0.88. The average test information is also very similar among the three item pools. The amount of information on the direction of

Table 5.7: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control
(2-dimension simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(-0.01, 0.00)	(-0.01, 0.00)	(-0.01, 0.01)
RMSE	(0.45, 0.46)	(0.45, 0.46)	(0.45, 0.46)
Correlation	(0.89, 0.89)	(0.89, 0.89)	(0.89, 0.89)
Average test information	$\begin{bmatrix} 3.58 & 0 \\ 0 & 3.58 \end{bmatrix}$	$\begin{bmatrix} 3.57 & 0 \\ 0 & 3.58 \end{bmatrix}$	$\begin{bmatrix} 3.58 & 0 \\ 0 & 3.58 \end{bmatrix}$
Overall Pool Usage	28.47	31.69	66.65
Overlap rate	0.18	0.32	0.20
% of overexposed item (r > 0.2)	16%	34%	10%
% of underexposed item (r < 0.02)	32%	29%	55%

Table 5.8: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control
(2-dimension simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(-0.01, 0.00)	(-0.01, 0.00)	(-0.01, 0.00)
RMSE	(0.46, 0.47)	(0.47, 0.47)	(0.47, 0.48)
Correlation	(0.88, 0.88)	(0.88, 0.88)	(0.88, 0.87)
Average test information	$\begin{bmatrix} 3.31 & 0 \\ 0 & 3.31 \end{bmatrix}$	$\begin{bmatrix} 3.27 & 0 \\ 0 & 3.29 \end{bmatrix}$	$\begin{bmatrix} 3.29 & 0 \\ 0 & 3.02 \end{bmatrix}$
Overall Pool Usage	3.55	1.59	13.48
Overlap rate	0.10	0.13	0.09
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r < 0.02)	0%	0%	26%

θ_1 and $\theta_2$is around 3.58. In general, the results suggest that the *.96*- and *.86*-optimal item pool can provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as the baseline pool.

Table 5.7 also presents the results about item pool usage. The overall pool usage index for the *.96*-optimal item pool is slightly smaller than that of the *.86*-optimal item pool, and the index

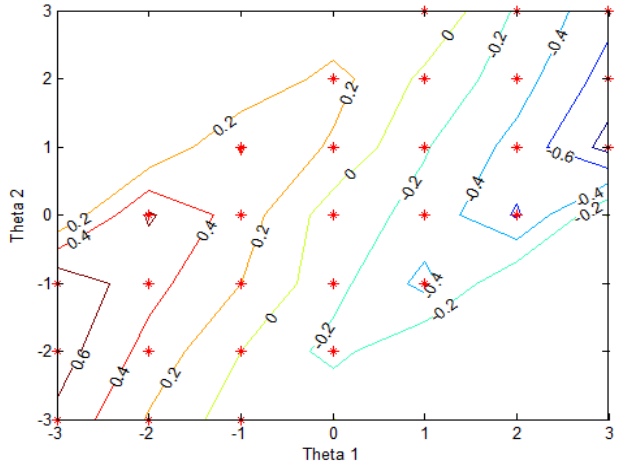
for the baseline pool is more than twice as much as the *.96*- and *.86*-optimal item pool. The results suggest that the *.96*-optimal item pool has slightly better usage than the *.86*-optimal item pool, and the two optimal item pools have much better usage than the baseline pool. More specifically, for the *.96*-optimal item pool, the overlap rate is 0.18, and the percentage of overexposed and underexposed item are 16% and 32%, respectively. For the *.86*-optimal item pool, the results are: 32% of items overlap, 34% overexposed, and 29 % under exposed. Because more items from the *.86*-optimal item pool are overlapped and overexposed, the *.86*-optimal item pool is less secure than the *.96*-optimal item pool. The overlap rate for the baseline pool is 0.20, which is slightly higher than the *.96*-optimal item pool and lower than the *.86*-optimal item pool. Although a smaller number of items (10%) from the baseline pool are overexposed, more than half of the items (55%) are rarely used. It implies many items in the baseline pool are wasted. In brief, based on these pool usage results, the item pool usage for the *.96*- and *.86*-optimal item pool is much better than the baseline pool.

When item exposure control is implemented (see Table 5.8), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimation as the baseline pool, and yield better item pool usage than the baseline pool. Compared with the condition without item exposure control, item exposure control only results in a 0.01 to 0.02 increase for the RMSE, and about 0.3 decrease for the average test information. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed item and overlapped item are also decreased. The *.96*- and the *.86*-optimal item pool has been fully used with no item underexposed. The overall pool usage index for the *.96*-, *.86*-optimal item pool and the baseline pool are 3.55, 1.59, and 13.48, respectively. The value is much smaller than the condition without item exposure control. Thus, the results suggest the item

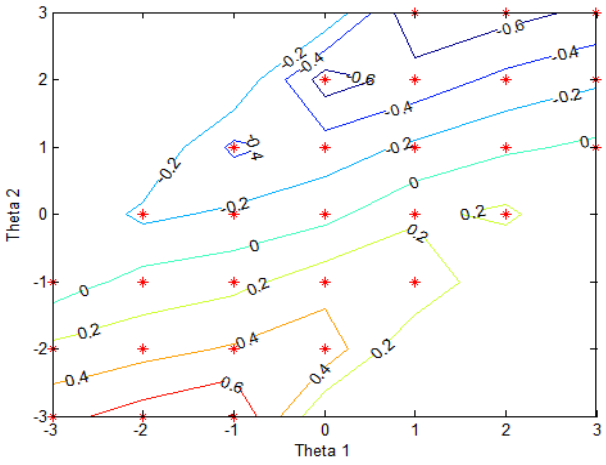
exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

In addition to the overall performance, the conditional bias and RMSE at the 29 (θ_1, θ_2) points are also calculated in this study to evaluate the ability estimation at each $\boldsymbol{\theta}$ point. The conditional bias for each $\boldsymbol{\theta}$ point is plotted in Figure 5.10 and 5.11, for the MCAT without and with item exposure control, respectively. The conditional RMSE is plotted in Figure 5.12 and 5.13.

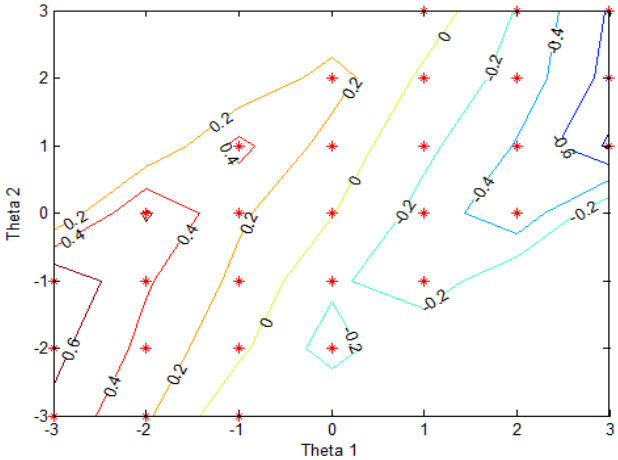
Under the condition without item exposure control (see Figure 5.10 for bias and 5.12 for RMSE), it is obvious that the plot for the .96-, .86-optimal item pool, and the baseline pool are very similar. This finding supports the results for the overall bias and RMSE, and also suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. Similar to the results in Section 5.1.1, larger bias and RMSE occurs when θ_1 and θ_2 are very large or very small, which is the upper right corner and lower left corner in the contour plot. In addition to the value of $\boldsymbol{\theta}$, the difference between θ_1 and θ_2 also affects the estimation accuracy. More specifically, when θ_1 is within (-1, 1) and θ_2 is near θ_1, the bias for θ_1 is close to 0 and the RMSE is less than 0.4. Negative bias and large RMSE appear when the value of θ_1 increases and the difference between θ_1 and θ_2 increases. For example, at point (3, 1) and (3, 2) in the plot, the bias for $\theta_1$is about -0.7 and RMSE for $\theta_1$is about 0.8. Meanwhile, positive bias and large RMSE appear when the value of θ_1 decreases and the difference between θ_1 and θ_2 increases. At point (-3, -1) and (-3, -2), the bias for $\theta_1$is about 0.7 and RMSE for $\theta_1$is about 0.8. Similar results for θ_2 can be observed from the right panel of Figure 5.10 and 5.12. When θ_2 is within (-1, 1) and θ_1 is near θ_2, the bias and RMSE for θ_2 is very small. When the value of θ_2 becomes more extreme and θ_1 is away from θ_2, large bias and RMSE values appear.



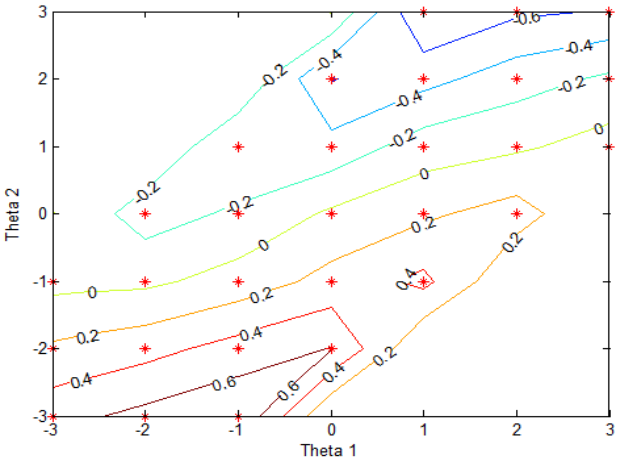
(a) Bias for θ_1 for the *.96*- optimal item pool



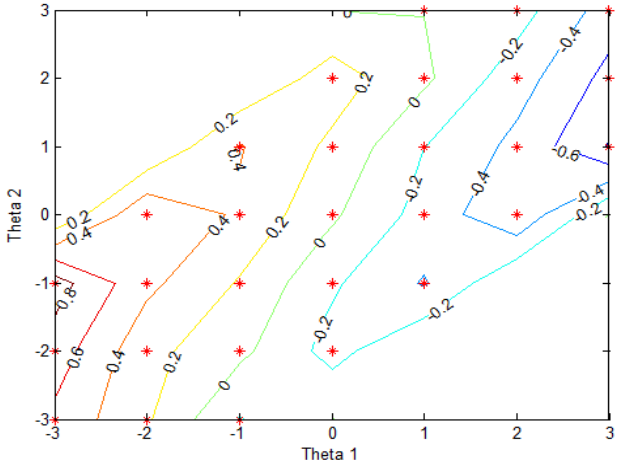
(b) Bias for θ_2 for the *.96*- optimal item pool



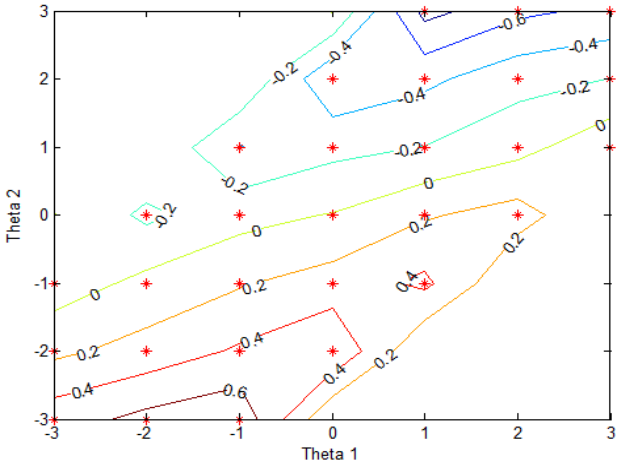
(c) Bias for θ_1 for the *.86*- optimal item pool



(d) Bias for θ_2 for the *.86*- optimal item pool

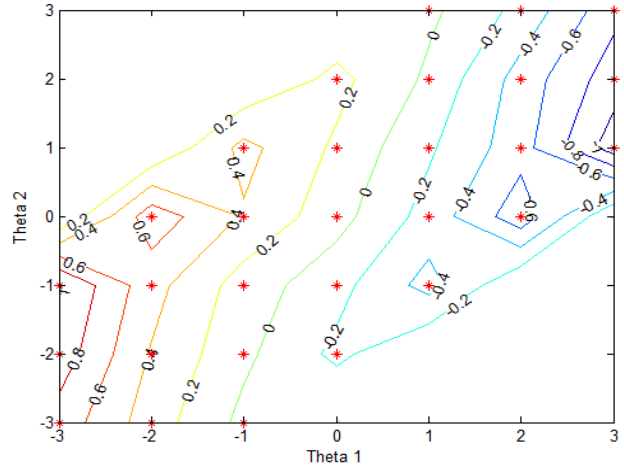


(e) Bias for θ_1 for the comparison pool

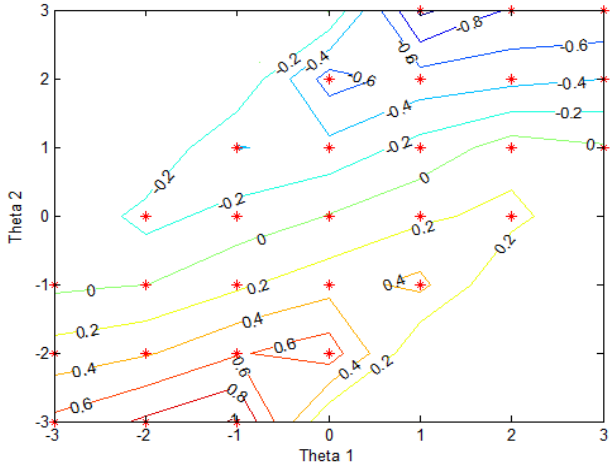


(f) Bias for θ_2 for the comparison pool

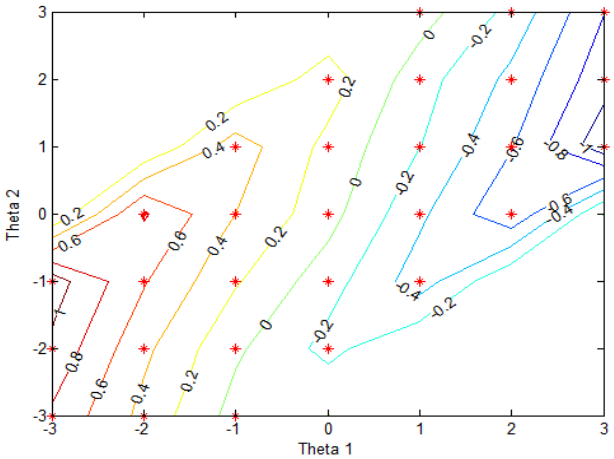
Figure 5.10: Conditional bias for the $\boldsymbol{\theta}$ estimates without exposure control (2-dimension simple structure, moderate correlation)



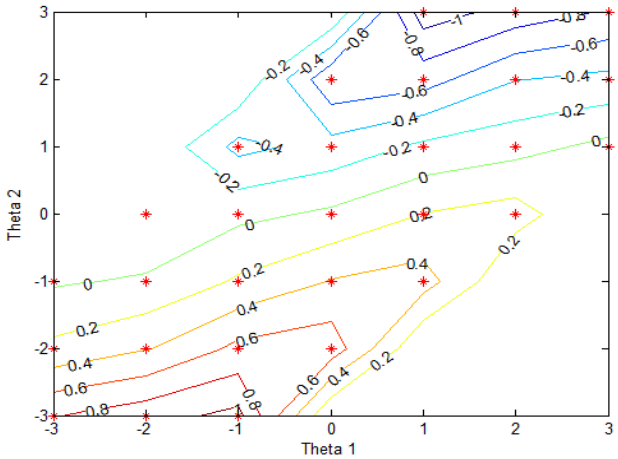
(a) Bias for θ_1 for the *.96*- optimal item pool



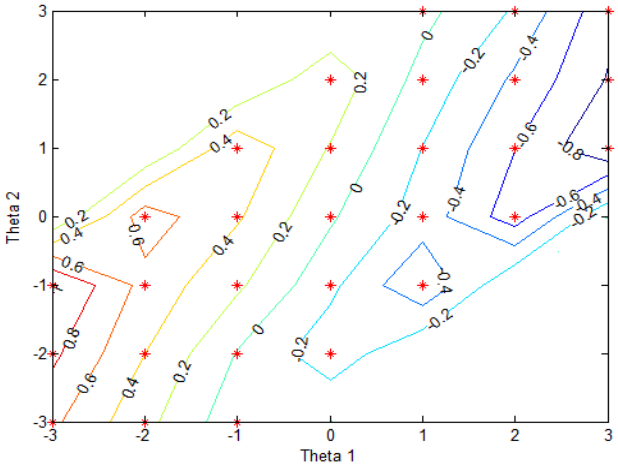
(b) Bias for θ_2 for the *.96*- optimal item pool



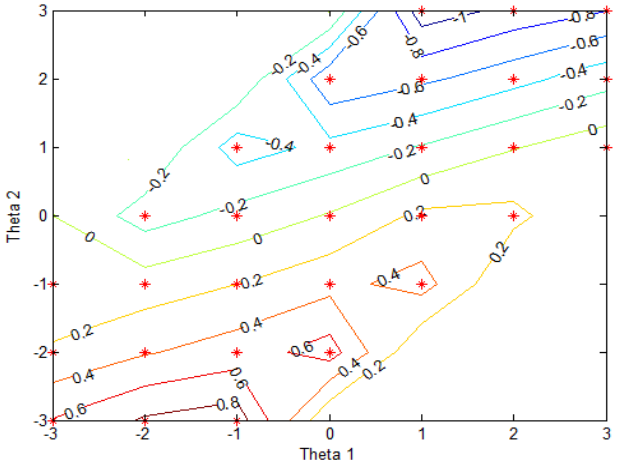
(c) Bias for θ_1 for the *.86*- optimal item pool



(d) Bias for θ_2 for the *.86*- optimal item pool

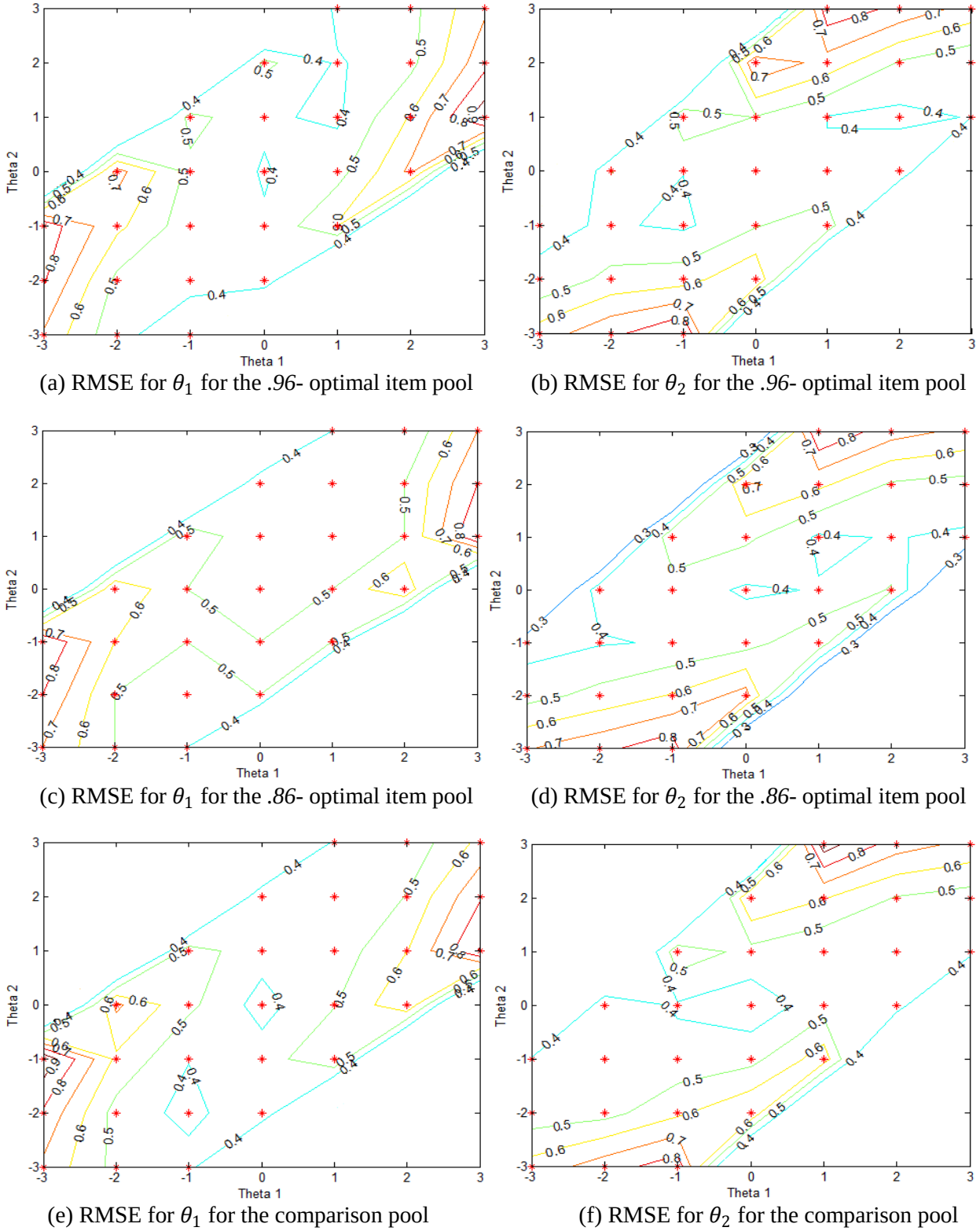


(e) Bias for θ_1 for the comparison pool



(f) Bias for θ_2 for the comparison pool

Figure 5.11: Conditional bias for the $\boldsymbol{\theta}$ estimates with exposure control
(2-dimension simple structure, moderate correlation)



(a) RMSE for θ_1 for the *.96*- optimal item pool

(b) RMSE for θ_2 for the *.96*- optimal item pool

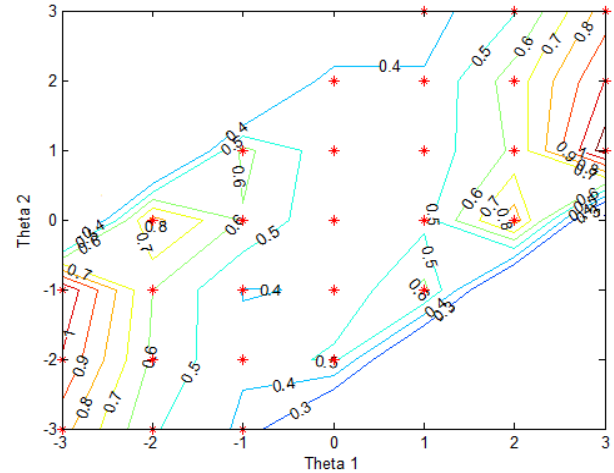
(c) RMSE for θ_1 for the *.86*- optimal item pool

(d) RMSE for θ_2 for the *.86*- optimal item pool

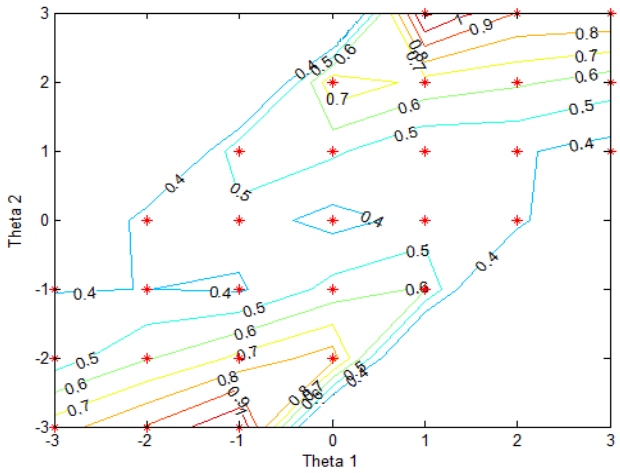
(e) RMSE for θ_1 for the comparison pool

(f) RMSE for θ_2 for the comparison pool

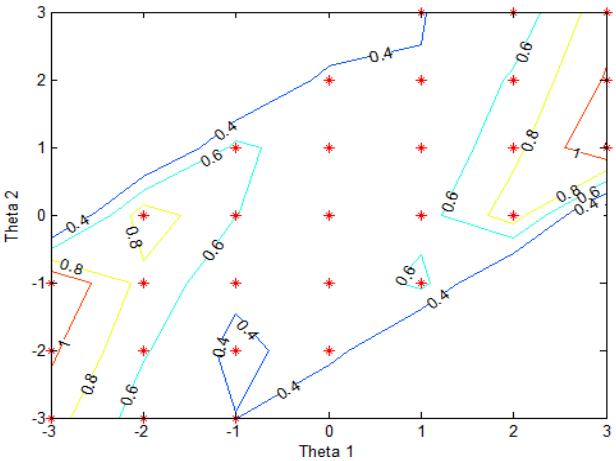
Figure 5.12: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (2-dimension simple structure, moderate correlation)



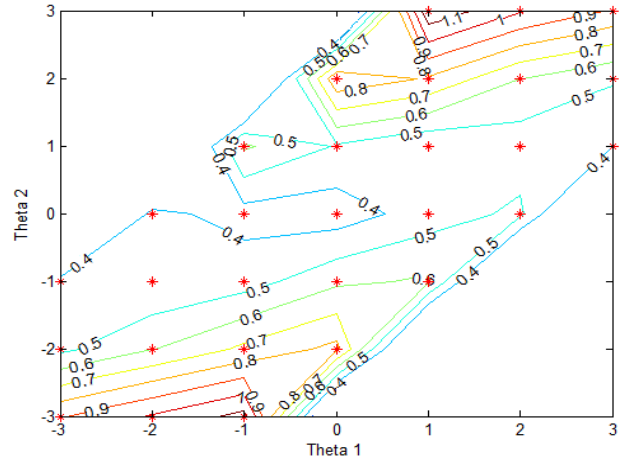
(a) RMSE for θ_1 for the *.96*- optimal item pool



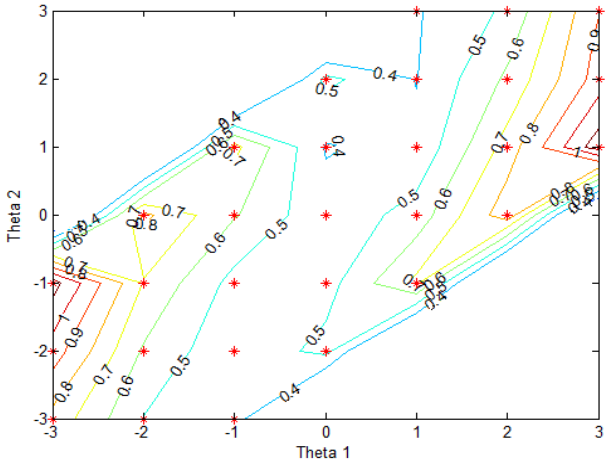
(b) RMSE for θ_2 for the *.96*- optimal item pool



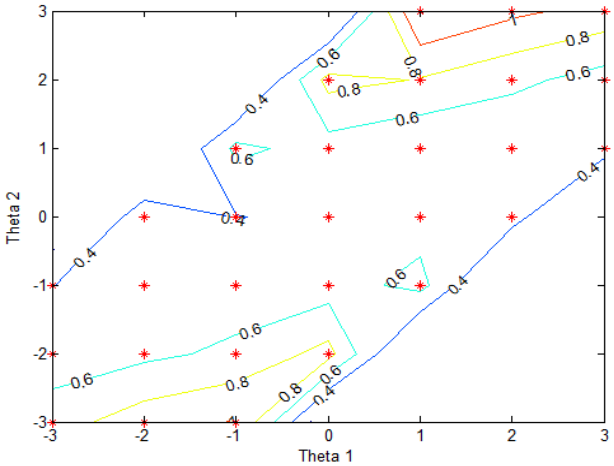
(c) RMSE for θ_1 for the *.86*- optimal item pool



(d) RMSE for θ_2 for the *.86*- optimal item pool



(e) RMSE for θ_1 for the comparison pool



(f) RMSE for θ_2 for the comparison pool

Figure 5.13: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (2-dimension simple structure, moderate correlation)

By comparing the contour plots in this section with the plots in Section 5.2.1 (when θ_1 and θ_2 are highly correlated), it is easy to see that the pattern of the contour plot is the same, but the magnitude of the bias and RMSE is smaller. When the correlation between θ_1 and θ_2 decreases, the prior weakly reduces the difference between θ_1 and θ_2. Therefore, when θ_1 and θ_2 are moderately correlated, the bias the RMSE values are slightly smaller at those points where θ_1 and θ_2 are away from each other, compared with the condition when θ_1 and θ_2 are highly correlated.

When item exposure control is implemented, similar findings can be observed from Figure 5.11 and 5.13. Again, there is nearly no difference between the two *p*-optimal item pools, and between the *p*-optimal item pools and the baseline pool. The results support the finding based on the overall bias and RMSE, and further suggest the three item pool performs similarly in terms of the ability estimation on the 29 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1 and θ_2 are very large or very small, and when θ_1 and θ_2 are away from each other. Similar to results under the high correlation condition in Section 5.2.1, when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger.

In summary, this section present the results for the MCAT with the test specification of two-dimension simple structure and with moderate correlation between θ_1 and θ_2. The *p*-optimal item pools perform similar as the baseline pool in terms of the accuracy of ability estimation, but the *p*-optimal item pools can save over 140 items and have a better item pool usage. When item exposure control is implemented, the *p*-optimal item pools still can provide accurate ability estimation and meanwhile the item exposure rate and item overlap rate can be well controlled.

In general, the findings from this section are similar to the finding in previous section. A close comparison between these two sections reveals that the measurement error in this section is

slightly larger. This result is due to the magnitude of the correlation between θ_1 and θ_2. Unlike the UIRT model estimating θ_1 and θ_2 one at a time, the MIRT model estimates θ_1 and θ_2 simultaneously, by borrowing information from one to another. When θ_1 and θ_2 are highly correlated, more variance in θ_1 can be explained by θ_2, so that more information can be borrowed for ability estimation. When the correlation between θ_1 and θ_2 decreases, the amount of information that can be borrowed reduces accordingly, and therefore the RMSE for $\boldsymbol{\theta}$ estimates increase. In addition to the accuracy of ability estimation, the pool usage for the two *p*-optimal item pool in this section is also slightly better. This is probably because of the pool size. When the correlation between θ_1 and θ_2 decreases, the pool size decreases as well. A smaller item pool is more likely to be fully used.

5.2.3 Performance for item pools based on Test Specification 2 (high correlation)

The results for the MCAT based on the three-dimension simple structure, and with θ_1 and θ_2 are highly correlated, are presented in Table 5.9 and 5.10. The results in Table 5.9 are under the condition without item exposure control; and Table 5.10 is with item exposure control. In both tables, there are three values for bias, RMSE and correlation, representing the results for (θ_1, θ_2, θ_3).

Under the condition without item exposure control (see Table 5.9), the *p*-optimal item pools and the baseline pool show nearly no bias on average. Also, the RMSE ranges from 0.41 to 0.46, and the correlations between the estimated $\boldsymbol{\theta}$ and the true $\boldsymbol{\theta}$ are around 0.90. The average test information is also very similar among the three item pools. The amount of information on the direction of θ_1, θ_2, and θ_3 is around 2.39. This value is very high for the three dimensional MCAT in this study, because only 10 items from each cluster are administered and the maximum amount of information an item can provide is 0.25. In general, the results suggest that the .*96*-

Table 5.9: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control
(3-dimension simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(-0.01, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)
RMSE	(0.44, 0.42, 0.45)	(0.44, 0.41, 0.45)	(0.44, 0.41, 0.46)
Correlation	(0.90, 0.91, 0.89)	(0.90, 0.91, 0.89)	(0.90, 0.91, 0.89)
Average test information	$\begin{bmatrix} 2.39 & 0 & 0 \\ 0 & 2.40 & 0 \\ 0 & 0 & 2.40 \end{bmatrix}$	$\begin{bmatrix} 2.38 & 0 & 0 \\ 0 & 2.39 & 0 \\ 0 & 0 & 2.39 \end{bmatrix}$	$\begin{bmatrix} 2.39 & 0 & 0 \\ 0 & 2.40 & 0 \\ 0 & 0 & 2.40 \end{bmatrix}$
Overall Pool Usage	28.47	31.69	66.65
Overlap rate	0.18	0.32	0.20
% of overexposed item (r > 0.2)	16%	34%	10%
% of underexposed item (r<0.02)	32%	29%	55%

Table 5.10: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control
(3-dimension simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(-0.01, 0.00, 0.01)	(0.00, 0.00, 0.01)	(0.00, 0.00, 0.01)
RMSE	(0.45, 0.43, 0.47)	(0.46, 0.44, 0.47)	(0.46, 0.43, 0.47)
Correlation	(0.89, 0.90, 0.89)	(0.89, 0.90, 0.88)	(0.89, 0.90, 0.89)
Average test information	$\begin{bmatrix} 2.17 & 0 & 0 \\ 0 & 2.18 & 0 \\ 0 & 0 & 2.17 \end{bmatrix}$	$\begin{bmatrix} 2.13 & 0 & 0 \\ 0 & 2.15 & 0 \\ 0 & 0 & 2.13 \end{bmatrix}$	$\begin{bmatrix} 2.08 & 0 & 0 \\ 0 & 2.03 & 0 \\ 0 & 0 & 2.15 \end{bmatrix}$
Overall Pool Usage	3.55	1.59	13.48
Overlap rate	0.10	0.13	0.09
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r<0.02)	0%	0%	26%

and *.86*-optimal item pool provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as baseline pool.

Table 5.9 also presents the results about item pool usage. Compared with the MCAT based on the Test Specification 1 in 5.2.1 and 5.2.2, similar results can be drawn from Table 5.9. The

item pool usage for the *.96*-optimal item pool is slightly better than the *86*-optimal item pool. And the two *p*-optimal item pools are much better used than the baseline pool.

When item exposure control is implemented (see Table 5.10), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimation as the baseline pool, and yield better item pool usage than the baseline pool. Compared with the condition without item exposure control, item exposure control only results in a 0.01 to 0.03 increase for the RMSE, and about 0.3 decrease for the average test information. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed item and overlapped item are also decreased. The two *p*-optimal item pools have been fully used with no item underexposed. The comparison between the condition with and without item exposure control suggests the item exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

In additional to the overall performance, the conditional bias and RMSE at 37 $(\theta_1, \theta_2, \theta_3)$ points are also calculated in this study to evaluate the ability estimation at each $\boldsymbol{\theta}$ point. The 3-dimensional bias and RMSE cannot be plotted in a contour plot. The conditional bias for each $\boldsymbol{\theta}$ point is presented in Table 5.11 and 5.12, for the MCAT without and with item exposure control, respectively. In each table, the conditional bias is color coded based on the value. Negative bias is colored in blue and positive bias is in red. Deeper color represents larger bias. The conditional RMSE is presented in Table 5.13 and 5.14 in the same manner. Small RMSE is colored in green and large RMSE is colored in red.

Under the condition without item exposure control (see Table 5.11 for bias and 5.13 for RMSE), the conditional bias and RMSE for the .96-, .86-optimal item pool, and the baseline pool are quite similar. This finding supports the results for the overall bias and RMSE, and also

Table 5.11: Conditional Bias for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.48	0.52	0.56	0.41	0.44	0.50	0.53	0.51	0.56
-3	-3	-2	0.65	0.60	0.69	0.66	0.61	0.68	-0.06	-0.07	-0.05
-3	-2	-3	0.82	0.80	0.76	-0.13	-0.13	-0.19	0.81	0.79	0.74
-3	-2	-2	0.94	0.91	0.93	0.06	0.03	0.08	0.20	0.16	0.21
-2	-3	-3	-0.07	-0.02	-0.11	0.72	0.77	0.72	0.72	0.72	0.72
-2	-3	-2	0.14	0.02	0.05	0.96	0.86	0.89	0.12	0.08	0.10
-2	-2	-3	0.16	0.15	0.10	0.07	0.05	0.01	0.93	0.90	0.89
-2	-2	-2	0.29	0.30	0.28	0.23	0.27	0.26	0.26	0.33	0.27
-2	-2	-1	0.45	0.46	0.46	0.48	0.52	0.49	-0.23	-0.18	-0.24
-2	-1	-2	0.56	0.54	0.62	-0.36	-0.39	-0.27	0.52	0.54	0.60
-2	-1	-1	0.72	0.76	0.72	-0.12	-0.07	-0.12	-0.01	0.07	-0.01
-1	-2	-2	-0.19	-0.23	-0.24	0.58	0.56	0.49	0.50	0.49	0.45
-1	-2	-1	-0.09	-0.05	-0.09	0.73	0.79	0.75	-0.12	-0.03	-0.06
-1	0	-1	0.32	0.42	0.38	-0.58	-0.47	-0.51	0.29	0.36	0.31
-1	0	0	0.58	0.60	0.57	-0.25	-0.21	-0.25	-0.16	-0.12	-0.14
0	-1	-1	-0.38	-0.41	-0.40	0.42	0.37	0.42	0.32	0.30	0.32
0	-1	0	-0.25	-0.27	-0.33	0.61	0.62	0.55	-0.27	-0.20	-0.26
0	0	-1	-0.18	-0.16	-0.15	-0.26	-0.22	-0.22	0.52	0.55	0.50
0	0	0	0.07	0.01	-0.03	0.05	0.02	-0.03	0.03	0.03	-0.02
0	0	1	0.12	0.20	0.15	0.16	0.29	0.22	-0.58	-0.49	-0.52
0	1	0	0.24	0.27	0.26	-0.61	-0.59	-0.62	0.24	0.25	0.19
0	1	1	0.47	0.48	0.39	-0.32	-0.33	-0.38	-0.28	-0.29	-0.26
1	0	0	-0.56	-0.58	-0.51	0.26	0.26	0.30	0.16	0.16	0.17
1	0	1	-0.42	-0.38	-0.41	0.48	0.53	0.48	-0.35	-0.32	-0.35
1	2	1	0.09	0.08	0.10	-0.77	-0.75	-0.76	0.06	0.05	0.03
1	2	2	0.29	0.27	0.25	-0.52	-0.50	-0.54	-0.46	-0.47	-0.45
2	1	1	-0.73	-0.67	-0.73	0.08	0.14	0.11	-0.06	0.00	0.01
2	1	2	-0.54	-0.54	-0.56	0.34	0.30	0.33	-0.56	-0.62	-0.58
2	2	1	-0.46	-0.50	-0.46	-0.48	-0.51	-0.52	0.25	0.20	0.19
2	2	2	-0.32	-0.30	-0.30	-0.28	-0.27	-0.27	-0.32	-0.33	-0.33
2	2	3	-0.15	-0.13	-0.15	-0.05	-0.04	-0.07	-0.91	-0.87	-0.92
2	3	2	-0.11	-0.09	-0.07	-0.96	-0.92	-0.91	-0.17	-0.10	-0.08
2	3	3	0.04	0.05	0.11	-0.72	-0.72	-0.68	-0.69	-0.66	-0.67
3	2	2	-0.95	-0.91	-1.00	-0.08	-0.03	-0.11	-0.25	-0.20	-0.21
3	2	3	-0.72	-0.74	-0.75	0.25	0.18	0.20	-0.66	-0.72	-0.76
3	3	2	-0.64	-0.61	-0.68	-0.65	-0.62	-0.68	0.08	0.10	-0.03
3	3	3	-0.49	-0.57	-0.48	-0.40	-0.50	-0.43	-0.48	-0.57	-0.47

Note: .96 represents .*96*-optimal pool; .86 represents .*86*-optimal pool; C represents baseline pool

Table 5.12: Conditional Bias for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.65	0.79	0.71	0.61	0.72	0.64	0.69	0.80	0.69
-3	-3	-2	0.78	0.94	0.82	0.78	0.94	0.81	0.04	0.14	0.05
-3	-2	-3	0.84	1.03	0.82	-0.12	0.07	-0.12	0.86	1.02	0.87
-3	-2	-2	1.02	1.11	0.93	0.12	0.19	0.04	0.22	0.31	0.20
-2	-3	-3	-0.04	0.02	0.07	0.79	0.86	0.89	0.80	0.89	0.88
-2	-3	-2	0.13	0.24	0.16	1.00	1.14	1.03	0.14	0.28	0.19
-2	-2	-3	0.19	0.24	0.32	0.11	0.16	0.21	1.01	1.09	1.10
-2	-2	-2	0.35	0.45	0.35	0.32	0.40	0.33	0.36	0.45	0.38
-2	-2	-1	0.54	0.60	0.52	0.57	0.61	0.51	-0.21	-0.14	-0.24
-2	-1	-2	0.65	0.60	0.63	-0.29	-0.34	-0.28	0.60	0.59	0.64
-2	-1	-1	0.79	0.84	0.79	-0.10	-0.02	-0.06	-0.04	0.05	0.06
-1	-2	-2	-0.23	-0.21	-0.21	0.56	0.63	0.59	0.53	0.56	0.56
-1	-2	-1	-0.05	-0.12	-0.02	0.80	0.80	0.85	-0.05	-0.02	0.03
-1	0	-1	0.47	0.34	0.44	-0.42	-0.54	-0.48	0.43	0.38	0.39
-1	0	0	0.64	0.64	0.65	-0.21	-0.18	-0.22	-0.15	-0.09	-0.14
0	-1	-1	-0.34	-0.46	-0.36	0.47	0.38	0.45	0.40	0.33	0.39
0	-1	0	-0.21	-0.22	-0.25	0.64	0.66	0.66	-0.21	-0.19	-0.18
0	0	-1	-0.09	-0.15	-0.20	-0.16	-0.21	-0.26	0.65	0.59	0.54
0	0	0	-0.07	-0.06	-0.08	-0.04	-0.06	-0.10	-0.02	-0.05	-0.08
0	0	1	0.17	0.12	0.24	0.22	0.16	0.32	-0.58	-0.60	-0.46
0	1	0	0.26	0.23	0.27	-0.62	-0.66	-0.61	0.21	0.21	0.23
0	1	1	0.46	0.47	0.37	-0.35	-0.36	-0.44	-0.31	-0.31	-0.36
1	0	0	-0.56	-0.64	-0.61	0.27	0.23	0.24	0.21	0.14	0.16
1	0	1	-0.45	-0.48	-0.42	0.44	0.43	0.48	-0.41	-0.45	-0.38
1	2	1	0.00	0.06	0.04	-0.85	-0.81	-0.84	0.00	0.04	0.01
1	2	2	0.24	0.26	0.20	-0.58	-0.57	-0.60	-0.58	-0.56	-0.55
2	1	1	-0.73	-0.86	-0.82	0.14	0.04	0.06	0.02	-0.01	-0.03
2	1	2	-0.68	-0.70	-0.67	0.26	0.25	0.25	-0.61	-0.68	-0.66
2	2	1	-0.54	-0.63	-0.54	-0.57	-0.67	-0.55	0.16	0.07	0.23
2	2	2	-0.35	-0.47	-0.35	-0.32	-0.44	-0.30	-0.38	-0.50	-0.36
2	2	3	-0.21	-0.32	-0.25	-0.13	-0.24	-0.17	-1.03	-1.13	-1.09
2	3	2	-0.18	-0.24	-0.16	-1.04	-1.11	-1.05	-0.19	-0.27	-0.15
2	3	3	0.01	-0.11	-0.04	-0.80	-0.92	-0.88	-0.82	-0.93	-0.90
3	2	2	-1.02	-1.12	-1.06	-0.10	-0.20	-0.15	-0.21	-0.31	-0.27
3	2	3	-0.91	-1.05	-0.92	0.06	-0.07	0.06	-0.90	-1.04	-0.92
3	3	2	-0.76	-0.94	-0.84	-0.74	-0.94	-0.82	0.05	-0.14	-0.05
3	3	3	-0.65	-0.84	-0.66	-0.60	-0.77	-0.62	-0.71	-0.84	-0.71

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.13: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.58	0.63	0.68	0.52	0.57	0.62	0.63	0.63	0.66
-3	-3	-2	0.72	0.71	0.77	0.74	0.70	0.77	0.36	0.34	0.41
-3	-2	-3	0.89	0.91	0.83	0.39	0.43	0.37	0.89	0.89	0.80
-3	-2	-2	1.02	0.98	1.00	0.38	0.35	0.35	0.44	0.38	0.41
-2	-3	-3	0.42	0.37	0.38	0.82	0.85	0.79	0.79	0.81	0.79
-2	-3	-2	0.38	0.39	0.35	1.01	0.95	0.94	0.36	0.45	0.34
-2	-2	-3	0.39	0.40	0.38	0.33	0.33	0.35	0.99	0.96	0.96
-2	-2	-2	0.47	0.44	0.45	0.42	0.42	0.41	0.47	0.50	0.43
-2	-2	-1	0.56	0.58	0.55	0.57	0.63	0.57	0.40	0.40	0.41
-2	-1	-2	0.66	0.65	0.71	0.51	0.55	0.46	0.62	0.67	0.72
-2	-1	-1	0.80	0.81	0.82	0.39	0.31	0.41	0.38	0.33	0.41
-1	-2	-2	0.43	0.38	0.44	0.68	0.63	0.59	0.62	0.59	0.57
-1	-2	-1	0.38	0.37	0.35	0.81	0.87	0.82	0.40	0.32	0.37
-1	0	-1	0.46	0.56	0.51	0.67	0.59	0.63	0.47	0.52	0.48
-1	0	0	0.67	0.69	0.67	0.41	0.41	0.45	0.37	0.38	0.39
0	-1	-1	0.50	0.56	0.53	0.53	0.50	0.54	0.49	0.48	0.48
0	-1	0	0.42	0.43	0.48	0.69	0.72	0.64	0.44	0.42	0.44
0	0	-1	0.38	0.39	0.42	0.40	0.39	0.43	0.59	0.65	0.62
0	0	0	0.37	0.37	0.36	0.37	0.35	0.32	0.40	0.37	0.33
0	0	1	0.40	0.45	0.40	0.37	0.51	0.42	0.67	0.64	0.62
0	1	0	0.43	0.44	0.43	0.71	0.68	0.71	0.46	0.43	0.41
0	1	1	0.59	0.61	0.52	0.46	0.49	0.51	0.45	0.48	0.47
1	0	0	0.68	0.66	0.59	0.47	0.40	0.42	0.43	0.36	0.38
1	0	1	0.53	0.53	0.56	0.58	0.64	0.60	0.51	0.47	0.50
1	2	1	0.35	0.34	0.37	0.84	0.81	0.83	0.34	0.31	0.38
1	2	2	0.47	0.46	0.41	0.62	0.62	0.63	0.59	0.59	0.56
2	1	1	0.82	0.75	0.81	0.38	0.38	0.36	0.38	0.39	0.37
2	1	2	0.64	0.66	0.66	0.47	0.48	0.47	0.63	0.72	0.68
2	2	1	0.58	0.61	0.58	0.58	0.63	0.61	0.42	0.43	0.40
2	2	2	0.48	0.45	0.49	0.44	0.44	0.47	0.46	0.48	0.52
2	2	3	0.40	0.38	0.36	0.37	0.34	0.34	0.97	0.93	0.98
2	3	2	0.40	0.41	0.34	1.03	0.99	0.97	0.42	0.40	0.36
2	3	3	0.36	0.40	0.37	0.81	0.81	0.77	0.78	0.76	0.75
3	2	2	1.02	0.99	1.06	0.35	0.38	0.38	0.43	0.44	0.42
3	2	3	0.81	0.82	0.84	0.45	0.40	0.44	0.75	0.80	0.83
3	3	2	0.75	0.70	0.76	0.74	0.70	0.76	0.36	0.37	0.41
3	3	3	0.61	0.72	0.60	0.55	0.64	0.54	0.63	0.68	0.58

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.14: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.74	0.90	0.84	0.71	0.84	0.79	0.79	0.91	0.82
-3	-3	-2	0.88	1.03	0.92	0.88	1.02	0.92	0.42	0.43	0.45
-3	-2	-3	0.91	1.11	0.93	0.37	0.39	0.46	0.93	1.10	0.98
-3	-2	-2	1.10	1.18	1.01	0.45	0.44	0.40	0.46	0.51	0.44
-2	-3	-3	0.39	0.43	0.45	0.87	0.97	1.00	0.88	0.99	0.99
-2	-3	-2	0.41	0.47	0.43	1.07	1.20	1.09	0.40	0.47	0.42
-2	-2	-3	0.42	0.45	0.50	0.40	0.42	0.43	1.08	1.15	1.15
-2	-2	-2	0.51	0.57	0.55	0.49	0.55	0.53	0.52	0.58	0.55
-2	-2	-1	0.65	0.72	0.64	0.68	0.74	0.62	0.44	0.44	0.41
-2	-1	-2	0.76	0.73	0.76	0.49	0.55	0.50	0.72	0.74	0.75
-2	-1	-1	0.87	0.94	0.89	0.36	0.41	0.41	0.40	0.39	0.42
-1	-2	-2	0.44	0.43	0.41	0.69	0.73	0.68	0.67	0.68	0.64
-1	-2	-1	0.35	0.40	0.36	0.87	0.89	0.92	0.34	0.37	0.38
-1	0	-1	0.59	0.51	0.62	0.56	0.67	0.65	0.59	0.53	0.58
-1	0	0	0.72	0.72	0.76	0.39	0.37	0.44	0.36	0.35	0.43
0	-1	-1	0.49	0.58	0.51	0.57	0.52	0.58	0.53	0.48	0.54
0	-1	0	0.45	0.45	0.46	0.75	0.76	0.76	0.41	0.40	0.40
0	0	-1	0.37	0.42	0.40	0.37	0.43	0.42	0.76	0.73	0.66
0	0	0	0.35	0.37	0.44	0.34	0.36	0.43	0.36	0.37	0.42
0	0	1	0.42	0.37	0.49	0.41	0.39	0.51	0.70	0.68	0.61
0	1	0	0.44	0.48	0.51	0.71	0.77	0.75	0.44	0.46	0.47
0	1	1	0.60	0.58	0.52	0.53	0.49	0.57	0.48	0.47	0.52
1	0	0	0.67	0.73	0.71	0.45	0.43	0.42	0.45	0.41	0.40
1	0	1	0.56	0.61	0.60	0.54	0.57	0.64	0.52	0.57	0.56
1	2	1	0.35	0.44	0.38	0.92	0.91	0.93	0.36	0.39	0.38
1	2	2	0.42	0.51	0.44	0.68	0.70	0.71	0.69	0.69	0.68
2	1	1	0.81	0.94	0.90	0.39	0.37	0.36	0.35	0.34	0.37
2	1	2	0.81	0.83	0.77	0.49	0.49	0.46	0.74	0.80	0.75
2	2	1	0.66	0.73	0.69	0.68	0.76	0.71	0.40	0.36	0.49
2	2	2	0.52	0.65	0.51	0.50	0.63	0.46	0.54	0.67	0.52
2	2	3	0.41	0.48	0.43	0.39	0.43	0.40	1.10	1.20	1.16
2	3	2	0.42	0.46	0.42	1.10	1.18	1.11	0.42	0.48	0.41
2	3	3	0.39	0.47	0.36	0.89	1.04	0.95	0.91	1.05	0.97
3	2	2	1.08	1.20	1.14	0.37	0.46	0.44	0.42	0.54	0.50
3	2	3	1.00	1.18	0.99	0.38	0.53	0.39	0.96	1.17	1.00
3	3	2	0.88	1.04	0.96	0.85	1.05	0.96	0.42	0.48	0.48
3	3	3	0.73	0.94	0.81	0.67	0.87	0.77	0.76	0.94	0.84

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. In general, larger bias and RMSE occurs when θ_1, θ_2, and θ_3 are very large or very small, which is the top and the bottom of each table. In addition to the value of $\boldsymbol{\theta}$, the difference between two $\theta's$ also affects the estimation accuracy. More specifically, when θ_1 is around 0, and θ_2 and θ_3 are near θ_1, the bias for θ_1 is close to 0 and the RMSE is less than 0.4. Negative bias and large RMSE appear when the value of θ_1 increases and the difference between θ_1 and θ_2, and between θ_1 and θ_3, increases. For example, at point (3, 2, 2) in the table, the bias for $\theta_1$is almost -1.0 and RMSE for $\theta_1$is around 1.0. Meanwhile, positive bias and large RMSE appear when the value of θ_1 decreases and the difference between θ_1 and θ_2, and between θ_1 and θ_3, increases. At point (-3, -2, -2), the bias for $\theta_1$is about 0.93 and RMSE for $\theta_1$is around 1.0. Similar results for θ_2 can be observed from the three columns in the middle of Table 5.11 and 5.13. When θ_2 is around 0, and θ_1 and θ_3 is near θ_2, the bias and RMSE for θ_2 is very small. When the value of θ_2 becomes more extreme and θ_1 and $\theta_3$is away from θ_2, large bias and RMSE values appear. Again, similar results can be found for θ_3 from the three columns on the right side of Table 5.11 and 5.13. As described in Section 5.2.1, this finding is probably due to the Bayesian MAP estimation method. The prior for $\boldsymbol{\theta}$ estimation is a multivariate normal distribution with a mean vector of (0, 0) and a high correlation among θ_1, θ_2, and θ_3. The prior will shrink the ability estimation into the middle and reduce the difference among each θ. Under this condition, the overall test length is 30 so that about 10 items are selected from each cluster. The effect of the likelihood function is relatively weak comparing to the effect of the prior. If the test length further increases, the effect of the likelihood function will dominate the effect of the prior eventually, and therefore reduce the bias and RMSE in those extreme cases.

When item exposure control is implemented, similar findings can be observed from Table 5.12 and 5.14. Again, there is nearly no difference between the two *p*-optimal item pools, and between the *p*-optimal item pools and the baseline pool. The results support the finding based on the overall bias and RMSE, and further suggest the three item pools perform similarly in terms of the ability estimation on the 37 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1, θ_2, and θ_3 are very large or very small, and when θ's are away from each other. A comparison between the condition with and without item exposure control shows, when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger. The increase of estimation error is due to the item exposure control. As explained in Section 5.2.1, because the item exposure control prevents the most informative item from being frequently selected, the information available for ability estimation reduces slightly. Thus, the measurement error at extreme $\boldsymbol{\theta}$ points becomes larger if item exposure control is built into the item selection process.

In summary, this section present the results for the MCAT with the test specification of three-dimension simple structure and with high correlation among θ_1, θ_2, and θ_3. In general, the *p*-optimal item pools perform similarly as the baseline pool in terms of both overall and conditional accuracy of ability estimation, but the *p*-optimal item pools can save about 100 items and have a better item pool usage. When item exposure control is implemented, the item exposure rate and item overlap rate can be controlled very well. The *p*-optimal item pools still can provide reliable ability estimation with a relatively small pool size.

5.2.4 Performance for item pools based on Test Specification 2 (moderate correlation)

The results for the MCAT with the same test specification, but with θ_1, θ_2, and θ_3 are moderately correlated, are presented in Table 5.15 and 5.15. The results in Table 5.15 are under

Table 5.15: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control
(3-dimension simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.01, 0.00, 0.00)	(0.01, 0.01, 0.00)	(0.00, 0.00, 0.00)
RMSE	(0.52, 0.49, 0.49)	(0.51, 0.49, 0.49)	(0.52, 0.49, 0.49)
Correlation	(0.86, 0.87, 0.87)	(0.86, 0.87, 0.87)	(0.86, 0.87, 0.87)
Average test information	$\begin{bmatrix} 2.38 & 0 & 0 \\ 0 & 2.39 & 0 \\ 0 & 0 & 2.39 \end{bmatrix}$	$\begin{bmatrix} 2.38 & 0 & 0 \\ 0 & 2.38 & 0 \\ 0 & 0 & 2.38 \end{bmatrix}$	$\begin{bmatrix} 2.38 & 0 & 0 \\ 0 & 2.38 & 0 \\ 0 & 0 & 2.39 \end{bmatrix}$
Overall Pool Usage	29.38	31.95	67.46
Overlap rate	0.18	0.33	0.20
% of overexposed item (r > 0.2)	15%	35%	8%
% of underexposed item (r<0.02)	31%	29%	53%

Table 5.16: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control
(3-dimension simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.01, -0.01, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)
RMSE	(0.54, 0.51, 0.50)	(0.54, 0.51, 0.50)	(0.55, 0.52, 0.50)
Correlation	(0.85, 0.86, 0.86)	(0.84, 0.86, 0.86)	(0.84, 0.85, 0.86)
Average test information	$\begin{bmatrix} 2.14 & 0 & 0 \\ 0 & 2.16 & 0 \\ 0 & 0 & 2.17 \end{bmatrix}$	$\begin{bmatrix} 2.12 & 0 & 0 \\ 0 & 2.15 & 0 \\ 0 & 0 & 2.15 \end{bmatrix}$	$\begin{bmatrix} 2.07 & 0 & 0 \\ 0 & 2.00 & 0 \\ 0 & 0 & 2.15 \end{bmatrix}$
Overall Pool Usage	1.50	0.47	9.82
Overlap rate	0.10	13%	0.08
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r<0.02)	0%	0%	21%

the condition without item exposure control; and Table 5.16 is with item exposure control. In both tables, there are three values for bias, RMSE and correlation, representing the results for $(\theta_1, \theta_2, \theta_3)$.

Under the condition without item exposure control (see Table 5.15), the *p*-optimal item pools and the baseline pool show nearly no bias on the $\boldsymbol{\theta}$ estimates. Also, the RMSE are all at 0.50, and correlations between estimated $\boldsymbol{\theta}$ and true $\boldsymbol{\theta}$ are around 0.87. The average test information is also very similar among the three item pools. The amount of information on the direction of each θ is around 2.38. In general, the results suggest that the *.96*- and *.86*-optimal item pool can provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as baseline pool.

Table 5.15 also presents the results about item pool usage. The overall pool usage index for the *.96*-optimal item pool is slightly smaller than that of the *.86*-optimal item pool, and the index for the baseline pool is more than twice as much as the *.96*- and *.86*-optimal item pool. The results suggest that the *.96*-optimal item pool has been slightly better used than the *.86*-optimal item pool, and the two optimal item pools have been much better used than the baseline pool. More specifically, for the *.96*-optimal item pool, the overlap rate is 0.18, and the percentage of overexposed and underexposed item are 15% and 31%, respectively. For the *.86*-optimal item pool, the results are: 33% of items overlap, 35% overexposed, and 29 % under exposed. Because more items from the *.86*-optimal item pool are overlapped and overexposed, the *.86*-optimal item pool is less secure than the *.96*-optimal item pool. The overlap rate for the baseline pool is 0.20, which is slightly higher than the *.96*-optimal item pool and lower than the *.86*-optimal item pool. Although a smaller number of items (8%) from the baseline pool are overexposed, more than half of the items (53%) are rarely used. It implies many items in the baseline pool are wasted. In brief, based on these pool usage results, the item pool usage for the *.96*- and *.86*-optimal item pool is much better than the baseline pool.

When item exposure control is implemented (see Table 5.16), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimation as the baseline pool, and

yield better item pool usage than the baseline pool. Compared with the condition without item exposure control, item exposure control only results in a 0.01 to 0.03 increase for the RMSE, about 0.01 to 0.02 decrease in correlation, and about 0.3 decrease on the average test information. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed item and overlapped item are also decreased. The *.96*- and the *.86*-optimal item pool has been fully used with no item underexposed. The overall pool usage index for the *.96*-, *.86*-optimal item pool and the baseline pool are 1.50, 0.47, and 9.82, respectively. The value is much smaller than the condition without item exposure control. Thus, the item exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

In addition to the overall pool performance, the conditional bias and RMSE at 37 ($\theta_1, \theta_2, \theta_3$) points are also reported. The conditional bias for each $\boldsymbol{\theta}$ point is presented in Table 5.17 and 5.18, for the MCAT without and with exposure control, respectively. Negative bias is colored in blue and positive bias is in red. Deeper color represents larger bias. The conditional RMSE is presented in Table 5.19 and 5.20 in the same manner. Small RMSE is colored in green and large RMSE is colored in red.

Under the condition without item exposure control (see Table 5.17 for bias and 5.19 for RMSE), the conditional bias and RMSE for the .96-, .86-optimal item pool, and the baseline pool are quite similar. This finding supports the results for the overall bias and RMSE, and also suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. Similar to the condition that dimensions are highly correlated, larger bias and RMSE occurs when θ_1, θ_2, and θ_3 are very large or very small. The difference between two $\theta's$ also affects the estimation accuracy. More specifically, when θ_1 is around 0, and θ_2 and θ_3 are

Table 5.17: Conditional Bias for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.61	0.60	0.68	0.48	0.59	0.67	0.48	0.55	0.57
-3	-3	-2	0.70	0.78	0.74	0.70	0.75	0.74	0.01	0.03	0.05
-3	-2	-3	0.78	0.71	0.70	0.11	0.10	0.09	0.77	0.70	0.73
-3	-2	-2	0.89	0.83	0.84	0.29	0.22	0.26	0.20	0.17	0.20
-2	-3	-3	0.15	0.13	0.15	0.65	0.66	0.69	0.64	0.70	0.69
-2	-3	-2	0.33	0.26	0.28	0.86	0.80	0.84	0.20	0.19	0.18
-2	-2	-3	0.20	0.23	0.18	0.18	0.28	0.11	0.83	0.83	0.77
-2	-2	-2	0.42	0.32	0.33	0.42	0.30	0.29	0.40	0.29	0.27
-2	-2	-1	0.48	0.54	0.44	0.44	0.56	0.50	-0.16	-0.13	-0.16
-2	-1	-2	0.48	0.50	0.49	-0.19	-0.10	-0.14	0.45	0.46	0.52
-2	-1	-1	0.60	0.60	0.70	-0.04	0.08	0.09	-0.02	0.00	0.03
-1	-2	-2	-0.06	-0.06	0.03	0.41	0.37	0.51	0.44	0.35	0.52
-1	-2	-1	0.03	0.09	0.00	0.57	0.62	0.54	-0.02	-0.04	-0.10
-1	0	-1	0.31	0.27	0.26	-0.27	-0.24	-0.39	0.36	0.30	0.31
-1	0	0	0.44	0.39	0.45	-0.09	-0.14	-0.08	-0.06	-0.11	-0.03
0	-1	-1	-0.22	-0.28	-0.24	0.31	0.24	0.21	0.33	0.23	0.24
0	-1	0	-0.06	-0.04	-0.09	0.45	0.49	0.45	-0.24	-0.13	-0.13
0	0	-1	-0.16	-0.08	-0.14	-0.19	-0.14	-0.11	0.46	0.48	0.47
0	0	0	0.06	-0.06	0.01	0.09	-0.04	-0.04	0.09	-0.02	-0.01
0	0	1	0.15	0.10	0.09	0.18	0.15	0.16	-0.47	-0.48	-0.45
0	1	0	0.10	0.12	0.16	-0.46	-0.43	-0.37	0.14	0.13	0.18
0	1	1	0.13	0.25	0.21	-0.29	-0.29	-0.28	-0.36	-0.29	-0.27
1	0	0	-0.39	-0.40	-0.40	0.09	0.09	0.10	0.09	0.07	0.15
1	0	1	-0.27	-0.22	-0.32	0.25	0.29	0.26	-0.38	-0.28	-0.37
1	2	1	-0.11	-0.04	-0.04	-0.68	-0.61	-0.68	-0.08	-0.01	-0.03
1	2	2	0.04	0.00	0.06	-0.47	-0.45	-0.47	-0.48	-0.47	-0.49
2	1	1	-0.62	-0.67	-0.60	-0.06	-0.04	-0.05	-0.01	-0.07	-0.01
2	1	2	-0.45	-0.44	-0.41	0.11	0.09	0.13	-0.45	-0.49	-0.49
2	2	1	-0.47	-0.52	-0.54	-0.48	-0.52	-0.49	0.17	0.10	0.16
2	2	2	-0.30	-0.31	-0.38	-0.38	-0.35	-0.38	-0.32	-0.35	-0.31
2	2	3	-0.21	-0.30	-0.32	-0.13	-0.24	-0.20	-0.80	-0.86	-0.83
2	3	2	-0.33	-0.30	-0.27	-0.86	-0.81	-0.87	-0.20	-0.16	-0.16
2	3	3	-0.11	-0.14	-0.17	-0.62	-0.63	-0.62	-0.64	-0.71	-0.61
3	2	2	-0.91	-0.87	-0.91	-0.24	-0.28	-0.22	-0.22	-0.24	-0.23
3	2	3	-0.76	-0.68	-0.75	-0.16	-0.05	-0.11	-0.77	-0.72	-0.74
3	3	2	-0.76	-0.73	-0.71	-0.75	-0.69	-0.76	-0.02	-0.03	-0.11
3	3	3	-0.65	-0.57	-0.56	-0.56	-0.53	-0.61	-0.56	-0.52	-0.56

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.18: Conditional Bias for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.81	0.90	0.73	0.69	0.82	0.69	0.69	0.84	0.66
-3	-3	-2	0.98	1.05	0.97	0.94	1.02	0.94	0.17	0.17	0.20
-3	-2	-3	0.92	0.99	0.89	0.22	0.20	0.17	0.94	0.99	0.81
-3	-2	-2	1.06	1.19	1.00	0.36	0.44	0.30	0.29	0.39	0.24
-2	-3	-3	0.28	0.24	0.18	0.88	0.95	0.86	0.86	0.90	0.82
-2	-3	-2	0.35	0.40	0.37	1.01	1.09	1.02	0.22	0.33	0.21
-2	-2	-3	0.43	0.38	0.31	0.32	0.30	0.32	1.04	1.09	0.93
-2	-2	-2	0.43	0.48	0.44	0.37	0.43	0.45	0.35	0.43	0.39
-2	-2	-1	0.58	0.54	0.58	0.57	0.56	0.55	-0.10	-0.19	-0.14
-2	-1	-2	0.61	0.54	0.57	-0.05	-0.15	-0.13	0.55	0.53	0.51
-2	-1	-1	0.69	0.78	0.67	0.05	0.17	0.09	0.08	0.11	0.09
-1	-2	-2	-0.08	0.00	-0.02	0.51	0.57	0.51	0.53	0.60	0.52
-1	-2	-1	0.14	0.15	0.11	0.74	0.79	0.75	0.05	0.08	0.01
-1	0	-1	0.41	0.32	0.38	-0.20	-0.30	-0.33	0.40	0.36	0.34
-1	0	0	0.42	0.45	0.40	-0.12	-0.13	-0.16	-0.11	-0.12	-0.15
0	-1	-1	-0.25	-0.27	-0.20	0.29	0.29	0.35	0.31	0.30	0.34
0	-1	0	-0.10	-0.15	-0.02	0.50	0.49	0.55	-0.14	-0.18	-0.13
0	0	-1	-0.17	-0.10	-0.16	-0.15	-0.13	-0.22	0.49	0.51	0.49
0	0	0	-0.02	-0.07	0.05	0.07	0.00	0.05	0.00	0.02	0.08
0	0	1	0.22	0.14	0.15	0.29	0.16	0.22	-0.44	-0.47	-0.42
0	1	0	0.13	0.21	0.10	-0.46	-0.47	-0.51	0.22	0.16	0.15
0	1	1	0.28	0.25	0.21	-0.31	-0.29	-0.36	-0.34	-0.30	-0.34
1	0	0	-0.40	-0.43	-0.49	0.14	0.16	0.13	0.17	0.18	0.11
1	0	1	-0.28	-0.37	-0.32	0.34	0.25	0.28	-0.33	-0.40	-0.39
1	2	1	-0.09	-0.17	-0.10	-0.78	-0.75	-0.74	-0.03	-0.06	-0.09
1	2	2	-0.03	0.00	0.03	-0.58	-0.60	-0.55	-0.54	-0.62	-0.54
2	1	1	-0.65	-0.73	-0.72	-0.07	-0.13	-0.08	-0.03	-0.10	-0.03
2	1	2	-0.60	-0.64	-0.57	0.06	0.07	0.05	-0.63	-0.66	-0.64
2	2	1	-0.63	-0.56	-0.65	-0.67	-0.58	-0.60	0.10	0.14	0.07
2	2	2	-0.53	-0.51	-0.44	-0.39	-0.45	-0.46	-0.38	-0.48	-0.47
2	2	3	-0.32	-0.40	-0.40	-0.28	-0.37	-0.31	-1.04	-1.11	-1.07
2	3	2	-0.32	-0.43	-0.35	-1.01	-1.14	-1.06	-0.26	-0.39	-0.31
2	3	3	-0.19	-0.32	-0.34	-0.84	-1.04	-0.89	-0.84	-1.00	-0.89
3	2	2	-1.05	-1.18	-1.07	-0.34	-0.40	-0.32	-0.28	-0.34	-0.33
3	2	3	-0.91	-0.99	-0.92	-0.13	-0.24	-0.18	-0.88	-1.00	-0.92
3	3	2	-0.92	-1.05	-0.98	-0.95	-1.04	-0.95	-0.19	-0.25	-0.17
3	3	3	-0.79	-0.99	-0.74	-0.69	-0.90	-0.70	-0.71	-0.87	-0.68

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.19: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.74	0.74	0.78	0.62	0.73	0.77	0.60	0.70	0.72
-3	-3	-2	0.81	0.87	0.85	0.80	0.86	0.84	0.45	0.39	0.41
-3	-2	-3	0.88	0.80	0.82	0.45	0.41	0.39	0.85	0.79	0.82
-3	-2	-2	0.97	0.91	0.93	0.47	0.49	0.46	0.44	0.44	0.44
-2	-3	-3	0.41	0.41	0.43	0.77	0.76	0.80	0.74	0.79	0.82
-2	-3	-2	0.53	0.49	0.50	0.96	0.88	0.91	0.43	0.45	0.45
-2	-2	-3	0.44	0.43	0.43	0.50	0.49	0.41	0.93	0.91	0.85
-2	-2	-2	0.63	0.51	0.51	0.58	0.49	0.47	0.55	0.51	0.49
-2	-2	-1	0.63	0.66	0.60	0.57	0.66	0.63	0.43	0.41	0.41
-2	-1	-2	0.62	0.62	0.61	0.46	0.40	0.41	0.61	0.59	0.68
-2	-1	-1	0.72	0.72	0.81	0.40	0.38	0.38	0.41	0.41	0.40
-1	-2	-2	0.42	0.41	0.37	0.55	0.53	0.65	0.58	0.48	0.65
-1	-2	-1	0.44	0.40	0.42	0.67	0.72	0.69	0.40	0.38	0.40
-1	0	-1	0.48	0.44	0.47	0.47	0.45	0.57	0.52	0.49	0.51
-1	0	0	0.61	0.54	0.65	0.40	0.46	0.41	0.35	0.40	0.40
0	-1	-1	0.44	0.48	0.49	0.45	0.43	0.41	0.50	0.42	0.45
0	-1	0	0.44	0.36	0.43	0.60	0.61	0.59	0.50	0.36	0.35
0	0	-1	0.43	0.40	0.42	0.47	0.36	0.43	0.59	0.65	0.61
0	0	0	0.41	0.41	0.42	0.37	0.37	0.39	0.41	0.36	0.41
0	0	1	0.45	0.39	0.44	0.43	0.38	0.43	0.63	0.60	0.60
0	1	0	0.36	0.38	0.40	0.62	0.59	0.51	0.37	0.46	0.41
0	1	1	0.41	0.47	0.45	0.50	0.51	0.45	0.52	0.49	0.46
1	0	0	0.57	0.55	0.57	0.46	0.36	0.43	0.43	0.38	0.48
1	0	1	0.48	0.46	0.47	0.48	0.45	0.49	0.58	0.47	0.56
1	2	1	0.39	0.41	0.41	0.79	0.72	0.77	0.39	0.41	0.41
1	2	2	0.37	0.46	0.42	0.60	0.59	0.61	0.60	0.61	0.63
2	1	1	0.73	0.79	0.71	0.39	0.41	0.40	0.38	0.41	0.35
2	1	2	0.58	0.64	0.60	0.44	0.42	0.42	0.58	0.66	0.63
2	2	1	0.63	0.68	0.63	0.62	0.63	0.63	0.48	0.37	0.37
2	2	2	0.52	0.49	0.53	0.55	0.50	0.55	0.48	0.54	0.50
2	2	3	0.39	0.50	0.50	0.37	0.45	0.42	0.88	0.93	0.91
2	3	2	0.47	0.49	0.50	0.92	0.87	0.95	0.47	0.39	0.46
2	3	3	0.47	0.38	0.46	0.73	0.76	0.76	0.74	0.81	0.74
3	2	2	1.02	0.96	1.00	0.47	0.49	0.48	0.47	0.47	0.49
3	2	3	0.87	0.79	0.85	0.42	0.38	0.42	0.89	0.82	0.82
3	3	2	0.84	0.82	0.81	0.84	0.78	0.84	0.38	0.40	0.43
3	3	3	0.78	0.68	0.71	0.66	0.63	0.70	0.69	0.63	0.66

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.20: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.90	0.97	0.84	0.80	0.92	0.80	0.78	0.92	0.77
-3	-3	-2	1.07	1.11	1.07	1.02	1.11	1.02	0.47	0.47	0.44
-3	-2	-3	1.01	1.06	0.99	0.51	0.42	0.40	1.05	1.05	0.90
-3	-2	-2	1.14	1.26	1.09	0.51	0.58	0.49	0.50	0.54	0.45
-2	-3	-3	0.50	0.49	0.54	0.96	1.03	1.00	0.95	0.98	0.95
-2	-3	-2	0.55	0.60	0.56	1.09	1.17	1.12	0.43	0.55	0.48
-2	-2	-3	0.61	0.57	0.49	0.54	0.48	0.52	1.11	1.16	1.01
-2	-2	-2	0.58	0.63	0.61	0.53	0.60	0.64	0.52	0.59	0.57
-2	-2	-1	0.71	0.67	0.72	0.73	0.69	0.68	0.45	0.48	0.39
-2	-1	-2	0.76	0.67	0.71	0.43	0.46	0.41	0.71	0.68	0.64
-2	-1	-1	0.82	0.86	0.81	0.40	0.45	0.43	0.44	0.41	0.45
-1	-2	-2	0.46	0.42	0.44	0.64	0.69	0.66	0.70	0.74	0.66
-1	-2	-1	0.41	0.45	0.41	0.86	0.89	0.85	0.42	0.38	0.39
-1	0	-1	0.57	0.53	0.56	0.48	0.54	0.54	0.58	0.55	0.56
-1	0	0	0.58	0.61	0.55	0.43	0.43	0.43	0.43	0.39	0.39
0	-1	-1	0.49	0.46	0.46	0.48	0.49	0.57	0.52	0.48	0.57
0	-1	0	0.40	0.46	0.43	0.65	0.63	0.69	0.44	0.45	0.44
0	0	-1	0.42	0.42	0.46	0.46	0.45	0.45	0.67	0.68	0.65
0	0	0	0.41	0.42	0.44	0.44	0.37	0.38	0.44	0.40	0.39
0	0	1	0.45	0.44	0.45	0.52	0.42	0.52	0.61	0.65	0.60
0	1	0	0.45	0.47	0.43	0.62	0.62	0.66	0.47	0.42	0.43
0	1	1	0.49	0.46	0.48	0.49	0.47	0.56	0.52	0.46	0.55
1	0	0	0.64	0.60	0.70	0.44	0.44	0.45	0.45	0.50	0.45
1	0	1	0.52	0.55	0.54	0.52	0.48	0.53	0.53	0.56	0.61
1	2	1	0.44	0.46	0.43	0.89	0.88	0.87	0.40	0.40	0.44
1	2	2	0.47	0.47	0.43	0.73	0.75	0.68	0.69	0.74	0.69
2	1	1	0.76	0.88	0.84	0.39	0.47	0.42	0.38	0.49	0.36
2	1	2	0.73	0.82	0.71	0.40	0.47	0.42	0.75	0.80	0.75
2	2	1	0.76	0.72	0.76	0.77	0.73	0.71	0.40	0.43	0.45
2	2	2	0.64	0.66	0.66	0.56	0.65	0.64	0.52	0.62	0.63
2	2	3	0.55	0.64	0.58	0.50	0.62	0.50	1.12	1.19	1.14
2	3	2	0.53	0.64	0.54	1.09	1.23	1.16	0.50	0.58	0.56
2	3	3	0.48	0.60	0.54	0.93	1.17	0.98	0.93	1.14	0.99
3	2	2	1.13	1.27	1.15	0.52	0.58	0.56	0.45	0.56	0.56
3	2	3	1.00	1.10	1.03	0.42	0.52	0.44	0.96	1.09	1.00
3	3	2	1.01	1.16	1.08	1.04	1.14	1.07	0.47	0.55	0.47
3	3	3	0.89	1.11	0.88	0.79	1.04	0.80	0.84	1.01	0.82

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

near θ_1, the bias for θ_1 is close to 0 and the RMSE is around 0.4. Negative bias and large RMSE appear when the value of θ_1 increases and the difference between θ_1 and θ_2, and between θ_1 and θ_3, increases. For example, at point (3, 2, 2) in the table, the bias for $\theta_1$is around -0.74 and RMSE for $\theta_1$is around 0.79. Meanwhile, positive bias and large RMSE appear when the value of θ_1 decreases and the difference between θ_1 and θ_2, and between θ_1 and θ_3, increases. At point (-3, -2, -2), the bias for $\theta_1$is about 0.74 and RMSE for $\theta_1$is around 0.79. Similar results for θ_2 can be observed from the three columns in the middle of Table 5.17 and 5.19. When θ_2 is around 0, and θ_1 and θ_3 is near θ_2, the bias and RMSE for θ_2 is very small. When the value of θ_2 becomes more extreme and θ_1 and $\theta_3$is away from θ_2, large bias and RMSE values appear. Again, similar results can be found for θ_3 from the three columns on the right side of Table 5.17 and 5.19. As described in Section 5.2.1, this finding is probably due to the Bayesian MAP estimation method. By comparing the conditional Bias and RMSE in this section with the results in Section 5.2.3, when θ_1, θ_2, and θ_3 are highly correlated, it is easy to observe that the pattern of these tables are the same, but the magnitude of the bias and RMSE in this section is smaller. When the correlation among θ_1, θ_2, and θ_3 decrease, the prior will only weakly reduce the difference among θ_1, θ_2, and θ_3. Therefore, the bias the RMSE values are slightly smaller at those points where θ_1 and θ_2 are away from each other, compared with the condition when θ_1, θ_2, and θ_3 are highly correlated.

When item exposure control is implemented, similar findings can be observed from Table 5.18 and 5.20. Again, there is nearly no difference between the two *p*-optimal item pools, and between the *p*-optimal item pools and the baseline pool. The results suggest the three item pools perform similarly in terms of the ability estimation on the 37 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1, θ_2, and θ_3 are very large or very small, and when $\theta's$ are away

from each other. Similar to results under the high correlation condition in Section 5.2.3, when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger.

In summary, this section present the results for the MCAT with the test specification of three-dimension simple structure and with moderate correlation among θ_1, θ_2, and θ_3. The *p*-optimal item pools perform similarly as the baseline pool in terms of the accuracy of ability estimation, but the *p*-optimal item pools can save over 140 items and have a better item pool usage. When item exposure control is implemented, the *p*-optimal item pools still can provide accurate ability estimation and meanwhile the item exposure rate and item overlap rate can be well controlled.

In general, the findings from this section are similar to the previous, when θ_1, θ_2, and θ_3 are highly correlated. A closely comparison between these two sections reveal that the measurement error in this section is slightly larger. This result is due to the correlation among θ_1, θ_2, and θ_3. As explained in Section 5.2.2, the MIRT model estimates all the θ's simultaneously by borrowing information from one to another. When θ_1, θ_2, and θ_3 are highly correlated, more information can be borrowed for ability estimation. When the correlation decreases, the amount of information that can be borrowed can be reduced, and therefore the RMSE increase. In addition to the accuracy of ability estimation, the pool usage for the two *p*-optimal item pool in this section is also slightly better. This is probably because of the pool size. When the correlation decreases, the pool size decreases as well. A smaller item pool is more likely to be fully used.

5.2.5 Performance for item pools based on Test Specification 3 (high correlation)

The results of the ability estimates and item pool utilization for the *.96*-optimal item pool, the *.86*-optimal item pool, and the baseline pool based on the three-dimension non-simple

structure test specification with θ_1, θ_2, and θ_3 are highly correlated are presented in Table 5.21 and 5.22. The results in Table 5.21 is under the condition that no item exposure control is implemented; and Table 5.22 is when item exposure control is implemented. In both tables, there are three values for bias, RMSE and correlation, representing the results for (θ_1, θ_2, θ_3).

Under the condition without item exposure control (see Table 5.21), the *p*-optimal item pools and the baseline pool show no bias on the $\boldsymbol{\theta}$ estimates. Also, the RMSE are between 0.31 and 0.37, and correlations between estimated $\boldsymbol{\theta}$ and true $\boldsymbol{\theta}$ are around 0.94. The average test information between the *.96*-optimal item pool and the baseline pool is very similar, but the information for the *.86*-optimal item pool is slightly smaller. The amount of information on the direction of θ_1, θ_2, and θ_3 (i.e., the value on the diagonal) is about 3.50 for the *.96*-optimal item pool and the baseline pool, and about 3.39 for the *.86*-optimal item pool. These values are over one unit higher than the values under the three-dimensional non-simple structure case. The additional information comes from items in Cluster 4 with $\boldsymbol{a} = (1, 1, 1)$. Because these items measure all the θ's, they provide information on the direction of θ_1, θ_2, and θ_3, as well as on the direction of the diagonal in the three dimensional space (see Figure 5.1). For this reason, the off-diagonal values in the information matrix are no longer zero. The values on the off-diagonal represent the amount of information on the direction of the θ_1- θ_2 composite, θ_1- θ_3 composite, and θ_2- θ_3 composite. In general, the results suggest that the *.96*- and *.86*-optimal item pool can provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as baseline pool, but the average test information for the *.86*-optimal item pool is slightly small than the other two.

Table 5.21 also presents the results about item pool usage. Compared with the MCAT based on Test Specification 1 and 2 in Section 5.2.1 to 5.2.4, similar results can be drawn from Table 5.21. The item pool usage for the *.96*-optimal item pool is slightly better than the *86*-optimal

Table 5.21: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control (3-dimension non-simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.00, 0.00, 0.00)	(0.00, 0.00, -0.01)	(-0.01, 0.00, 0.00)
RMSE	(0.35, 0.31, 0.37)	(0.35, 0.31, 0.37)	(0.35, 0.31, 0.37)
Correlation	(0.94, 0.95, 0.93)	(0.94, 0.95, 0.93)	(0.94, 0.95, 0.93)
Average test information	$\begin{bmatrix} 3.50 & 1.75 & 1.75 \\ 1.75 & 3.48 & 1.75 \\ 1.75 & 1.75 & 3.50 \end{bmatrix}$	$\begin{bmatrix} 3.40 & 1.61 & 1.61 \\ 1.61 & 3.37 & 1.61 \\ 1.61 & 1.61 & 3.39 \end{bmatrix}$	$\begin{bmatrix} 3.52 & 1.77 & 1.77 \\ 1.77 & 3.49 & 1.77 \\ 1.77 & 1.77 & 3.52 \end{bmatrix}$
Overall Pool Usage	28.06	28.69	35.26
Overlap rate	0.14	0.25	0.12
% of overexposed item (r > 0.2)	3%	30%	1%
% of underexposed item (r<0.02)	33%	31%	43%

Table 5.22: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control (3-dimension non-simple structure, high correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)
RMSE	(0.37, 0.33, 0.39)	(0.37, 0.33, 0.39)	(0.36, 0.32, 0.38)
Correlation	(0.93, 0.94, 0.92)	(0.93, 0.94, 0.92)	(0.93, 0.95, 0.93)
Average test information	$\begin{bmatrix} 2.94 & 1.30 & 1.30 \\ 1.30 & 2.94 & 1.30 \\ 1.30 & 1.30 & 2.94 \end{bmatrix}$	$\begin{bmatrix} 2.89 & 1.30 & 1.30 \\ 1.30 & 2.89 & 1.30 \\ 1.30 & 1.30 & 2.89 \end{bmatrix}$	$\begin{bmatrix} 3.18 & 1.61 & 1.61 \\ 1.61 & 3.13 & 1.61 \\ 1.61 & 1.61 & 3.24 \end{bmatrix}$
Overall Pool Usage	3.64	1.65	8.26
Overlap rate	0.08	0.12	0.07
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r<0.02)	1%	0%	17%

item pool. And the two *p*-optimal item pools are much better used than the baseline pool.

When item exposure control is implemented (see Table 5.22), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimates as the baseline pool, and yield better item pool usage than the baseline pool. Compared with the condition without item

exposure control, item exposure control only results in 0.01 to 0.02 increase for the RMSE, 0.01 decrease in correlation, and about 0.5 decrease for the average test information. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed items and overlapped items are also decreased. The two *p*-optimal item pools have been fully used. No item from the .86-optimal item pool is underexposed, and only 1% of items from *96*-optimal item pool are underexposed. The comparison between the condition with and without item exposure control suggests the item exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

In addition to the overall pool performance, the conditional bias and RMSE at the 37 ($\theta_1, \theta_2, \theta_3$) points are also calculated. The conditional bias for each $\boldsymbol{\theta}$ point is presented in Table 5.23 and 5.24, for the MCAT without and with item exposure control, respectively. Negative bias is colored in blue and positive bias is in red. Deeper color represents larger bias. The conditional RMSE is presented in Table 5.25 and 5.26 in the same manner. Small RMSE is colored in green and large RMSE is colored in red.

Under the condition without item exposure control (see Table 5.23 for bias and 5.25 for RMSE), the conditional bias and RMSE for the .96-, .86-optimal item pool, and the baseline pool are quite similar. This finding supports the results of the overall bias and RMSE, and suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. Similar to the results from the Test Specification 2 (three-dimension simple structure), larger bias and RMSE occurs when θ_1, θ_2, and θ_3 are very large or very small, which is the top and the bottom of each table. The difference between θ_1 and θ_2, and between θ_1 and θ_3, also affects the estimation accuracy.

Table 5.23: Conditional Bias for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension non-simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.40	0.41	0.24	0.33	0.33	0.19	0.42	0.43	0.28
-3	-3	-2	0.59	0.52	0.44	0.59	0.53	0.44	-0.15	-0.21	-0.30
-3	-2	-3	0.55	0.60	0.50	-0.36	-0.35	-0.46	0.61	0.61	0.50
-3	-2	-2	0.73	0.75	0.74	-0.19	-0.15	-0.18	-0.02	0.00	-0.05
-2	-3	-3	-0.29	-0.27	-0.38	0.51	0.51	0.42	0.49	0.53	0.44
-2	-3	-2	-0.13	-0.13	-0.14	0.71	0.73	0.70	-0.11	-0.13	-0.13
-2	-2	-3	-0.02	-0.02	-0.10	-0.14	-0.12	-0.19	0.73	0.75	0.72
-2	-2	-2	0.08	0.13	0.11	0.05	0.10	0.07	0.11	0.19	0.14
-2	-2	-1	0.30	0.37	0.33	0.31	0.36	0.35	-0.46	-0.44	-0.39
-2	-1	-2	0.39	0.38	0.43	-0.54	-0.55	-0.51	0.40	0.37	0.41
-2	-1	-1	0.62	0.62	0.64	-0.23	-0.27	-0.25	-0.11	-0.18	-0.18
-1	-2	-2	-0.47	-0.43	-0.42	0.36	0.40	0.41	0.29	0.38	0.38
-1	-2	-1	-0.22	-0.22	-0.18	0.64	0.64	0.68	-0.22	-0.18	-0.19
-1	0	-1	0.38	0.31	0.36	-0.54	-0.61	-0.55	0.32	0.29	0.39
-1	0	0	0.57	0.56	0.57	-0.26	-0.29	-0.29	-0.16	-0.18	-0.20
0	-1	-1	-0.48	-0.55	-0.49	0.36	0.29	0.33	0.26	0.24	0.25
0	-1	0	-0.32	-0.27	-0.30	0.58	0.62	0.59	-0.26	-0.26	-0.29
0	0	-1	-0.22	-0.23	-0.21	-0.29	-0.29	-0.28	0.50	0.51	0.52
0	0	0	0.02	-0.01	0.04	0.04	-0.02	0.00	0.05	-0.03	0.00
0	0	1	0.19	0.21	0.24	0.26	0.27	0.29	-0.53	-0.54	-0.52
0	1	0	0.26	0.26	0.31	-0.63	-0.61	-0.59	0.23	0.25	0.25
0	1	1	0.46	0.52	0.45	-0.34	-0.33	-0.36	-0.27	-0.29	-0.29
1	0	0	-0.50	-0.52	-0.54	0.35	0.32	0.30	0.25	0.22	0.18
1	0	1	-0.32	-0.33	-0.37	0.59	0.55	0.55	-0.30	-0.36	-0.29
1	2	1	0.22	0.19	0.22	-0.65	-0.68	-0.64	0.19	0.17	0.20
1	2	2	0.41	0.42	0.42	-0.41	-0.41	-0.38	-0.36	-0.39	-0.32
2	1	1	-0.63	-0.64	-0.66	0.23	0.25	0.22	0.11	0.15	0.12
2	1	2	-0.42	-0.42	-0.43	0.52	0.50	0.51	-0.42	-0.44	-0.42
2	2	1	-0.33	-0.33	-0.32	-0.35	-0.36	-0.34	0.41	0.42	0.44
2	2	2	-0.12	-0.11	-0.14	-0.06	-0.08	-0.09	-0.12	-0.14	-0.17
2	2	3	0.02	0.09	0.08	0.13	0.16	0.19	-0.77	-0.72	-0.71
2	3	2	0.17	0.13	0.10	-0.68	-0.74	-0.75	0.15	0.04	0.05
2	3	3	0.19	0.27	0.35	-0.60	-0.49	-0.47	-0.57	-0.48	-0.48
3	2	2	-0.70	-0.74	-0.75	0.21	0.14	0.18	0.07	0.01	0.03
3	2	3	-0.63	-0.56	-0.51	0.35	0.37	0.46	-0.59	-0.60	-0.51
3	3	2	-0.56	-0.50	-0.41	-0.57	-0.49	-0.42	0.18	0.21	0.34
3	3	3	-0.42	-0.32	-0.20	-0.34	-0.26	-0.14	-0.44	-0.37	-0.22

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.24: Conditional Bias for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension non-simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.44	0.65	0.50	0.38	0.61	0.45	0.45	0.72	0.53
-3	-3	-2	0.67	0.81	0.65	0.66	0.80	0.66	-0.12	-0.01	-0.13
-3	-2	-3	0.75	0.88	0.71	-0.23	-0.12	-0.26	0.77	0.86	0.74
-3	-2	-2	0.90	0.96	0.82	-0.03	0.01	-0.10	0.07	0.13	0.05
-2	-3	-3	-0.20	-0.09	-0.20	0.65	0.78	0.64	0.70	0.79	0.64
-2	-3	-2	0.02	-0.01	-0.05	0.89	0.91	0.84	0.03	0.03	0.00
-2	-2	-3	0.09	0.14	0.08	-0.01	0.06	-0.03	0.88	1.01	0.90
-2	-2	-2	0.22	0.32	0.23	0.18	0.26	0.19	0.25	0.30	0.23
-2	-2	-1	0.46	0.51	0.39	0.47	0.52	0.42	-0.31	-0.30	-0.34
-2	-1	-2	0.52	0.53	0.46	-0.43	-0.43	-0.50	0.52	0.54	0.47
-2	-1	-1	0.72	0.71	0.74	-0.17	-0.23	-0.17	-0.06	-0.14	-0.08
-1	-2	-2	-0.37	-0.33	-0.42	0.47	0.53	0.41	0.44	0.53	0.39
-1	-2	-1	-0.20	-0.18	-0.11	0.69	0.72	0.76	-0.20	-0.14	-0.14
-1	0	-1	0.38	0.41	0.39	-0.56	-0.54	-0.56	0.38	0.36	0.33
-1	0	0	0.61	0.57	0.54	-0.27	-0.30	-0.34	-0.23	-0.22	-0.24
0	-1	-1	-0.50	-0.48	-0.49	0.35	0.38	0.35	0.29	0.31	0.27
0	-1	0	-0.32	-0.24	-0.29	0.61	0.64	0.61	-0.25	-0.28	-0.31
0	0	-1	-0.21	-0.22	-0.28	-0.26	-0.29	-0.30	0.57	0.54	0.57
0	0	0	-0.03	0.00	-0.01	-0.02	0.00	-0.01	-0.01	0.01	0.00
0	0	1	0.19	0.18	0.20	0.27	0.21	0.25	-0.53	-0.62	-0.61
0	1	0	0.24	0.31	0.30	-0.64	-0.59	-0.59	0.26	0.30	0.30
0	1	1	0.51	0.47	0.51	-0.34	-0.39	-0.34	-0.28	-0.33	-0.29
1	0	0	-0.60	-0.63	-0.57	0.28	0.26	0.30	0.19	0.18	0.20
1	0	1	-0.40	-0.39	-0.31	0.51	0.55	0.60	-0.39	-0.38	-0.32
1	2	1	0.11	0.14	0.16	-0.77	-0.75	-0.75	0.08	0.10	0.10
1	2	2	0.38	0.33	0.37	-0.47	-0.52	-0.48	-0.46	-0.50	-0.46
2	1	1	-0.75	-0.78	-0.75	0.17	0.12	0.17	0.10	0.03	0.12
2	1	2	-0.49	-0.54	-0.46	0.46	0.42	0.48	-0.49	-0.55	-0.44
2	2	1	-0.43	-0.51	-0.46	-0.46	-0.52	-0.48	0.35	0.29	0.32
2	2	2	-0.25	-0.32	-0.23	-0.20	-0.30	-0.18	-0.29	-0.35	-0.22
2	2	3	-0.05	-0.16	-0.11	0.04	-0.10	0.01	-0.86	-1.04	-0.92
2	3	2	0.01	-0.11	-0.03	-0.88	-1.00	-0.91	-0.03	-0.12	-0.05
2	3	3	0.14	0.11	0.11	-0.67	-0.77	-0.75	-0.67	-0.81	-0.76
3	2	2	-0.86	-0.99	-0.88	0.07	-0.03	0.07	-0.07	-0.17	-0.06
3	2	3	-0.70	-0.91	-0.77	0.29	0.09	0.23	-0.71	-0.91	-0.78
3	3	2	-0.66	-0.77	-0.69	-0.66	-0.75	-0.66	0.11	0.04	0.16
3	3	3	-0.59	-0.59	-0.57	-0.53	-0.55	-0.51	-0.62	-0.63	-0.61

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.25: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension non-simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.48	0.50	0.37	0.42	0.42	0.32	0.51	0.50	0.39
-3	-3	-2	0.63	0.59	0.51	0.64	0.59	0.51	0.31	0.34	0.42
-3	-2	-3	0.60	0.67	0.56	0.43	0.44	0.51	0.66	0.67	0.56
-3	-2	-2	0.78	0.80	0.79	0.31	0.31	0.31	0.24	0.29	0.28
-2	-3	-3	0.41	0.39	0.47	0.59	0.57	0.50	0.56	0.61	0.54
-2	-3	-2	0.29	0.29	0.28	0.75	0.77	0.74	0.29	0.31	0.28
-2	-2	-3	0.25	0.22	0.29	0.28	0.24	0.33	0.77	0.79	0.78
-2	-2	-2	0.29	0.30	0.29	0.29	0.26	0.24	0.30	0.32	0.31
-2	-2	-1	0.38	0.43	0.42	0.37	0.42	0.42	0.53	0.50	0.46
-2	-1	-2	0.47	0.46	0.49	0.59	0.61	0.55	0.48	0.47	0.50
-2	-1	-1	0.68	0.66	0.68	0.33	0.35	0.35	0.27	0.30	0.34
-1	-2	-2	0.52	0.50	0.50	0.43	0.47	0.47	0.40	0.47	0.45
-1	-2	-1	0.32	0.32	0.31	0.68	0.67	0.73	0.33	0.31	0.34
-1	0	-1	0.47	0.39	0.44	0.59	0.65	0.59	0.41	0.39	0.46
-1	0	0	0.62	0.62	0.63	0.36	0.37	0.38	0.31	0.32	0.33
0	-1	-1	0.54	0.61	0.55	0.42	0.35	0.41	0.35	0.34	0.37
0	-1	0	0.41	0.38	0.38	0.63	0.66	0.62	0.37	0.35	0.38
0	0	-1	0.32	0.34	0.32	0.36	0.37	0.36	0.56	0.56	0.58
0	0	0	0.23	0.23	0.26	0.22	0.23	0.24	0.24	0.25	0.24
0	0	1	0.29	0.32	0.35	0.35	0.36	0.38	0.60	0.60	0.58
0	1	0	0.36	0.35	0.42	0.67	0.65	0.64	0.36	0.37	0.36
0	1	1	0.54	0.57	0.50	0.43	0.39	0.41	0.40	0.38	0.38
1	0	0	0.55	0.58	0.58	0.42	0.39	0.35	0.35	0.34	0.29
1	0	1	0.40	0.40	0.45	0.62	0.59	0.59	0.39	0.44	0.38
1	2	1	0.33	0.29	0.32	0.68	0.72	0.68	0.32	0.30	0.32
1	2	2	0.51	0.49	0.49	0.52	0.47	0.44	0.47	0.47	0.41
2	1	1	0.68	0.69	0.70	0.33	0.34	0.32	0.27	0.32	0.28
2	1	2	0.50	0.51	0.48	0.57	0.56	0.56	0.49	0.55	0.51
2	2	1	0.43	0.40	0.41	0.42	0.42	0.41	0.48	0.49	0.50
2	2	2	0.25	0.28	0.25	0.24	0.23	0.26	0.31	0.27	0.29
2	2	3	0.24	0.25	0.29	0.27	0.27	0.34	0.82	0.76	0.77
2	3	2	0.29	0.33	0.30	0.72	0.79	0.79	0.31	0.29	0.28
2	3	3	0.34	0.36	0.45	0.65	0.54	0.54	0.63	0.56	0.56
3	2	2	0.74	0.79	0.79	0.31	0.30	0.29	0.27	0.29	0.28
3	2	3	0.67	0.61	0.57	0.41	0.45	0.52	0.64	0.66	0.58
3	3	2	0.62	0.57	0.50	0.61	0.55	0.50	0.31	0.35	0.45
3	3	3	0.51	0.46	0.37	0.44	0.41	0.31	0.53	0.50	0.36

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.26: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension non-simple structure, high correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.59	0.80	0.59	0.54	0.76	0.56	0.59	0.84	0.63
-3	-3	-2	0.75	0.91	0.74	0.74	0.90	0.75	0.38	0.38	0.38
-3	-2	-3	0.82	0.99	0.78	0.41	0.48	0.42	0.84	0.98	0.82
-3	-2	-2	0.96	1.02	0.89	0.35	0.37	0.37	0.35	0.43	0.34
-2	-3	-3	0.42	0.38	0.41	0.74	0.85	0.73	0.78	0.86	0.73
-2	-3	-2	0.36	0.39	0.36	0.97	0.98	0.90	0.41	0.39	0.30
-2	-2	-3	0.36	0.40	0.38	0.32	0.38	0.35	0.94	1.08	0.96
-2	-2	-2	0.36	0.48	0.37	0.35	0.42	0.33	0.42	0.45	0.36
-2	-2	-1	0.55	0.62	0.50	0.56	0.62	0.52	0.45	0.49	0.46
-2	-1	-2	0.61	0.61	0.55	0.53	0.53	0.58	0.60	0.64	0.54
-2	-1	-1	0.78	0.78	0.79	0.35	0.41	0.31	0.35	0.34	0.28
-1	-2	-2	0.50	0.47	0.53	0.57	0.63	0.50	0.55	0.61	0.49
-1	-2	-1	0.34	0.40	0.27	0.75	0.80	0.80	0.35	0.39	0.32
-1	0	-1	0.49	0.52	0.46	0.64	0.62	0.60	0.49	0.48	0.43
-1	0	0	0.68	0.66	0.60	0.38	0.44	0.42	0.38	0.40	0.35
0	-1	-1	0.58	0.59	0.55	0.46	0.51	0.43	0.42	0.45	0.39
0	-1	0	0.46	0.38	0.39	0.69	0.71	0.66	0.40	0.40	0.40
0	0	-1	0.37	0.39	0.38	0.39	0.43	0.40	0.63	0.63	0.63
0	0	0	0.30	0.32	0.23	0.30	0.31	0.23	0.32	0.33	0.26
0	0	1	0.39	0.34	0.36	0.41	0.34	0.38	0.60	0.68	0.66
0	1	0	0.40	0.44	0.39	0.71	0.66	0.65	0.41	0.44	0.40
0	1	1	0.59	0.57	0.59	0.44	0.48	0.45	0.41	0.43	0.42
1	0	0	0.66	0.70	0.63	0.40	0.39	0.41	0.35	0.34	0.33
1	0	1	0.52	0.49	0.41	0.59	0.62	0.66	0.51	0.48	0.44
1	2	1	0.30	0.38	0.29	0.82	0.83	0.79	0.31	0.36	0.27
1	2	2	0.54	0.46	0.49	0.60	0.60	0.56	0.58	0.59	0.55
2	1	1	0.81	0.84	0.80	0.35	0.34	0.32	0.35	0.35	0.29
2	1	2	0.58	0.62	0.54	0.55	0.53	0.56	0.58	0.63	0.54
2	2	1	0.53	0.59	0.56	0.56	0.59	0.58	0.48	0.44	0.48
2	2	2	0.39	0.44	0.38	0.37	0.43	0.34	0.45	0.48	0.39
2	2	3	0.37	0.37	0.36	0.35	0.37	0.32	0.92	1.10	0.96
2	3	2	0.40	0.43	0.32	0.95	1.09	0.97	0.37	0.41	0.36
2	3	3	0.41	0.43	0.32	0.77	0.87	0.80	0.77	0.91	0.82
3	2	2	0.92	1.06	0.94	0.37	0.36	0.32	0.38	0.39	0.34
3	2	3	0.79	0.99	0.82	0.47	0.41	0.38	0.80	0.99	0.84
3	3	2	0.78	0.88	0.77	0.78	0.87	0.74	0.43	0.42	0.35
3	3	3	0.69	0.76	0.66	0.64	0.73	0.60	0.73	0.79	0.69

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

When item exposure control is implemented, similar findings can be observed from Table 5.24 and 5.26. The results suggest the three item pools perform similarly in terms of the ability estimation on the 37 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1, θ_2, and θ_3 are very large or very small, and when θ's are away from each other. A comparison between the condition with and without item exposure control shows, when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger. The reason why the item exposure control increases the estimation error is explained in previous sections.

In summary, this section present the results for the MCAT with the test specification of three-dimension non-simple structure and with high correlation among θ_1, θ_2, and θ_3. In general, the *p*-optimal item pools perform similarly as the baseline pool in terms of both overall and conditional accuracy of ability estimation, but the *p*-optimal item pools can save over 100 items and have a better item pool usage. When item exposure control is implemented, the item exposure rate and item overlap rate can be controlled very well. The *p*-optimal item pools still can provide reliable ability estimation with a relatively small pool size.

A comparison for the results between Test Specification 2 and 3 suggests $\boldsymbol{\theta}$ can be more accurately estimated under the condition of three-dimension non-simple structure. Under the condition of simple structure, the RMSE value is from 0.42 to 0.47 for θ_1, θ_2, and θ_3, and the correlation is about 0.90; under the condition of non-simple structure, the RMSE is less than 0.4 and the correlation is about 0.94. The increase in the estimation accuracy is primary due to the items with $\boldsymbol{a} = (1, 1, 1)$. Because those items provide more information than the items that only measure one θ, there is more information available for the ability estimation. An increase in information will result in a decrease on the measurement error. Another possible explanation for the estimation accuracy is the pool size. The item pools with non-simple structure have about

40-item more than the item pools with simple structure. A larger item pool is expected to yield more accurate ability estimation, because there are more items available for selection.

5.2.6 Performance for item pools based on Test Specification 3 (moderate correlation)

The results for the MCAT based on the test specification of three-dimension non-simple structure, and with θ_1, θ_2, and θ_3 moderately correlated, are presented in Table 5.27 and 5.28. The results in Table 5.27 are under the condition without item exposure control; and Table 5.22 is when item exposure control is implemented. In both tables, there are three values for bias, RMSE and correlation, representing the results for (θ_1, θ_2, θ_3).

Under the condition without item exposure control (see Table 5.27), the *p*-optimal item pools and the baseline pool show no bias on the $\boldsymbol{\theta}$ estimates. Also, the RMSE are between 0.42 and 0.46, and correlations between estimated $\boldsymbol{\theta}$ and true $\boldsymbol{\theta}$ are around 0.89. The average test information between the *.96*-optimal item pool and the baseline pool is very similar, but the information for the *.86*-optimal item pool is slightly smaller. The amount of information on the direction of θ_1, θ_2, and θ_3 (i.e., the value on the diagonal) is about 3.49 for the *.96*-optimal item pool and the baseline pool, and about 3.38 for the *.86*-optimal item pool. In general, the results suggest that the *.96*- and *.86*-optimal item pool can provide accurate estimation for $\boldsymbol{\theta}$, and the level of accuracy is the same as baseline pool, but the average test information for the *.86*-optimal item pool is slightly small than the other two. Table 5.27 also presents the results about item pool usage. Compared with the MCAT based on Test Specification 3 with high correlation, similar results can be drawn from Table 5.27. The item pool usage for the *.96*-optimal item pool is slightly better than the *.86*-optimal item pool. And the two *p*-optimal item pools are much better used than the baseline pool.

Table 5.27: The performance of the *.96*- and *.86*-optimal pool and the baseline pool without exposure control
(3-dimension non-simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.01, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)
RMSE	(0.46, 0.43, 0.42)	(0.46, 0.43, 0.42)	(0.46, 0.43, 0.42)
Correlation	(0.89, 0.90, 0.90)	(0.89, 0.90, 0.90)	(0.89, 0.90, 0.91)
Average test information	$\begin{bmatrix} 3.49 & 1.74 & 1.74 \\ 1.74 & 3.48 & 1.74 \\ 1.74 & 1.74 & 3.48 \end{bmatrix}$	$\begin{bmatrix} 3.39 & 1.60 & 1.60 \\ 1.60 & 3.38 & 1.60 \\ 1.60 & 1.60 & 3.38 \end{bmatrix}$	$\begin{bmatrix} 3.51 & 1.76 & 1.76 \\ 1.76 & 3.49 & 1.76 \\ 1.76 & 1.76 & 3.50 \end{bmatrix}$
Overall Pool Usage	28.82	28.14	40.25
Overlap rate	0.16	0.27	0.13
% of overexposed item (r > 0.2)	12%	32%	2%
% of underexposed item (r<0.02)	33%	29%	45%

Table 5.28: The performance of the *.96*- and *.86*-optimal pool and the baseline pool with exposure control
(3-dimension non-simple structure, moderate correlation)

Statistics	*.96*-optimal pool	*.86*-optimal pool	Baseline pool
Bias	(0.00, -0.01, 0.00)	(0.00, 0.00, 0.00)	(-0.01, 0.00, 0.00)
RMSE	(0.48, 0.44, 0.44)	(0.48, 0.45, 0.43)	(0.47, 0.44, 0.43)
Correlation	(0.88, 0.89, 0.90)	(0.88, 0.89, 0.90)	(0.88, 0.89, 0.90)
Average test information	$\begin{bmatrix} 2.96 & 1.33 & 1.33 \\ 1.33 & 2.96 & 1.33 \\ 1.33 & 1.33 & 2.97 \end{bmatrix}$	$\begin{bmatrix} 2.92 & 1.33 & 1.33 \\ 1.33 & 2.93 & 1.33 \\ 1.33 & 1.33 & 2.94 \end{bmatrix}$	$\begin{bmatrix} 3.15 & 1.60 & 1.60 \\ 1.60 & 3.09 & 1.60 \\ 1.60 & 1.60 & 3.23 \end{bmatrix}$
Overall Pool Usage	3.09	1.12	9.02
Overlap rate	0.09	12%	0.07
% of overexposed item (r > 0.2)	0%	0%	0%
% of underexposed item (r<0.02)	0%	0%	21%

When item exposure control is implemented (see Table 5.28), similar results can be observed: the two *p*-optimal item pools provide as accurate ability estimates as the baseline pool, and yield better item pool usage than the baseline pool. Compared with the condition without item exposure control, item exposure control only results in 0.01 to 0.02 increase in the RMSE, 0.01

decrease in correlation, and about 0.5 decrease in the average test information. For the item pool usage, when the item exposure control is implemented, no item is overexposed, and the percentage of underexposed item and overlapped item are also decreased. The two *p*-optimal item pools have been fully used, and no item from the two p-optimal item pools is underexposed. The comparison between the condition with and without item exposure control suggests the item exposure control can effectively increase the item pool usage and reduce the item exposure rate without obvious loss on the accuracy of ability estimation.

In addition to the overall pool performance, the conditional bias and RMSE at 37 ($\theta_1, \theta_2, \theta_3$) points are also calculated. The conditional bias for each $\boldsymbol{\theta}$ point is presented in Table 5.29 and 5.30, for the MCAT without and with exposure control, respectively. In each table, the conditional bias is color coded based on the value. Negative bias is colored in blue and positive bias is in red. Deeper color represents larger bias. The conditional RMSE is presented in Figure 5.31 and 5.32 in the same manner. Small RMSE is colored in green and large RMSE is colored in red.

Under the condition without item exposure control (see Table 5.29 for bias and 5.31 for RMSE), the conditional bias and RMSE for the .96-, .86-optimal item pool, and the baseline pool are quite similar. This finding supports the results for the overall bias and RMSE, and also suggests the *p*-optimal item pools can provide as accurate ability estimation as the baseline pool at each $\boldsymbol{\theta}$ point. Similar to the results in previous sections, larger bias and RMSE occurs when θ_1, θ_2, and θ_3 are very large or very small, which is the top and the bottom of each table. The difference between θ_1 and θ_2, and between θ_1 and θ_3, also affects the estimation accuracy.

When item exposure control is implemented, similar findings can be observed from Table 5.30 and 5.32. The results suggest the three item pools perform similarly in terms of the ability

Table 5.29: Conditional Bias for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension non-simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.58	0.54	0.36	0.48	0.45	0.31	0.51	0.43	0.31
-3	-3	-2	0.65	0.61	0.54	0.66	0.61	0.47	-0.03	-0.13	-0.25
-3	-2	-3	0.65	0.54	0.49	-0.10	-0.10	-0.27	0.59	0.55	0.43
-3	-2	-2	0.75	0.72	0.66	0.01	0.03	0.04	0.03	-0.01	-0.05
-2	-3	-3	0.03	0.00	-0.18	0.50	0.56	0.38	0.56	0.50	0.41
-2	-3	-2	0.08	0.01	0.03	0.66	0.65	0.63	0.00	-0.01	-0.09
-2	-2	-3	0.12	0.10	-0.03	-0.01	0.00	-0.06	0.70	0.71	0.64
-2	-2	-2	0.16	0.13	0.16	0.16	0.17	0.11	0.20	0.08	0.13
-2	-2	-1	0.29	0.33	0.30	0.33	0.30	0.32	-0.39	-0.37	-0.39
-2	-1	-2	0.39	0.35	0.31	-0.31	-0.36	-0.33	0.38	0.38	0.31
-2	-1	-1	0.50	0.53	0.50	-0.16	-0.07	-0.06	-0.14	-0.14	-0.12
-1	-2	-2	-0.22	-0.28	-0.30	0.30	0.37	0.27	0.28	0.34	0.28
-1	-2	-1	-0.12	-0.08	-0.08	0.54	0.50	0.50	-0.19	-0.19	-0.13
-1	0	-1	0.20	0.18	0.24	-0.40	-0.43	-0.39	0.24	0.23	0.24
-1	0	0	0.44	0.37	0.42	-0.15	-0.15	-0.19	-0.22	-0.17	-0.15
0	-1	-1	-0.37	-0.34	-0.34	0.23	0.24	0.20	0.22	0.26	0.23
0	-1	0	-0.13	-0.14	-0.14	0.46	0.46	0.45	-0.21	-0.27	-0.28
0	0	-1	-0.12	-0.25	-0.17	-0.22	-0.23	-0.24	0.44	0.48	0.45
0	0	0	0.00	0.02	-0.01	0.01	0.00	-0.03	0.03	-0.02	-0.02
0	0	1	0.17	0.19	0.21	0.21	0.23	0.17	-0.51	-0.46	-0.44
0	1	0	0.19	0.12	0.11	-0.44	-0.48	-0.46	0.25	0.19	0.22
0	1	1	0.27	0.33	0.30	-0.19	-0.24	-0.20	-0.23	-0.21	-0.26
1	0	0	-0.40	-0.42	-0.46	0.19	0.15	0.11	0.15	0.16	0.18
1	0	1	-0.25	-0.25	-0.21	0.40	0.40	0.44	-0.34	-0.29	-0.27
1	2	1	0.11	0.14	0.08	-0.52	-0.49	-0.51	0.20	0.17	0.20
1	2	2	0.26	0.26	0.28	-0.29	-0.37	-0.29	-0.23	-0.37	-0.29
2	1	1	-0.55	-0.50	-0.56	0.09	0.07	0.07	0.08	0.12	0.12
2	1	2	-0.42	-0.32	-0.31	0.35	0.34	0.34	-0.36	-0.33	-0.34
2	2	1	-0.34	-0.38	-0.38	-0.40	-0.34	-0.41	0.37	0.35	0.35
2	2	2	-0.12	-0.15	-0.16	-0.14	-0.15	-0.11	-0.06	-0.16	-0.10
2	2	3	-0.01	-0.08	0.03	0.03	0.03	0.10	-0.65	-0.64	-0.61
2	3	2	-0.01	-0.08	-0.01	-0.63	-0.66	-0.62	0.09	0.03	0.13
2	3	3	0.05	0.02	0.15	-0.56	-0.54	-0.37	-0.55	-0.50	-0.37
3	2	2	-0.61	-0.75	-0.64	-0.03	-0.04	0.04	0.01	0.00	0.07
3	2	3	-0.53	-0.67	-0.50	0.11	0.05	0.22	-0.53	-0.63	-0.45
3	3	2	-0.61	-0.62	-0.51	-0.59	-0.62	-0.46	0.13	0.12	0.23
3	3	3	-0.51	-0.53	-0.38	-0.47	-0.48	-0.22	-0.43	-0.43	-0.20

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.30: Conditional Bias for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension non-simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.57	0.77	0.64	0.54	0.69	0.52	0.53	0.70	0.51
-3	-3	-2	0.80	0.88	0.72	0.77	0.83	0.73	-0.12	-0.03	-0.15
-3	-2	-3	0.74	0.82	0.73	-0.07	-0.04	-0.07	0.68	0.81	0.67
-3	-2	-2	0.87	0.97	0.84	0.08	0.13	0.17	0.05	0.15	0.09
-2	-3	-3	0.04	0.08	0.05	0.70	0.79	0.70	0.66	0.77	0.66
-2	-3	-2	0.16	0.24	0.19	0.87	0.96	0.85	0.11	0.16	0.04
-2	-2	-3	0.25	0.24	0.15	0.12	0.21	0.06	0.92	1.01	0.76
-2	-2	-2	0.31	0.35	0.25	0.25	0.37	0.25	0.21	0.31	0.23
-2	-2	-1	0.44	0.51	0.42	0.49	0.53	0.46	-0.25	-0.26	-0.28
-2	-1	-2	0.50	0.49	0.47	-0.29	-0.22	-0.29	0.46	0.46	0.44
-2	-1	-1	0.59	0.66	0.59	-0.08	-0.08	-0.06	-0.09	-0.07	-0.09
-1	-2	-2	-0.23	-0.22	-0.15	0.43	0.44	0.43	0.47	0.43	0.39
-1	-2	-1	-0.01	0.00	-0.06	0.62	0.67	0.58	-0.12	-0.03	-0.09
-1	0	-1	0.29	0.29	0.25	-0.41	-0.41	-0.42	0.34	0.32	0.26
-1	0	0	0.42	0.49	0.49	-0.17	-0.13	-0.16	-0.19	-0.13	-0.18
0	-1	-1	-0.32	-0.31	-0.27	0.31	0.27	0.28	0.31	0.30	0.25
0	-1	0	-0.20	-0.10	-0.18	0.47	0.47	0.53	-0.21	-0.24	-0.17
0	0	-1	-0.22	-0.26	-0.15	-0.22	-0.29	-0.24	0.49	0.53	0.49
0	0	0	-0.02	0.03	0.01	0.00	0.01	0.01	-0.04	0.00	0.03
0	0	1	0.27	0.22	0.14	0.24	0.25	0.30	-0.49	-0.42	-0.45
0	1	0	0.17	0.19	0.21	-0.47	-0.47	-0.53	0.22	0.21	0.25
0	1	1	0.33	0.41	0.30	-0.32	-0.29	-0.29	-0.36	-0.21	-0.31
1	0	0	-0.44	-0.44	-0.46	0.22	0.17	0.23	0.22	0.22	0.22
1	0	1	-0.32	-0.29	-0.34	0.33	0.41	0.39	-0.36	-0.30	-0.33
1	2	1	0.05	0.03	0.06	-0.56	-0.69	-0.61	0.19	0.07	0.13
1	2	2	0.22	0.12	0.21	-0.43	-0.47	-0.44	-0.40	-0.50	-0.42
2	1	1	-0.64	-0.63	-0.60	0.02	0.08	0.08	0.07	0.04	0.10
2	1	2	-0.47	-0.54	-0.45	0.22	0.20	0.25	-0.46	-0.59	-0.43
2	2	1	-0.51	-0.48	-0.46	-0.45	-0.56	-0.47	0.24	0.26	0.29
2	2	2	-0.32	-0.43	-0.33	-0.28	-0.28	-0.22	-0.23	-0.35	-0.27
2	2	3	-0.18	-0.26	-0.14	-0.11	-0.22	-0.07	-0.87	-0.98	-0.91
2	3	2	-0.15	-0.27	-0.11	-0.86	-1.03	-0.86	-0.11	-0.19	-0.05
2	3	3	-0.02	-0.05	-0.06	-0.69	-0.80	-0.78	-0.71	-0.80	-0.77
3	2	2	-0.81	-1.01	-0.89	-0.03	-0.26	-0.16	-0.04	-0.19	-0.13
3	2	3	-0.78	-0.89	-0.81	0.02	-0.11	-0.03	-0.73	-0.85	-0.79
3	3	2	-0.72	-0.93	-0.76	-0.70	-0.88	-0.81	0.12	-0.03	-0.05
3	3	3	-0.61	-0.77	-0.70	-0.56	-0.71	-0.66	-0.56	-0.68	-0.66

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.31: Conditional RMSE for the $\boldsymbol{\theta}$ estimates without exposure control (3-dimension non-simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.68	0.64	0.50	0.60	0.58	0.46	0.63	0.54	0.45
-3	-3	-2	0.74	0.68	0.64	0.72	0.68	0.57	0.32	0.31	0.40
-3	-2	-3	0.74	0.63	0.60	0.31	0.35	0.44	0.67	0.63	0.56
-3	-2	-2	0.83	0.79	0.75	0.28	0.32	0.29	0.27	0.31	0.29
-2	-3	-3	0.34	0.30	0.41	0.62	0.66	0.50	0.66	0.59	0.52
-2	-3	-2	0.37	0.33	0.33	0.73	0.74	0.72	0.32	0.35	0.34
-2	-2	-3	0.31	0.35	0.35	0.30	0.29	0.30	0.78	0.78	0.71
-2	-2	-2	0.41	0.35	0.34	0.37	0.36	0.34	0.36	0.36	0.34
-2	-2	-1	0.45	0.47	0.45	0.41	0.42	0.44	0.49	0.45	0.48
-2	-1	-2	0.52	0.48	0.45	0.40	0.49	0.44	0.48	0.49	0.43
-2	-1	-1	0.61	0.63	0.59	0.31	0.34	0.29	0.33	0.35	0.34
-1	-2	-2	0.42	0.41	0.45	0.41	0.50	0.41	0.40	0.49	0.43
-1	-2	-1	0.33	0.34	0.30	0.62	0.59	0.58	0.37	0.34	0.32
-1	0	-1	0.38	0.39	0.36	0.50	0.53	0.47	0.39	0.38	0.38
-1	0	0	0.53	0.48	0.53	0.33	0.32	0.36	0.36	0.30	0.34
0	-1	-1	0.50	0.46	0.44	0.38	0.42	0.40	0.39	0.38	0.38
0	-1	0	0.35	0.35	0.34	0.55	0.55	0.55	0.36	0.40	0.41
0	0	-1	0.38	0.39	0.36	0.36	0.38	0.40	0.53	0.57	0.54
0	0	0	0.28	0.32	0.32	0.30	0.29	0.31	0.27	0.30	0.31
0	0	1	0.34	0.38	0.35	0.35	0.37	0.35	0.59	0.54	0.53
0	1	0	0.34	0.34	0.37	0.54	0.56	0.55	0.39	0.36	0.37
0	1	1	0.44	0.44	0.44	0.38	0.38	0.37	0.36	0.35	0.38
1	0	0	0.51	0.54	0.55	0.37	0.34	0.28	0.34	0.35	0.33
1	0	1	0.39	0.40	0.37	0.48	0.48	0.56	0.47	0.42	0.38
1	2	1	0.35	0.35	0.35	0.61	0.59	0.61	0.35	0.37	0.36
1	2	2	0.42	0.41	0.44	0.42	0.49	0.40	0.38	0.49	0.43
2	1	1	0.62	0.58	0.66	0.32	0.35	0.31	0.30	0.32	0.34
2	1	2	0.51	0.48	0.48	0.48	0.46	0.48	0.47	0.48	0.45
2	2	1	0.48	0.50	0.50	0.53	0.48	0.53	0.49	0.48	0.45
2	2	2	0.37	0.41	0.39	0.33	0.39	0.33	0.34	0.35	0.33
2	2	3	0.33	0.32	0.38	0.36	0.37	0.32	0.71	0.69	0.68
2	3	2	0.36	0.35	0.34	0.70	0.74	0.68	0.30	0.31	0.34
2	3	3	0.36	0.34	0.37	0.63	0.61	0.48	0.64	0.58	0.50
3	2	2	0.67	0.81	0.70	0.33	0.32	0.33	0.33	0.36	0.34
3	2	3	0.63	0.74	0.61	0.31	0.33	0.41	0.64	0.70	0.53
3	3	2	0.71	0.70	0.60	0.65	0.70	0.57	0.36	0.34	0.40
3	3	3	0.65	0.65	0.52	0.61	0.60	0.37	0.57	0.55	0.37

Note: .96 represents *.96*-optimal pool; .86 represents *.86*-optimal pool; C represents baseline pool

Table 5.32: Conditional RMSE for the $\boldsymbol{\theta}$ estimates with exposure control (3-dimension non-simple structure, moderate correlation)

37 Points			θ_1			θ_2			θ_3		
θ_1	θ_2	θ_3	.96	.86	C	.96	.86	C	.96	.86	C
-3	-3	-3	0.69	0.88	0.73	0.65	0.82	0.65	0.64	0.82	0.63
-3	-3	-2	0.92	0.96	0.85	0.88	0.95	0.82	0.48	0.41	0.43
-3	-2	-3	0.83	0.91	0.82	0.43	0.40	0.40	0.79	0.91	0.76
-3	-2	-2	0.95	1.06	0.93	0.35	0.40	0.38	0.40	0.43	0.36
-2	-3	-3	0.35	0.44	0.33	0.81	0.90	0.78	0.77	0.89	0.73
-2	-3	-2	0.45	0.45	0.40	0.95	1.06	0.93	0.41	0.48	0.36
-2	-2	-3	0.49	0.48	0.37	0.41	0.41	0.33	0.99	1.08	0.84
-2	-2	-2	0.51	0.54	0.47	0.47	0.54	0.43	0.44	0.51	0.43
-2	-2	-1	0.58	0.62	0.56	0.59	0.63	0.57	0.44	0.43	0.42
-2	-1	-2	0.61	0.62	0.59	0.47	0.44	0.47	0.60	0.58	0.52
-2	-1	-1	0.69	0.75	0.68	0.36	0.38	0.33	0.32	0.38	0.32
-1	-2	-2	0.44	0.43	0.42	0.58	0.55	0.53	0.59	0.56	0.51
-1	-2	-1	0.36	0.34	0.35	0.69	0.75	0.66	0.36	0.38	0.33
-1	0	-1	0.48	0.42	0.42	0.54	0.52	0.51	0.49	0.48	0.40
-1	0	0	0.54	0.61	0.56	0.39	0.37	0.35	0.38	0.35	0.35
0	-1	-1	0.49	0.49	0.45	0.43	0.44	0.41	0.45	0.46	0.41
0	-1	0	0.40	0.38	0.38	0.60	0.59	0.63	0.39	0.41	0.36
0	0	-1	0.41	0.42	0.36	0.39	0.45	0.43	0.57	0.62	0.60
0	0	0	0.39	0.35	0.38	0.33	0.30	0.29	0.32	0.31	0.33
0	0	1	0.41	0.41	0.38	0.40	0.44	0.43	0.57	0.52	0.54
0	1	0	0.37	0.39	0.38	0.59	0.58	0.62	0.36	0.39	0.39
0	1	1	0.47	0.52	0.43	0.45	0.44	0.40	0.49	0.42	0.43
1	0	0	0.54	0.55	0.56	0.40	0.34	0.38	0.39	0.40	0.37
1	0	1	0.48	0.46	0.47	0.47	0.53	0.54	0.50	0.46	0.42
1	2	1	0.38	0.35	0.30	0.67	0.78	0.71	0.43	0.36	0.36
1	2	2	0.46	0.41	0.38	0.55	0.60	0.55	0.53	0.61	0.53
2	1	1	0.74	0.73	0.69	0.35	0.39	0.36	0.34	0.34	0.34
2	1	2	0.59	0.66	0.58	0.41	0.42	0.44	0.58	0.71	0.56
2	2	1	0.61	0.63	0.58	0.54	0.67	0.59	0.42	0.46	0.46
2	2	2	0.47	0.58	0.48	0.46	0.48	0.40	0.43	0.51	0.44
2	2	3	0.46	0.50	0.36	0.41	0.49	0.34	0.96	1.06	0.97
2	3	2	0.48	0.49	0.37	0.94	1.11	0.93	0.44	0.39	0.35
2	3	3	0.40	0.43	0.36	0.79	0.89	0.87	0.81	0.90	0.86
3	2	2	0.90	1.09	0.97	0.40	0.48	0.41	0.40	0.46	0.39
3	2	3	0.91	1.00	0.87	0.44	0.43	0.33	0.86	0.96	0.87
3	3	2	0.83	1.05	0.85	0.80	0.97	0.88	0.43	0.45	0.40
3	3	3	0.71	0.89	0.80	0.66	0.80	0.76	0.67	0.79	0.75

Note: .96 represents .*96*-optimal pool; .86 represents .*86*-optimal pool; C represents baseline pool

estimation on the 37 $\boldsymbol{\theta}$ points. In addition, larger bias and RMSE also occurs when θ_1, θ_2, and θ_3 are very large or very small, and when θ's are away from each other. A comparison between the condition with and without item exposure control shows, when item exposure control is built in, the magnitude of the bias and RMSE at some extreme points becomes larger. The reason why the item exposure control increases the estimation error is explained in the previous sections.

In summary, this section presents the results for the MCAT with the test specification of three-dimension non-simple structure and with moderate correlation between θ_1, θ_2, and θ_3. In general, the *p*-optimal item pools perform similarly as the baseline pool in terms of both overall and conditional accuracy of ability estimation, but the *p*-optimal item pools can save over 100 items and have a better item pool usage. When item exposure control is implemented, the item exposure rate and item overlap rate can be controlled very well. The *p*-optimal item pools still can provide reliable ability estimation with a relatively small pool size.

A comparison between the high correlation condition and moderate correlation for Test Specification 3 suggests that, the measurement error significantly increases as the correlation among dimensions decreases. The RMSE increases about one unit, and the correlation decreases about 0.5. One possible explanation is that, when the correlation decreases, the amount of information that can be borrowed among each θ reduces, and thus the estimation accuracy decreases. Although the measurement error for the Test Specification 3 with moderate correlation is large, it is still smaller than the error for the Test Specification 2 with moderate correlation. When θ_1, θ_2, and θ_3 are moderately correlated, adding the cluster of items with $\boldsymbol{a} = (1, 1, 1)$ decreases the RMSE by 0.5 and increase the correlation by 0.3 on average. Therefore, item pools with non-simples structure characteristic yield more accurate ability estimation than the item pools with simple structure.

Chapter 6 Discussion and Conclusion

In this chapter, the simulation results and their implications are discussed. Section 6.1 first summarizes the findings from the simulation study and addresses the research questions. Section 6.2 presents the discussion of results. The implications for item pool development and management are then described in Section 6.3. Finally, the limitations and suggestions for the future research are discussed in Section 6.4

6.1 Summary of Results

This study aimed to generalize the *p*-optimal item pool design method (Reckase, 2003 & 2007) to multidimensional CAT (MCAT). The reason why the *p*-optimal item pool is "*p*-optimal" is because the item pool design is specifically tailored to the adaptive test. And because of this, no single *p*-optimal item pool is universally *p*-optimal. The characteristics of the *p*-optimal item pool are determined by a number of factors such as the examinee population and the algorithms for the adaptive test. Therefore, this study not only designs *p*-optimal item pools for MCAT, but also examines how the *p*-optimal item pool is affected by the test specifications, item exposure control, correlation among dimensions, and bin sizes. The results based on a simulation study are summarized below.

A total of 24 *p*-optimal item pools were designed and then developed in this study. Generally speaking, the item difficulty (i.e., the MDIFF value) was symmetrically distributed with more items located on the middle of the MDIFF scale, and fewer items located on each side. The standard deviation of the MDIFF value was 1.5 to 2.3 times larger than the standard deviation of the target examinee population and the distribution of the MDIFF value was flatter than a standard normal distribution.

The performance of the MCAT using the 24 *p*-optimal item pools was evaluated by comparison with the MCAT using baseline pools through a simulation study. The results showed the MCAT using the *p*-optimal item pools and the MCAT using the baseline pools performed very similarly in terms of the ability estimation accuracy, but the pool size for the *p*-optimal item pools was more than 100-item smaller than the baseline pools. In addition, the item pool usage for all the *p*-optimal item pools was better than the baseline pools.

Specifically, when bin size increased from 0.4 to 0.8, item pool size decreased by 40% on average. Bin size also determined the how much information the best available item in the item pool could provide for ability estimation. A bin size of 0.4 implies the best available item can provide at least 96% of the maximum possible information, and therefore the item pool is called the *.96*-optimal item pool. Similarly, a bin size of 0.8 implies a *.86*-optimal item pool. This is the reason why the average test information yielded for the *.86*-optimal item pools was smaller than the *.96*-optimal item pools in the simulation study. Even though the pool size and the average test information for the *.86*-optimal item pools were smaller, the MCAT using the two types of item pools performed very similar in terms of the accuracy in ability estimation. Similar findings were observed for a unidimensional CAT in Reckase (2010). Because of the small pool size, the item overexposure rate and the item overlap rage for the *.86*-optimal item pools was larger than the *.96*-optimal item pools.

The 24 *p*-optimal item pools were designed based on three test specifications. The pool size for the two-dimension simple structure condition and three-dimension simple structure condition are very similar. For the two-dimensional case, half of the items in the item pool measured θ_1 and another half measured θ_2; For the three-dimensional case, one third of items in the item pool measured each of the three θ's. Therefore, when the test length was the same, the pool size for

the *p*-optimal item pools did change if a cluster of items measuring a different ability was added to the current test. However, when test specification changed from simple structure to non-simple structure, the pool size for the *p*-optimal item pools increased by about 9%. For the three-dimension non-simple structure case, the proportion of items measuring three ability was slightly larger than items measuring only one ability. The measurement error for ability estimation yielded from the *p*-optimal item pools were all within the acceptable range for these three test specifications. The error in the two-dimension simple structure condition was slightly smaller than the three-dimension simple structure condition. This was due to one more θ is estimated in the three dimensional test, but overall test length was the same for the two tests. The error in the three-dimension non-simple structure condition was also smaller than simple structure condition, because the items measuring three ability provided more information for $\boldsymbol{\theta}$ estimation.

A unique factor that influenced the functioning of the MCAT is the correlation among θ's. If ability were highly correlated, the size of the *p*-optimal item pool is about 10% larger than the condition when ability were moderately correlated. Those 10% of items were mainly located on each side of the MDIFF scale, with relatively high or low item difficulty. That is to say, for a MCAT measuring highly correlated ability, a larger number of difficult items and easy items should be created for the *p*-optimal item pool. The ability estimation accuracy in the high correlation condition was better than the moderate correlation condition. Similar results can be found for multidimensional linear test and MCAT in Liu (2007), Segall (2005), Yao (2010), and Yao and Boughton (2007).

When item exposure control was built into the item selection process, the most informative items will not be too frequently selected. In this situation, to ensure the ability estimation accuracy for the adaptive test, another equally informative item should be available in the item

pool. If item exposure control is necessary for a MCAT, the *p*-optimal item pool design can take the item exposure rate into account and adjust the number of item within each MDIFF-bin. The goal is to make sure the there is sufficient number of item in the *p*-optimal item pool to ensure both ability estimation accuracy and test security. Based on the simulation results, when the bin size was 0.4, item exposure control had nearly no influence on the pool size. When the bin size was 0.8, about 20% more items were needed if item exposure control was implemented. These 20% of items were all located on the middle of the MDIFF scale with item difficulty close to 0. If item exposure control was implemented, the measurement error yielded from all *p*-optimal item pools only slightly decreased. This finding suggests the *p*-optimal item pool design is able to balance the ability estimation accuracy and the test security.

6.2 Discussion of Results

The *p*-optimal item pools produced in this study was a union of items that meet all the predetermined psychometrical specifications, and that target to a predetermined examinee population. van der Linen (1999) provided three criteria for an optimal item pool: 1) an optimal item pool should be sufficiently large to allow several thousand overlapping subtests to be drawn from its items; 2) an optimal item pool should consist of items spanning the entire range of item difficulty relative to the population of interest; and 3) an optimal item pool should consist of an appropriate mix of high and low discriminating items to lower the item creation cost while meeting the needs of the ability estimation accuracy.

The first criterion addresses the issues of the item pool size. The findings from the simulation study suggest that the size of the *p*-optimal item pools was affected by a number of factors. For different MCAT programs, the lower limit of optimal item pool size is different. For example, Stocking (1994) recommended the item pool size for a high-stakes CAT should be

approximately 12 times the test length. A longer CAT requires a larger item pool. Also, for high-stakes CAT, item exposure rate is an important issue for test validity and security. When item exposure control is implemented, the item pool should consist of a larger number of items in order to prevent items from being overexposed to examinees. Because a larger item pool tends to solve all these issues, many adaptive testing programs usually develop a very large item pool for operational use. However, a larger item pool does not necessarily increase the ability estimation accuracy. Instead, the pool usage for very large item pool might be undesirable. In this study, for example, the three baseline pools consisted of more than 100 items than the corresponding *p*-optimal item pools. According to the simulation results, baseline pools yielded similar level of measurement accuracy as the *p*-optimal item pools. When item exposure control is not implemented, about half of the items in the baseline pools had exposure rate less than 2%; when item exposure control is implemented, still 20% of items were underexposed. Those underexposed items were wasted because they were very unlikely to be selected. Therefore, item pool design should seek a balance between the demands for a larger item pool, and the potential risk of items being wasted in a larger item pool. The results in this study suggest the design for *p*-optimal item pools can achieve such a balance when item exposure control is considered. The *p*-optimal item pools ensure the ability estimation accuracy and let all the items in the item pool to be fully used.

The second criterion is about the range of item difficulty. As stated in the criterion, the range of item difficulty of a optimal item pool is determined by the examinee population. The standard deviation of the *p*-optimal item pools in this study was 1.5 to 2.3 times larger than the standard deviation of examinees' ability. The range of the item difficulty is from -4.0 to 4.0. Similar results were found by Gu (2009), Reckase (2010) and Zhou (2012) for unidimensional *p*-optimal

item pools. For the baseline pools in this study, the standard deviation of item difficulty was more than 2.5 times larger than that of the ability distribution. Baseline pools consisted of a number of items with extremely high or extremely low item difficulty. These items are useful for examinees with very high or very low ability. However, since those examinees are rare in the population, most of those extreme items are underexposed. Therefore, although the optimal item pool should span the entire range of item difficulty, only a couple of very difficult or very easy items are sufficient.

In addition to the examinee population, the range of item difficulty also depends on the purpose of the test. For licensure exams, the purpose of the test is to classify examinees into two or more categories. If the cut score is in the middle of the θ scale, a large number of items with middle item difficulty should be included in the item pool to ensure the measurement error at the cut score is sufficiently low. In this situation, it is acceptable to drop items with very high or very low item difficulty from the item pool, because they don't contribute much to the measurement accuracy at the cut score. However, if the purpose of the test is to selected gifted student, or to indentify low achieving students, a large number of difficult items or easy items should be added into the item pool, respectively.

The third criterion addresses the issue of item discrimination. Because the MCAT in this study is based on the multidimensional Rasch model, the magnitude of the item discrimination is fixed, but the direction of the item discrimination is not fixed and it can affect the test precision and item creation cost. If an item only loads on one dimension, for instance θ_1, the direction that is best measured by this item is along θ_1. In other words, this item can only discriminate examines with variations on θ_1. If an item loads on more than one dimension such as an item from Cluster 4 with $\boldsymbol{a} = (1,1,1)$, the direction that is best measured by this item is along

direction of the θ_1, θ_2, and θ_3 composite. This item can most effectively discriminate examinees located on different points along the θ_1, θ_2, and $\theta_3$composite line, and can moderate effectively discriminate examinees with variations on θ_1, θ_2, or θ_3. The simulation results in this study suggested that, for a test with simple structure to meet the ability estimation accuracy for all the θ's, the *p*-optimal item pool should consist of the same proportion of items measuring each θ. Compared with tests with simple structure, consisting of items with $\boldsymbol{a} = (1,1,1)$ in the item pool yielded better ability estimation accuracy but increased the pool size at the same time. A larger item pool would cost more to create. Moreover, compared with items measuring only one ability, items with $\boldsymbol{a} = (1,1,1)$ are relatively more difficult to write and cost more to create. Although these items are desirable in psychometrical perspective, they might not be the best choice in practice considering the cost of item creation.

Overall, the *p*-optimal item pools developed in this study met these three criteria, because the item pool design process considered the features of the examinee population, ability estimation accuracy, item pool usage, and the purposes for the test. The size of the *p*-optimal item pools was sufficient for a large number of examinee. Items in the *p*-optimal item pools spanned the entire range of item difficulty. Also, the *p*-optimal item pools yielded acceptable ability estimation accuracy and fairly good item pool usage. Even though the item creation cost is not directly addressed in the item pool design process, it can be controlled by adding a content balancing constrain. For example, if there is an upper limit for the proportion of the expensive items in the item pool, content balancing algorithms are able to control the number of expensive items from being frequently selected, and therefore control the proportion of the expensive items in the item pool.

6.3 Implications

The end product of the *p*-optimal item pool design for MCAT is a bin-count table, which tells the proportion of item from each cluster and the minimum number of items in each item bin. The bin-count table serves as an instructive guide for item creation, item pool development, and item pool management.

Similar to the function of a test blueprint for a linear paper-and-pencil test, the bin-count table is also a target for item creation. Item writers should create items that meet the requirements of the bin-count table. For items measuring only one ability, they can be treated as unidimensional items, so that item writers can create them in the same way they create unidimensional items. Items measuring more than one ability, however, can be difficult to write. When creating items measure more than one ability, the first thing to consider is the direction that is best measured by this item. Items with $\boldsymbol{a} = (1,1,1)$ should measure the three abilities with the same level of discrimination power. In practical, this is very hard to control because more than one strategy may be used to solve an item, and different strategy may require different combinations of these three abilities. Therefore, in this situation, it might be helpful to provide some examples to item writers and give them instructions on how to write items measuring multidimensional abilities. Even though we assume a set of items with $\boldsymbol{a} = (1,1,1)$ is successfully created, these items still cannot be guaranteed to function the same for different groups of examinees. As emphasized in Reckase (2009), "dimensionality is a property of the data matrix, not the test." Although these items are sensitive to differences along the three dimensions, the response data matrix may not be three-dimensional unless there is adequate amount of variation in the examinee sample along each dimension. Because dimensionality is sample-specific, the quality of the examinee sample

for field test is very important. If the sample is not representative, the characteristics of multidimensional items may be greatly affected.

Because items measuring more than one dimension are expensive to create and may be unstable in practice, a *p*-optimal item pool with simple structure might be easier to develop in practice. For an item pool only consisting of items measuring only one ability, some may argue for fitting this item pool with a unidimensional IRT model and treating each cluster of items as one content area. It is feasible to do so, but the advantages of using a multidimensional IRT model are apparent. First, if a unidimensional IRT model is fitted to this item pool, the assumption of unidimensionality might be violated as items are measuring different content areas. Second, if subscores are reported to examinees, MCAT will yield more accurate subscores than UCAT. Third, MCAT can estimate all the θ's simultaneously, but UCAT needs to estimate each θ separately and one at a time. Therefore, MCAT is more efficient in terms of subscore reporting than UCAT. Because of these advantages for MCAT, multidimensional *p*-optimal item pools are more desirable than unidimensional item pools.

In practice, operational item pools are always being renewed. Obsolete items are removed from time to time and new items are filled in accordingly. van der Linden and Veldkamp (2000) summarized that monitoring item usage and replenishing new items are two important tasks for item pool management. The *p*-optimal item pool design presented in this study can be adapted for use in item pool management. If an item located in bin X is retired, a new item should be added to bin X. Because items within each bin are considered to be equivalent in terms of the amount of information they provide for ability estimation, the new item does not need to be identical to the old item; rather, any item that fits into bin X can be used to replace the old item. In this situation, the concept of item bin can reduce the cost for item replenishing. In addition,

when there is a need to create a master pool which supplies several operational item pools, the *p*-optimal item pool design can be used to design the master pool as well. If the master pool needs to supply *N* operational item pools, the size of the master pool would be at least *N* times the *p*-optimal item pool.

6.4 Limitation and Future Studies

The results of this study demonstrated the advantages of using the *p*-optimal item pool design to develop item pools for several MCATs with different features. The results indicate the *p*-optimal item pools can ensure ability estimation accuracy as well as a good item pool usage. This conclusion, however, is restricted by the fact that items are fit by the multidimensional Rasch model. The item discrimination parameter for the multidimensional Rasch model is fixed by test developers, instead of estimated from the data matrix. If inaccurate item discrimination parameters are assigned to some items in the item pool, the extent the ability estimation accuracy would be affected is unknown. Future study can examine the consequences of item discrimination being inaccurately identified. In addition, compared with the multidimensional Rasch model, the multidimensional 2PL or 3PL model tends to fit the data better in practice. Gu (2007) has generalized the *p*-optimal item pool design method to unidimensional 3PL model. It is also worthwhile to generalize Gu's methodology to multidimensional 2PL or 3PL model.

This study is based on the assumption that examinees are multivariate normally distributed. However, in reality, the distribution of examinees is not always normal, and the expected distribution may not always match the reality. The question raised is how robust the *p*-optimal item pool design is to the violation to the shape of the examinee distribution. That is, if a *p*-optimal item pool is developed based on a multivariate normal distribution, but the actual

examinee is not normally distributed, how the performance of the MCAT using this *p*-optimal item pool will be affected. Future study can investigate this issue by a simulation study.

Two bin sizes were considered in this study for the *p*-optimal item pool design. An increase in the bin size will results in a smaller *p*-optimal item pool. The *.86*-optimal item pool in this study yielded similar level of ability estimation accuracy as the *.96*-optimal item pool, if item exposure control was not implemented. If item exposure rate is not an important issue, a smaller item pool is desirable because a smaller item pool will cost less to create. Therefore, it might be interesting to investigate how large the bin size can get before the MCAT does not function well. The results could be useful to determine the bin size in the future.

The item type is this study is purely dichotomous. As the educational measurement area is changing towards to the next generation of assessments, new types of items have emerged and brought significant challenges for test developers. For example, new types of items have been created for the Smarter Balanced tests and will be used operationally. Performance task questions, for example, are one type of new item. A performance task usually requires students to follow several steps to accomplish it. Each step can be treated as one item and the entire task is considered to be a testlet. At this point, it is still unknown how to develop a *p*-optimal item pool for adaptive test consisting of this item type. Since a number of states will soon adopt the Smarter Balanced assessments to replace their current K-12 large-scale standardized assessments, the quality of the item pool is an important issue from both psychometric and policy perspectives. Therefore, the *p*-optimal item pool design for new types of items is definitely a promising direction for future research as well.

APPENDIX

Table A.1: Bin count table for the *.96*-optimal item pool (Test Specification 1, high correlation, without item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0)	3	7	9	11	13	13	14	15	15	14	14	13	12	11	9	7	4
a = (0, 1)	3	7	9	11	13	13	14	15	15	15	14	13	12	11	9	7	4

Table A.2: Bin count table for the *.86*-optimal item pool (Test Specification 1, high correlation, without item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0)	6	11	13	14	15	14	13	11	6
a = (0, 1)	6	11	13	14	15	14	13	11	6

Table A.3: Bin count table for the *.96*-optimal item pool (Test Specification 1, moderate correlation, without item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0)	2	5	8	9	11	12	13	14	14	14	14	12	11	10	8	5	2
a = (0, 1)	2	5	8	9	11	12	13	14	14	14	14	12	11	10	8	5	2

Table A.4: Bin count table for the *.86*-optimal item pool (Test Specification 1, moderate correlation, without item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0)	3	10	13	14	15	14	13	10	4
a = (0, 1)	4	10	13	14	15	14	13	10	3

Table A.5: Bin count table for the *.96*-optimal item pool (Test Specification 2, high correlation, without item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0, 0)	2	4	6	7	8	9	10	10	10	10	10	9	8	7	6	4	2
a = (0, 1, 0)	2	4	6	7	8	9	10	10	10	10	10	9	8	7	6	4	3
a = (0, 0, 1)	1	4	6	7	8	9	10	10	10	10	10	9	8	7	6	4	2

Table A.6: Bin count table for the *.86*-optimal item pool (Test Specification 2, high correlation, without item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0, 0)	3	7	9	10	10	10	9	7	3
a = (0, 1, 0)	4	8	9	10	10	10	9	8	3
a = (0, 0, 1)	3	7	9	10	10	10	9	7	3

Table A.7: Bin count table for the *.96*-optimal item pool (Test Specification 2, moderate correlation, without item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0, 0)	0	3	5	6	7	8	9	10	10	9	9	8	7	6	5	2	1
a = (0, 1, 0)	1	3	5	6	7	8	9	10	10	10	9	8	7	6	5	3	1
a = (0, 0, 1)	1	3	5	6	7	8	9	10	10	10	9	9	7	6	5	3	1

Table A.8: Bin count table for the *.86*-optimal item pool (Test Specification 2, moderate correlation, without item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0, 0)	1	6	8	10	10	10	8	6	1
a = (0, 1, 0)	2	7	9	10	10	10	9	6	1
a = (0, 0, 1)	2	7	9	10	10	10	9	7	2

Table A.9: Bin count table for the *.96*-optimal item pool (Test Specification 3, high correlation, without item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0, 0)	2	4	5	6	7	7	7	8	8	8	7	7	7	6	5	4	2
a = (0, 1, 0)	3	4	5	6	7	7	7	8	8	8	7	7	7	6	5	4	3
a = (0, 0, 1)	2	4	5	6	7	7	7	8	8	8	7	7	7	6	5	4	2
MDIFF	**-5.6**	**-4.9**	**-4.2**	**-3.5**	**-2.8**	**-2.1**	**-1.4**	**-0.7**	**0**	**0.7**	**1.4**	**2.1**	**2.8**	**3.5**	**4.2**	**4.9**	**5.6**
a = (1, 1, 1)	3	4	6	6	7	7	8	8	8	8	8	7	7	6	5	4	3

Table A.10: Bin count table for the *.86*-optimal item pool (Test Specification 3, high correlation, without item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0, 0)	3	6	7	8	8	8	7	6	3
a = (0, 1, 0)	3	6	7	8	8	8	7	6	3
a = (0, 0, 1)	3	6	7	8	8	8	7	6	3
MDIFF	**-5.6**	**-4.2**	**-2.8**	**-1.4**	**0**	**1.4**	**2.8**	**4.2**	**5.6**
a = (1, 1, 1)	4	7	8	9	9	9	8	7	4

Table A.11: Bin count table for the *.96*-optimal item pool (Test Specification 3, moderate correlation, without item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0, 0)	1	3	4	5	6	7	7	7	8	7	7	7	6	5	4	3	1
a = (0, 1, 0)	1	3	4	5	6	7	7	7	7	7	7	7	6	5	4	3	1
a = (0, 0, 1)	2	3	4	5	6	7	7	7	8	7	7	7	6	6	5	3	1
MDIFF	**-5.6**	**-4.9**	**-4.2**	**-3.5**	**-2.8**	**-2.1**	**-1.4**	**-0.7**	**0**	**0.7**	**1.4**	**2.1**	**2.8**	**3.5**	**4.2**	**4.9**	**5.6**
a = (1, 1, 1)	1	3	5	6	7	7	8	8	8	8	7	7	7	6	5	3	1

Table A.12: Bin count table for the *.86*-optimal item pool (Test Specification 3, moderate correlation, without item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0, 0)	2	5	7	8	8	8	7	5	2
a = (0, 1, 0)	2	5	7	8	8	8	7	5	2
a = (0, 0, 1)	2	5	7	8	8	8	7	5	3
MDIFF	**-5.6**	**-4.2**	**-2.8**	**-1.4**	**0**	**1.4**	**2.8**	**4.2**	**5.6**
a = (1, 1, 1)	2	6	8	9	9	9	8	6	2

Table A.13: Bin count table for the *.96*-optimal item pool (Test Specification 1, high correlation, with item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0)	4	6	9	11	12	13	14	15	17	15	14	13	12	11	9	7	3
a = (0, 1)	4	7	9	11	12	13	14	15	17	15	14	13	12	11	9	7	3

Table A.14: Bin count table for the *.86*-optimal item pool (Test Specification 1, high correlation, with item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0)	6	11	13	17	32	17	13	11	6
a = (0, 1)	6	11	13	17	32	17	13	11	6

Table A.15: Bin count table for the *.96*-optimal item pool (Test Specification 1, moderate correlation, with item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0)	2	5	8	10	11	12	13	14	18	14	13	12	11	9	8	5	2
a = (0, 1)	2	5	8	9	11	12	13	14	18	14	13	12	11	10	7	5	2

Table A.16: Bin count table for the *.86*-optimal item pool (Test Specification 1, moderate correlation, with item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0)	4	10	13	18	33	18	13	10	4
a = (0, 1)	4	10	13	18	33	18	13	10	4

Table A.17: Bin count table for the *.96*-optimal item pool (Test Specification 2, high correlation, with item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0, 0)	2	4	6	7	8	9	10	10	12	10	10	9	8	7	6	4	2
a = (0, 1, 0)	2	4	6	7	8	9	10	10	12	10	10	9	8	7	6	4	2
a = (0, 0, 1)	1	4	6	7	8	9	10	10	12	10	10	9	8	7	6	4	1

Table A.18: Bin count table for the *.86*-optimal item pool (Test Specification 2, high correlation, with item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0, 0)	3	7	9	12	21	12	9	7	3
a = (0, 1, 0)	3	8	9	12	21	12	9	8	3
a = (0, 0, 1)	3	7	9	12	21	12	9	7	3

Table A.19: Bin count table for the *.96*-optimal item pool (Test Specification 2, moderate correlation, with item exposure control)

MDIFF	-3.2	-2.8	-2.4	-2	-1.6	-1.2	-0.8	-0.4	0	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2
a = (1, 0, 0)	0	2	4	6	7	8	9	10	12	10	9	8	7	6	5	2	1
a = (0, 1, 0)	1	3	5	6	7	9	9	10	12	10	9	8	7	6	5	3	1
a = (0, 0, 1)	1	3	5	6	8	9	9	10	12	10	9	9	7	6	5	3	1

Table A.20: Bin count table for the *.86*-optimal item pool (Test Specification 2, moderate correlation, with item exposure control)

MDIFF	-3.2	-2.4	-1.6	-0.8	0	0.8	1.6	2.4	3.2
a = (1, 0, 0)	1	6	8	12	22	12	8	6	1
a = (0, 1, 0)	2	6	9	12	22	12	9	6	2
a = (0, 0, 1)	2	6	9	12	22	12	9	6	2

Table A.21: Bin count table for the *.96*-optimal item pool (Test Specification 3, high correlation, with item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0, 0)	2	4	5	6	7	7	7	8	8	8	7	7	7	6	5	4	2
a = (0, 1, 0)	3	4	5	6	7	7	7	8	8	7	7	7	7	6	5	4	3
a = (0, 0, 1)	2	4	5	6	7	7	7	8	8	8	7	7	7	6	5	4	2
MDIFF	**-5.6**	**-4.9**	**-4.2**	**-3.5**	**-2.8**	**-2.1**	**-1.4**	**-0.7**	**0**	**0.7**	**1.4**	**2.1**	**2.8**	**3.5**	**4.2**	**4.9**	**5.6**
a = (1, 1, 1)	3	4	6	6	7	7	8	8	9	8	8	7	7	6	5	4	3

Table A.22: Bin count table for the *.86*-optimal item pool (Test Specification 3, high correlation, with item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0, 0)	3	6	7	9	15	9	7	6	4
a = (0, 1, 0)	4	6	7	9	14	9	7	6	4
a = (0, 0, 1)	3	6	7	9	15	9	7	6	3
MDIFF	**-5.6**	**-4.2**	**-2.8**	**-1.4**	**0**	**1.4**	**2.8**	**4.2**	**5.6**
a = (1, 1, 1)	4	7	8	10	17	10	8	7	4

Table A.23: Bin count table for the *.96*-optimal item pool (Test Specification 3, moderate correlation, with item exposure control)

MDIFF	**-3.2**	**-2.8**	**-2.4**	**-2**	**-1.6**	**-1.2**	**-0.8**	**-0.4**	**0**	**0.4**	**0.8**	**1.2**	**1.6**	**2**	**2.4**	**2.8**	**3.2**
a = (1, 0, 0)	1	3	4	5	6	7	7	7	9	7	7	7	6	5	4	3	1
a = (0, 1, 0)	1	3	4	5	6	7	7	7	8	7	7	7	6	6	4	3	1
a = (0, 0, 1)	1	3	5	6	6	7	7	7	8	7	7	7	6	6	5	3	1
MDIFF	**-5.6**	**-4.9**	**-4.2**	**-3.5**	**-2.8**	**-2.1**	**-1.4**	**-0.7**	**0**	**0.7**	**1.4**	**2.1**	**2.8**	**3.5**	**4.2**	**4.9**	**5.6**
a = (1, 1, 1)	1	3	5	6	7	7	8	8	10	8	7	7	7	6	5	3	1

Table A.24: Bin count table for the *.86*-optimal item pool (Test Specification 3, moderate correlation, with item exposure control)

MDIFF	**-3.2**	**-2.4**	**-1.6**	**-0.8**	**0**	**0.8**	**1.6**	**2.4**	**3.2**
a = (1, 0, 0)	2	5	7	9	15	9	7	5	2
a = (0, 1, 0)	2	5	7	9	15	9	7	5	2
a = (0, 0, 1)	2	5	7	9	15	9	7	5	2
MDIFF	**-5.6**	**-4.2**	**-2.8**	**-1.4**	**0**	**1.4**	**2.8**	**4.2**	**5.6**
a = (1, 1, 1)	2	6	8	10	18	10	8	6	2

REFERENCES

Adams RJ, Wilson M, Wang W (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement* 21:1-24.

Ansley, T.N. (1 984). Using a unidimensional latent trait model with multidimensional data: An empirical investigation of robustness. Unpublished doctoral dissertation, University of Iowa, Iowa city, IA.

Bejar, I. I, & Weis, D. J. (1979). *Computer programs for scoring test data with item characteristic curve models* (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric methods Program.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

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

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.

Buyske, S. (2005). Optimal design in educational testing. In M. P. F. Berger & W. K. Wong (Eds.), *Applied optimal designs* (pp. 1-19). West Sussex, UK: Wiley.

Common Core State Standards Initiative (2010). *Common Core State Standards for Mathematics*. Washington, DC: CCSSO & National Governors Association.

Chang, H.-H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58, 37-52.

Chang, H.-H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, *20*(3), 213–229.

Chang, H.-H., & Ying, Z. (1999). a-Stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement*, *23*(3), 211 –222.

Chen, S.-Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, 41(2), 149-174.

Chen, S.-Y., Ankenmann, R. D., & Chang, H.-H. (2000). A Comparison of Item Selection Rules at the Early Stages of Computerized Adaptive Testing. *Applied Psychological Measurement*, *24*(3), 241–255.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 63, 369-383.

Childs, R. A., & Oppler, S. H. (2000). Implications of Test Dimensionality for Unidimensional Irt Scoring: An Investigation of a High-Stakes Testing Program. *Educational and Psychological Measurement*, *60*(6), 939–955.

Diao, Q. (2009). *Comparison of ability estimation and item selection methods in MCAT* (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.

Frey, A., Cheng, Y. & Seitz, N. (2011). *Content Balancing with the Maximum Priority Index Method in Multidimensional Adaptive Testing*. Presented at the 2011 meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Frey, A., & Seitz, N.-N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation*, *35*, 89–94.

Georgiadou, E., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing from 1983 to 2005. *The Journal of Technology, Learning, and Assessment, 5(8).*

Gordon Commission (2013). To assess, to teach, to learn: a vision for the future of assessment. The Gordon Commission, Princeton, NJ. . Retrieved from: http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Gu, L. (2007). *Designing optimal item pools for computerized adaptive tests with exposure controls* (Unpublished doctoral dissertation). Michigan State University.

He, W. (2010). *Optimal item pool design for a highly constrained computerized adaptive test* (Unpublished doctoral dissertation). Michigan State University.

Hetter, R., & Sympson, B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. Wasters, & J. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22, 79-86.

Leung, C. K., Chang, H., & Hau, K. T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, 63, 257-270.

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

Luecht, R. M. (1998). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389–404.

Mao, L., Luo, X. & Zhou, X. (2013). *The Comparison of the Unidimensional and Multidimensional CAT in terms of Composite Score Estimation*. Paper presented at the 2013 Annual Meeting of the American Educational Research Association, San Francisco, CA.

Mulder, J., & van der Linden, W. (2009). Multidimensional Adaptive Testing with Optimal Design Criteria for Item Selection. *Psychometrika*, *74*(2), 273–296.

National Assessment of Educational Progress (NAEP). (2010). Retrieved from http://nces.ed.gov/nationsreportcard/tdw/analysis/2007/scaling_determination_correlations_math2007conditional.asp

Patsula, L. N., & Steffan, M. (1997). *Maintaining item and test security in a CAT environment: A simulation study*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Rasch G (1962). On general laws and the meaning of measurement in psychology. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* 4: 321-334.

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

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement Issues and Practice*, 8(3), 11-15.

Reckase, M. D. (2003). Item pool design for computerized adaptive tests. Paper presented at the 2003 Annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

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.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Segall, D. O. (2010). Principles of Multidimensional Adaptive Testing. In *Elements of Adaptive Testing* (pp. 57–75). New York: Springer.

Segall, D. O., Moreno, K. E., & Hetter, D. H. (1997). *Item pool development and evaluation*. In W. A.Sands, B. K.Waters, & J.R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation (pp. 117–130). Washington DC: American Psychological Association.

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling, 55*(1), 105–123.

Smarter Balanced Assessment Consortium (2013). *On Track and Moving Forward: The Smarter Balanced Assessment System*. Retrieved from http://www.smarterbalanced.org/resources-events/publications-resources/

Song, T. (2010). *The effect of fitting a unidimensional IRT model to multidimensional data in content-balanced computerized adaptive testing* (Unpublished doctoral dissertation). Michigan State University.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep No. 94-05). Princeton, NJ: Educational Testing Service.

Svetina, D. (2013). Assessing Dimensionality of Noncompensatory Multidimensional Item Response Theory With Complex Structures. *Educational and Psychological Measurement, 73*(2), 312–338.

Swanson, D., & Stocking, M. L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.

Sympson, J. (1978). *A model for testing with multidimensional items.* In D.J Weiss (Ed.), Proceedings of the 1977 computerized adaptive testing conference. Minneapolis MN: University of Minnesota.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 27th Annual Meeting of the Military Testing Association, San Diego, CA.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 14(2), 181-196.

van der Linden, W.J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24(4), 398–412.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.

van der Linden, W., & Glas, C. (2010). *Elements of Adaptive Testing*. New York, NY: Springer.

van der Linden, W. J., & Reese, L. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22 (3), 259-270.

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In *Computerized Adaptive Testing: Theory and Practice* (pp. 149–166). Dordrecht: Kluwer.

Veldkamp, B., & Linden, W. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575–588.

Wainer, H. (1990). Computerized adaptive testing: A primer. In. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, N. J.: Lawrence Erlbaum Associates.

Wang, C., & Chang, H. (2011). Item selection in multidimensional computer adaptive testing—gaining information from different angles. *Psychometrika*, *76*(3), 363–384.

Wang, W. C., & Chen, P. H. (2004). Implementation and Measurement Efficiency of Multidimensional Computerized Adaptive Testing. *Applied Psychological Measurement*, *28*, 450–480.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17, 17-27.

Weiss, A.A., 1982, Asymptotic theory for ARCH models: Stability, estimation and testing, Discussion paper 82-36 (University of California, San Diego, CA).

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, *18*(2), 93–113.

Yao, L. (2012). *Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints*. Paper presented at the 2012 Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada.

Yao, L. (2013). Comparing the Performance of Five Multidimensional CAT Selection Procedures With Different Stopping Rules. *Applied Psychological Measurement*, *37*(1), 3–23.