ITEM SELECTION METHODS IN MULTIDIMENSIONAL COMPUTERIZED ADAPTIVE TESTING ADOPTING POLY TOMOUSLY-SCORED ITEMS UNDER MULTIDIMENSIONAL GENERALIZED PARTIAL CREDIT MODEL

BY

HAIYAN LIN

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Educational Psychology in the Graduate College of the University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Katherine Ryan, Chair
Professor Hua-Hua Chang, Director of Research
Professor Carolyn Anderson
Professor Jeff Douglas
ABSTRACT

Four item selection methods are compared and investigated under three test formats in the context of Multidimensional Computerized Adaptive Testing (MCAT) delivering polytomous items partially or completely in tests. Item selection methods examined include Fisher information based D-optimality (D-optimality), Kullback-Leibler information index (KI), mutual information (MI), and continuous entropy method (CEM). The three test formats considered are the POLYTYPE format that contains polytomous items with three response categories, the DPMIX format that delivers dichotomous items at the beginning and polytomous items at the final stage, and the PDMIX format that has the reverse order as DPMIX. In general, D-optimality shows the best estimation accuracy and conditional estimation accuracy. D-optimality, MI, and CEM are similar in terms of ability estimation accuracy and tendency in selecting items when the item bank size is large. For both dichotomous and polytomous items, KI is mostly outperformed by the other three methods in terms of ability estimation precision. When sub-thetas in both dimensions are equal, however, KI shows the best performance for polytomous items. In this study, which item type, dichotomous or polytomous, being administered first does not affect the estimation accuracy. However, if the test length is much longer or shorter than the test length of the current study, it is possible that the estimation accuracy could be affected by the order of delivering different item types. Both DPMIX and PDMIX formats yield similar conditional estimation accuracy pattern and precision. In addition, the item bank size does affect the estimation precision. These conclusions, however, might not be applied to MCAT testing with different test designs or item pool structures. More studies are needed in MCAT combining with polytomous items to further facilitate the development and improvement of the next-generation assessments such as formative assessment or testing for diagnosis.
To my dear parents, my loving husband, and my adorable Aileen & Adam
ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance of my committee members, help from friends, and support from my family.

First and foremost, my utmost gratitude goes to Dr. Hua-Hua Chang and Dr. Katherine Ryan. They are the best teachers, advisors, and friends. They have been supporting and encouraging me without any holding back throughout the entire journey of my Ph.D. education. Dr. Ryan created an excellent atmosphere for learning and research, academically, spiritually, and financially. Dr. Chang led me to the most cutting-edge field and shared precious experiences without any reservations. My gratitude also goes to my committee members, Dr. Carolyn Anderson and Dr. Jeff Douglas for their insightful advice and help on my dissertation. Also, they have been providing great teaching and strong support in all aspects throughout my Ph.D. years.

I give special thanks to my parents, Enyuan Zhang and Meihua Lin for their unconditional love and support. I am very lucky for being their daughter.

I am extremely grateful to have the best husband in the world, Lei Zhao. He was always there cheering me up and stood by me through the good times and bad. He sacrificed many things in his life to support me to pursue Ph.D. education.

My sweetest thanks go to my daughter Aileen Yin Lin Zhao and my son Adam Haowei Zhao. They are the meaning of my life and the main reason motivating me to get the Ph.D. education. I did not have time for them over these years. My daughter got used to play with her daddy. I even did not have time combing her hair. I missed many important moments in my son's first two years. One of several sentences my son began to say is "Mom is busy". They, however, gave their mom the greatest tolerance and patience.

This work is dedicated to them.
# TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION ................................................................. 1  
CHAPTER 2 REVIEW OF THE LITERATURE ....................................... 10  
CHAPTER 3 METHODOLOGY ............................................................. 51  
CHAPTER 4 RESULTS, FINDINGS, AND DISCUSSIONS ......................... 64  
CHAPTER 5 LIMITATIONS AND FUTURE STUDIES ............................. 108  
REFERENCES .................................................................................. 112
CHAPTER 1
INTRODUCTION

A new generation of assessments have been gradually built up to meet the new capabilities of technology and to respond to accountability concerns of promoting learning through testing. The development in technology has provided irreplaceable efficiency and convenience so that it is imperative to embed technology into assessment systems. In addition to the previous function of educational testing as measuring student proficiency, it has been called on playing a role in improving instruction besides the previous function of measuring student proficiency. Quellmalz and Pellegrino (2009) pointed out that the next-generation of testing should offer "the potential for transforming what, how, when, where, and why testing occurs" (p.75) and "expand the potential for test to both probe and promote a broad spectrum of human learning, including the types of knowledge and competence advocated in various recent policy reports on education and the economy"(p.75). The next-generation of technology-powered assessment systems will adopt complex, multifaceted problem types and assessment approaches that were impossible missions during the era of paper-and-pencil tests. One example of innovative items types of investigating multiple abilities is in Figure 1.1 that was provided by Quellmalz and Pellegrino (2009). The question asked a student to determine how different payload masses influence the altitude of the balloon based on a provided scenario. To answer the question, students need to design an experiment, manipulate parameters, conduct experiments, record data, and show the results through graphing. Figure 1.1 also shows the availability of types of data offered to students before reaching the final conclusions and responses (Quellmalz & Pellegrino, 2009). Such innovative item types were already included in 2009 NAEP science exam (Quellmalz & Pellegrino, 2009).
Figure 1.1. One example of innovative items in assessments.

It is an exciting change toward item types and educational assessment systems that also could change the direction of education. However, new challenges come along with the technology-based revolutionary change. Accordingly, item response theory (IRT) models have been extended to measure several latent traits, the multidimensional case, and incorporate polytomously scored items.

Why Multidimensional Approaches?

Multidimensional item response theory (MIRT) has gained attention because it measures multiple abilities required by the items of the formative assessment, performance-based testing, and testing for diagnosis. There are two lines of research directions that diagnostic testing: the cognitive diagnostic (CD) approach and multidimensional item response theory (MIRT) approach. The CD approach has been studied in various fields, but most of the CD models only
provide dichotomous master/non-master reports (Wang, Chang, & Boughton, 2011). In contrast, the MIRT method could obtain a continuous estimate of each subscale so that as much information as possible about multiple abilities is extracted from every examinee (Wang, Chang, & Boughton, 2011). Multidimensional adaptive testing thus could provide greater efficiencies toward complicated and time-intensive tasks. MIRT has been studied for several decades (Mcdonald, 1967, 1997; Reckase, 1985, 1995; Samijima, 1974; etc.), but a lack of computational power was the main hindrance to the application and development of MIRT in the testing field (Mulder & van der Linden, 2010). However, with the current technology capacities, computational power is no longer a concern and the advantages of MIRT could be fully utilized (Mulder & van der Linden, 2010). The focus of this study is the MIRT approach.

MIRT stems from both factor analysis and unidimensional item response theory (UIRT) and the interpretation of MIRT analysis results is akin to the UIRT (Reckase, 2009). Generally speaking, IRT adopts a set of mathematical models to describe the interaction between a person and test items. In UIRT, the person latent trait variable is a scalar ability parameter denoted as \( \theta \). Geometrically, an examinee’s latent ability is located along a score scale in UIRT. In MIRT, however, the latent ability is multidimensional vector, denoted as \( \theta_j = (\theta_{j_1}, \theta_{j_2}, \ldots, \theta_{j_p})^T \) where \( p \) is the number of dimensions or subscales of latent abilities, analogous to the number of attributes in CD (Wang, Chang, & Boughton, 2011). Geometrically, an examinee’s latent trait is located on a plane or hyperplane in MIRT. An item is selected to ideally minimize the joint estimation errors for latent trait estimates on all ability dimensions or subscales.

Multidimensional computerized adaptive testing (MCAT) integrates MIRT into computerized adaptive testing (CAT). There are several unique and compelling advantages of MCAT over UCAT: (a) MCAT provides more information than UCAT. The multiple ability
dimensions measured in MCAT are often correlated. Added information provided by items of correlated dimensions would result in greater measurement efficiency such as greater precision or reduced test-length (Segall, 1996). For example, Segall (1996) compared multidimensional Bayesian simulation results with unidimensional simulation results based on the computerized adaptive testing version of Armed Services Vocational Aptitude Battery (CAT-ASVAB) test. The study found out that when dimensions are correlated MCAT can achieve equivalent to higher levels of precisions with around one-third fewer items than UCAT. When the test length is the same in MCAT and UCAT, MCAT provides a substantially higher reliability. (b) MCAT can automatically provide an efficient choice of items ensuring adequate content coverage without fully forcing the content balancing techniques commonly used in UCAT. In MCAT, content balance is based on the intermediately estimated level of proficiency and examinees at different levels are administered an appropriately tailored mixture of item content. Moreover, MCAT item selection also provides items of appropriate difficulty levels. In contrast, UCAT has the problem of selecting items of inappropriate difficulty due to the content balancing constraints so that items administered from some content area might provide little information about the general level of proficiency of examinees. MCAT treats desired content areas as separate but highly intercorrelated dimensions and incorporates information from several sources on all dimensions simultaneously (Segall, 1996; Wang & Chang, 2011).

**Why Polytomously-Scored Items?**

With the understanding of the imperativeness of incorporating innovative items, the polychotomous IRT models will play an extremely important role in educational testings. Polytomously scored items include performance tasks, selected responses, brief or extended constructed response items, essay, fill-in-the-blank. Polytomous items require examinees to use
verbal, mathematical, or figural components to construct a response. A type of information about the examinee's knowledge that dichotomous items cannot detect could be potentially obtained. In addition, many of these polytomous or innovative item types can be incorporated within an adaptive computerized assessment (Parshall, Harmes, Davey, & Pashley, 2010). For dichotomous items, the probability of a correct response given by an examinee can be described by one of the logistic IRT models, such as Rasch model, one-, two- and three-parameter logistic (1PL, 2PL, and 3PL) IRT models. For polytomous items, the probability that an examinee reach a specific score category is provided by polytomous IRT models (Tang, 1996).

Polytomously-scored items have been broadly used in a variety of exams. For example, based on a survey by Lane (2005), 63% of the state assessments use both dichotomously-scored items and polytomously-scored items. Despite of higher cost, polytomous items have several desirable features over dichotomous items. From the point of view of accountability, with polytomous items, the issue of "teaching to the test" criticized heavily in the era of No Child Left Behind (NCLB) would be less of a problem (Lazer, 2010). There are several strengths from a psychometric perspective as well. First, polytomously-scored items provide more information than dichotomously scored items. Samejima (1976) showed that polytomous items provided considerably more IRT information than the optimal dichotomization of the same items scored dichotomously by using the graded response model (GRM). Birenbaum and Tatsuoka (1987) found that polytomously-scored items provide more diagnostic information than dichotomously-scored items in a study using real data. Donogue (1994) found that four-category polytomous items provided 2.1 to 3.1 times as much IRT information as dichotomous items. Polytomous items yielded the most information about examinees with moderately high proficiency when the population mean is zero (Donogue, 1994). Second, it is believed that polytomous items measure
concepts and skills at greater depth than dichotomous items (Ercikan et al., 1998). Furthermore, a common claim is that polytomous items might measure more than one trait, unlike dichotomous items (Traub, 1993; Yao & Schwarz, 2006). Third, De Ayala (1989, 1992) argued that polytomous items could reduce the test length while achieving the same effects, particularly under the CAT context. Fourth, van Rijin et al. (2002) found that dichotomous item banks showed more bias for extreme ability values than polytomous item banks and mixed item banks. Fifth, polytomous items might adjust item exposure. When mixed item banks are used, items with more categories are selected more frequently in terms of item exposure (van Rijin et al., 2002). Furthermore, the mixture of various innovative item types is required in tests. For example, in the Race to the Top Assessment Program (RTTT) program, the common core state (CCSS) requires information on students' abilities in a variety of areas, including problem solving, conducting critical analyses, etc. Numerous studies have investigated polytomous UCAT (PUCAT) (e.g. Dodd, Koch, & De Ayala, 1988; De Ayala, 1989, 1992; De Ayala, Dodd, & Koch, 1992; Koch, Dodd, & Fitzpatrick, 1990; Jodoin, 2003; etc.).

The measurement of multiple abilities and the employment of polytomously-scored items in exams becomes increasingly significant since it occurs both in K-12 system and other high-stakes exams. In K-12 context, the RTTT announced in 2010 by the U.S.Department of Education (USDE) urges the assessment consortia to develop an integrated state assessment system with dual needs of accountability and instructional improvement (K-12 Center, 2010). With one main purpose of obtaining a richer, more intelligent, more nuanced picture in terms of what students know and can do (Lazer et al., 2010), the formative or performance-based assessment becomes a mandatory assessment component under the RTTT framework. For such assessments, multiple abilities are required to solve the test items (Mulder & van der linden,
2009; Yao & Schwarz, 2006) and thus measuring multiple latent abilities is a necessary field to be addressed in the RTTT context. In addition, the technology-driven assessment systems coupling with the innovative assessment task design is a key RTTT requirement, making the adoption of polytomously-scored items a desirable trend. Another example is that NAEP 2011 writing assessment requires the usage of word processing and editing pool to write essays (Quellmalz & Pellegrino, 2009). In the field of high-stakes credentialing exams, several professional admission and certification boards have begun to utilize web-based diagnostic services to obtain more informative diagnostic profiles of latent abilities in order to improve the high-stakes tests (Mulder & van der linden, 2009). For instance, one part of an architecture licensure exam is to use computer assisted designed programs (Quellmalz & Pellegrino, 2009). Other examples include the AICPA, the NCARB, the USMLE (Parshall, Harmes, Davey, & Pashley, 2010). Therefore, it is important to explore MIRT models and polytomously-scored items in the current and future technology-embedded testing.

In polytomous CAT field, however, not many studies are using the real data. Most studies remain in the marketing field with likert-type survey data (i.e., integer scored items) assessing attitude (Dodd et al., 1995). Few studies have been discussed in K-12 setting, and few studies are in high-stakes environment. This might explain why most studies did not include the content constraint limitations. In addition, some promising item selection methods such as mutual information or continuous entropy method have not been applied and compared.

The main motivation of this current study is to explore item selection issues in MCAT in the context of high-stakes educational testing, mainly for K-12 assessment system or similar versions of licensure exams. One approach is to explore the administration of innovative items in MCAT so that one test format studied is uni-type test format where only polytomously-scored
items are administered, called POLYTYPE test format. Another test format, mix-type test format that delivers both dichotomous and polytomous items, is studied because this is a more real testing context in that many types of assessment contain a mixture of dichotomous and polytomous items (Yao & Schwarz, 2006). Under the mix-type test format, two categories are examined. One category is the MCAT test delivering dichotomous items at the beginning of the test and administering polytomous items afterwards, named DPMIX test format. For the other category, the order of item types is reverse so that polytomous items and dichotomous items are the first and second delivered type respectively, called PDMIX test format in the study.

The CAT, regardless of the IRT model used, consists of four major components: 1) item bank; 2) item selection procedure; 3) latent trait estimation procedure; and 4) stopping rule (Kingsbury & Zara, 1989, 1991; Dodd, De Ayala, & Koch, 1995; Reckase, 1989; Wainer et al., 1990; Weiss, 1982). Among these four elements, the item selection procedure is a core element. This study aims to address the following research questions related to item selection methods.

**Research Questions**

Research Question 1: Among four item selection methods, (i.e., Fisher Information based D-optimality (D-optimality), Kullback-Leibler Information (KI), Mutual Information(MI), and Continuous Entropy method (CEM)), which method achieves the best estimation accuracy and conditional estimation accuracy for the POLYTYPE MCAT test? Are these item selection methods similar to each other in terms of ability estimation accuracy and item selection pattern?

Research Question 2: Among three test formats, POLYTYPE, DPMIX, PDMIX, which test format provides the best estimation accuracy given a fixed test length? For the mixed-type test formats containing both dichotomous items and polytomous items, does the order of delivering item types affect estimation accuracy and conditional estimation accuracy?
The four item selection methods are selected due to either high level of popularity or outstanding performance in previous studies. In addition, Wang and Chang (2011) explored these four methods using multidimensional three-parameter logistic model (M3PL) that represents responses to dichotomous items. This study intends to extend the study by Wang and Chang (2011) to contexts of only administering polytomous items and delivering both polytomous and dichotomous items in high-stakes MCAT.

In this study, UCAT means a unidimensional computerized adaptive test and MCAT refers to a multidimensional computerized adaptive test. DUCAT is UCAT using dichotomously-scored items and PUCAT means UCAT delivering polytomously-scored items. Similarly, PMCAT is MCAT delivering polytomously-scored items and DMCAT is MCAT administering dichotomously-scored items. The uni-type test format refers to a test administering only one type of items and the mix-type test format is a test administering more than one type of items. In the following chapters, $S_{k-1}$ is the set of first k-1 items administered in the test; $u_{k-1}$ represents the response vector to the first k-1 items; $R_k$ means the set of candidate items in the item bank from which item is picked. Also, $\theta$ is a scalar representing the true parameter describing the person characteristics of examinee j under UIRT models while $\theta$ is a vector representing the true parameter under MIRT models. Similarly, $\theta^{Hat}$ is a scalar representing the estimated parameter describing the person characteristics of examinee j under UIRT models while $\theta^{Hat}$ is a vector representing the estimated parameter under MIRT models.
CHAPTER 2

REVIEW OF THE LITERATURE

Four major components of a CAT include construction of an item bank, an item selection procedure, a latent trait estimation procedure, and a stopping rule (Kingsbury & Zara, 1989, 1991; Dodd, De Ayala, & Koch, 1995; Reckase, 1989; Wainer et al., 1990; Weiss, 1982). In addition, selecting an appropriate IRT model is a critical element. This chapter starts with the introduction of IRT models in unidimensional and multidimensional cases. Then item selection methods are extensively reviewed in the chapter. The rest remaining components of CAT procedures in multidimensional context and polytomous cases are discussed afterwards.

Item Response Theory Models

Selecting an appropriate IRT model for the data is an important precondition for operational procedure of CAT administration. This study discusses the mixed test format that consists of both dichotomous and polytomous items and the uniform test format containing only polytomous items. Most studies in MCAT so far have selected M3PL model for items with two score categories in high-stakes testing, or M2PL that is the special case of M3PL model (Yao & Schwarz, 2006; Wang, Chang, & Boughton, 2011; Wang & Chang, 2011; Segall, 1996, 2001; Mulder & van der Linden, 2009; van der Linden, 1999; etc.). Very few studies investigated polytomous items in MIRT or MCAT research. Yao and Schwarz (2006) used multidimensional generalized partial credit model (MGPCM) in the non-adaptive testing context. Thus, the discussion of polytomous items in MCAT starts from the review of polytomous items in UCAT context.

Examples of polytomous IRT models include the partial credit model (PCM; Master, 1982), the generalized partial credit model (GPCM; Muraki, 1992), the graded response model
(GRM; Samejima, 1969, 1972), the Muraki's rating scale model (MRSM; Muraki, 1990), the normal response model (NRM; Bock, 1972), the Andrich rating scale model (ARSM; Andrich, 1978), the multiple-choice model (MC; Thissen & Steinberg, 1984), the Model 6 (M6; Sympson, 1983), and the success intervals model (SIM; Rost, 1988) (Thissen & Steinberg, 1986; Dodd et al., 1995).

A variety of factors determine the selection of polytomous IRT models: the type of data, model data fit, philosophical considerations, model assumptions, and parsimony (Dodd et al., 1995). Studies have compared different polytomous UIRT models in various settings (Dodd, Koch, & De Ayala, 1988; De Ayala, 1989, 1992; De Ayala, Dodd, & Koch, 1992; Koch, Dodd, & Fitzpatrick, 1990; Jodoin, 2003).

According to Dodd, De Ayala, and Koch (1995), the GRM, GPCM, PCM could be adopted for data ordered to represent varying degrees of the latent trait measure. The PCM model could be used for items in mathematics, physics, and chemistry because points are awarded for the completion of steps leading to the correct answer (Dodd et al., 1995). The GPCM is a similar version of the PCM model with the exception of having varying slope parameters in the GPCM. One of the most commonly explored models is the GPCM in unidimensional and multidimensional cases toward educational testing (van Rijn et al., 2002; Yao & Schwartz, 2006). Thus, when the main interest is in high-stakes educational testing (e.g. K-12 system), the MGPCM model is chosen for polytomous items in this study.

MIRT is a generalization of unidimensional item response (UIRT), when an examinee in the former has a vector of latent abilities and where an examinee in the latter has single latent ability. A brief description of the unidimensional three-parameter-logistic model (U3PL) and unidimensional generalized partial credit model (UGPCM) are presented before the introduction
of their multidimensional counterparts.

The following notation will be used in unidimensional and multidimensional IRT models and to describe the testing process.

UIRT includes a set of mathematic models with the following general representation:

$$P(U = u \mid \theta, \eta) = f(\theta, \eta, u),$$

(2.1)

where $\eta$ is a vector of parameters describing the characteristics of the test items and could include discrimination parameter, difficulty parameter, and a guessing parameter; $f$ is a function showing the relationship between the parameters and the probability of the response (Reckase, 2009); and $u$ is the possible value of the score on the test item. The general relationship of MIRT has the same expression, but $\theta$ and discrimination parameters will be vectors instead of scalars.

The U3PL (Lord, 1980) model is a dichotomous UIRT model that models test items with two score categories. The model is

$$P(U_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}},$$

(2.2)

where $\theta_j$ is the person parameter for examinee j, $a_i$ is the discrimination parameter for item i, $b_i$ is the difficulty parameter for item i, and $c_i$ is the lower asymptote parameter for item i. In UIRT model, these parameters are all scalars.

Multidimensional three-parameter logistic model (M3PL) is an extension of the U3PL model. The mathematical formula is

$$P(U_{ij} = 1 \mid \theta_j, a_i, c_i, d_i) = c_i + (1 - c_i) \frac{\exp(a_i^T \theta_j + d_i)}{1 + \exp(a_i^T \theta_j + d_i)},$$

(2.3)

where $d_i$ is the intercept parameter for item i, and all other parameters have the same definitions as in (2.2). The exception is that $\theta_j$ and $a_i$ are vectors in MIRT models.
The unidimensional generalized partial credit model (UGPCM; Muraki, 1992) is as follows:

\[
P(U_{ij} = u \mid \theta_j) = \frac{\exp\left[\sum_{v=0}^{m_i} Da_i(\theta_j - b_{iv})\right]}{\sum_{j=0}^{m_i} \exp(\sum_{v=0}^{m_i} Da_i(\theta_j - b_{iv}))},
\]

where \( u = 0, 1, \ldots, m_i \), the response of examinee \( j \) to item \( i \), \( a_i \) is the slope or discrimination parameter for item \( i \), \( b_{iv} \) is the \( v \)th item category parameter for item \( i \), \( u = 0, 1, \ldots, m_i \) and \( b_{i0} \equiv 0 \). \( m_i \) is the highest score for item \( i \).

The multidimensional generalized partial credit model (MGPCM) has the following formula

\[
P(U_{ij} = u \mid \theta_j) = \frac{\exp(u a_i^T \theta_j - \sum_{l=0}^{u} \beta_{il})}{\sum_{v=0}^{m_i} \exp(va_i^T \theta_j - \sum_{l=0}^{u} \beta_{il})},
\]

where \( u = 0, 1, \ldots, m_i \), the score given to examinee \( j \) on the item \( i \), and \( m_i \) is the highest score for item \( i \). \( \beta_{il} \) is the threshold parameter for score category \( u \), \( \beta_{i0} = 0 \), \( \theta \) and \( a \) are vectors of latent trait and discrimination respectively in MIRT models.

The MGPCM is generalized from the GPCM in UIRT with two differences. First, unlike its unidimensional counterpart, MGPCM does not include separate difficulty and threshold parameters. Second, since \( \theta \) is a vector while \( \beta \)s are scalars, it is impossible to subtract the threshold parameter from theta vector (Reckase, 2009).

MIRT models generally could be categorized into two types defined by how the information from a vector of theta values is combined with item characteristics to specify the probability of response to the item (Reckase, 2009). One type is compensatory MIRT models
because the final sum \( \theta \) is based on a linear combination of sub-\( \theta \) values (Reckase, 2009). For one sum \( \theta \), there could be a variety of sub-\( \theta \) combinations. For two-dimensional case, if one sub-\( \theta \) on a certain \( \theta \)-coordinate is low, then if the other sub-\( \theta \) on the other \( \theta \)-coordinate is sufficiently high, the sum remains the same. The other type is non-compensatory models because they divide the cognitive tasks of a test item into parts and apply a unidimensional model for each part (Reckase, 2009). Currently, the MIRT models for polytomous items are all compensatory models (Reckase, 2009).

**Item Selection Methods**

One core component of adaptive testing is the item selection procedure that chooses items as a test progresses. This section starts to provide general pictures of item selection methods in UCAT and MCAT. The following section gives detailed descriptions of each type of item selection criterion. Discussions and comparisons of these methods are given afterwards.

Numerous studies have explored different item selection procedures. There are several approaches to classify item selection methods. Van der Linden and Pashley (2010) categorize the methods into classical and modern item selection criteria, where the Fisher expected information and the approximate Bayesian approach are grouped into the classical category and others are in the modern category. Another way to classify methods is as maximizing the information about the location of an examinee on the \( \theta \) values or through minimizing the error in the estimation of the location of an examinee on \( \theta \) values (Reckase, 2009).

Under the UCAT circumstances, the Fisher-information (FI) or maximum information (Lord, 1972, 1980) is the most commonly used approach, and it selects the items with the maximum Fisher information at the provisional ability level to be the next item. An extensively-explored alternative approach is a global information proposed by Chang and Ying (1996), the
Kullback-Leibler (KL) information. The KL information provides information when the estimator is not close to the true ability and the procedure is based on average global information, KL index (KI). The item selection methods taking a Bayesian approach adopts a prior or posterior distribution of ability combining with a Bayesian variant of information as the benchmark for item selection (van der Linden, 1998; van der Linden & Pashley, 2010). The concept of entropy that measures the uncertainty of the distribution of a random variable was introduced and employed in CD-CAT study by Xu, Chang, and Douglas (2005) and Cheng (2009). Wang and Chang (2011) also discussed the entropy-based method in MCAT. Note that the application of Robbins-Monron process into adaptive testing (Lord, 1970) is an important precursor of item selection method, but it is not discussed in this review. Detailed discussions could be found in Lord (1970) and Chang (2004).

Under MCAT circumstances, the general picture of item selection methods is similar to their UCAT counterparts with several enrichments. In MCAT, most recent investigations combine the Bayesian concept into KL information procedure. Examples include KL information with Bayesian update (KLB) (Wang & Chang, 2010), maximum KL distance between two subsequent posteriors ($K^p$) (Mulder & van der Linden, 2010), and mutual information (MI) that maximizes the mutual information between current posterior and predictive response distributions on the candidate items (Mulder & van der Linden, 2010).

Note that the discussion so far is mainly focused on dichotomously-scored items, both in UCAT and MCAT context. Although a few studies have explored polytomous MIRT (PMIRT) under non-adaptive testing (e.g. Muraki & Carlson, 1995; Yao & Schwarz, 2006), very few studies discuss the polytomous item selection in MCAT circumstances. Note that almost all the MCAT studies discussed in the literature apply to dichotomous MIRT models, most of which use
the multidimensional three-parameter logistic (M3PLM) and/or multidimensional two-parameter logistic MIRT model (M2PL). Thus, the literature with respect to polytomous item selection in UCAT is presented below to provide suggestions for polytomus MCAT item selection studies.

Item selection methods in polytomous UCAT (PUCAT) are mostly similar to dichotomous UCAT (DUCAT) with several differences. First, one item selection procedure in polytomous UCAT is called the closest-scale method, and contains a scale value of the item parameter for each item representing the location of the item along the theta continuum. The method selects the item with the closest scale value to the intermediate theta estimate (Dodd et al., 1995). The closest-scale method was studied for the ARSM by Dodd and De Ayala (1994) and the SIM by Kock and Dodd (1996). Its performance does not have significant differences from the performance of the FI procedure (Dodd & De Ayala, 1994; Kock & Dodd, 1996) and most polytomous studies used information-based item selection method (Dodd et al., 1995). Second, the dichotomous information function is defined at the item level whereas the polytomous information function may be defined at the response category or at the item level (Dodd et al., 1995). However, not many studies used a category-level information selection method. One exception by De Anala (1992) found that one less item on average is administered for NRM model when applying category-level information.

The following section is to present major item selection methods studied in the last several decades in details.

**Fisher Information**

The maximum Fisher information was proposed by Lord (1970) with the original motivation of matching items with the examinees' ability level $\theta$ so that the ability estimation process becomes the most efficient. Based on Lord (1980),
where \( P_i(\theta) \) is the probability that an examinee with true ability \( \theta \) will answer the item correctly assuming an IRT model, and \( Q_i(\theta) = 1 - P_i(\theta) \) is the probability examinee \( j \) will answer the item incorrectly.

One nice feature of Fisher information is that the contribution of each item to total test information is additive under the local independence. In another words, for a test that an examinee with ability \( \theta \) takes consisting items \( i=1, 2..., n \), the test information is the sum of individual item information.

\[
I(\theta) = \sum_{i=1}^{n} I_i(\theta). 
\]

(2.7)

This is a highly desirable advantage in CAT because test developers could separately calculate each item's information and then combine them to update test information at each stage (Chang, 2004).

In UIRT, Fisher information has connections with the maximum likelihood estimation of \( \theta \). An estimated theta based on MLE is asymptotically consistent and normally distributed around the true but unknown trait, \( \theta \). The variance of the maximum likelihood estimate, \( \theta_{Hat} \), about \( \theta \), is

\[
Var(\theta_{Hat} \mid \theta) = \frac{1}{E\left[\left(\frac{\partial \ln(L)}{\partial \theta}\right)^2\right]}. 
\]

Another expression of the Fisher information is \( I(\theta) = E\{\left[\frac{\partial}{\partial \theta} \ln(L(\theta; X_1, X_2, ... X_n))\right]^2 \mid \theta\} \)
so that the Fisher information is the reciprocal of the asymptotic variance of the ability estimator. As the number of items becomes large, the mean of the sampling distribution of estimates approaches to the true theta so that \( \theta_{\text{MLE}}^\text{Hat} \sim N(\theta, I^{-1}(\theta_{\text{MLE}}^\text{Hat})) \), and the information that the MLE about theta is

\[
I(\theta, \theta^\text{Hat}) = \sum_{i=1}^{n} \frac{\left( \frac{\partial P_i}{\partial \theta} \right)^2}{P_i(1 - P_i)}.
\]

Therefore, controlling the elements of the information measure controls the sampling distribution of the estimator (van der Linden, 1999). Note that this property does not hold during the early stages of a CAT where just a few items have been administered. For detailed discussions, see the section on attenuation paradox issue later in this document.

For the U3PL model, the Fisher information at the item level could be written in a closed form (Chang & Ying, 1999),

\[
I(\theta) = \frac{(1 - c)a^2 \exp[a(\theta^\text{Hat} - b)]}{1 + \exp[a(\theta^\text{Hat} - b)]} \left\{ 1 - c + c\{1 + \exp[a(\theta^\text{Hat} - b)]\} \right\}.
\] (2.8)

For the UGPCM model, the Fisher information at the item level is:

\[
I_i(\theta) = a_i^2 \left[ \sum_{k=0}^{m} k^2 P_{i,k}(\theta) - \left( \sum_{k=0}^{m} kP_{i,k}(\theta) \right)^2 \right] = a_i^2 \var(U_i).
\]

(Donoghue, 1994; van Rijin, Eggen, Hemker, & Sanders, 2002).

The test information in polytomous CAT is also the sum of individual polytomous item information (Samejima, 1969; van Rijin, Eggen, Hemker, & Sanders, 2002),

\[
I(\theta) = -E \left( \frac{\partial^2 \ln L}{\partial \theta^2} \right) = \sum_{i=1}^{n} \left[ -E \left( \frac{\partial^2 \ln P_{i}(\theta)}{\partial \theta^2} \right) \right] = \sum_{i=1}^{n} I_i(\theta).
\] (2.9)

In MIRT, the Fisher information criterion is a direct generalization from the Fisher
information in UIRT (Reckase & Mckinley, 1991). Due to the asymptotic multinormality, a confidence interval around $\hat{\theta}_{MLE}^{Hat}$ in multivariate analysis becomes an ellipsoid whose volume is proportional to the determinant of the inverse of the fisher information matrix at the true ability ($I^{-1}(\theta)$) as sample size goes to infinity (Anderson, 1984). Utilizing this conclusion in MCAT, Segall (1996) proposed to maximize the determinant of the Fisher information matrix and claimed that this method could maximize the decrement in the volume of the confidence ellipsoid around $\hat{\theta}_{MLE}^{Hat}$.

The additive feature of the Fisher information remains in the MIRT context so that the sum of the item information matrix of individual items forms the test information (Reckase & Mckinley, 1991).

Generally speaking, the Fisher information at the item level is

$$I_i(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta} \log(L(\theta; u_1, \ldots, u_n)) | \theta \right],$$

where

$$L(\theta; u_1, \ldots, u_n) = \prod_{i=1}^n \left[ P_{ji}(\theta)Q_{i1}^{-u_i}(\theta) \right].$$

For the M3PL model, the information matrix is

$$I_i(\theta) = \frac{Q_{ji}(\theta)[P_{ji}(\theta) - c_j]^2}{P_{ji}(\theta)(1 - c_j)} \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \cdots & a_{i2}a_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \cdots & a_{ip}^2 \end{bmatrix}, \quad (2.10)$$

where $p$ is the number of dimensions.

For the MGPCM model, Yao and Schwarz (2006) derived the Fisher information matrix:
\[ I_i(\Theta) = \left( \sum_{u=0}^{m} u^2 P_{iu}(\Theta) - E_i^2 \right) \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \cdots & a_{i2}a_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \cdots & a_{ip}^2 \end{bmatrix}, \] 

where

\[ E_i = \sum_{u=0}^{m} \frac{u \exp(ua_i^T \Theta - \sum_{t=0}^{u} \beta_{it})}{\sum_{u=0}^{m} \exp(ua_i^T \Theta - \sum_{t=0}^{u} \beta_{it})}. \]

In MIRT, the Fisher information is a matrix instead of a scalar. However, during the on-going CAT process, a scalar representing the Fisher information for each candidate item is needed to determine the maximum value to select items. Therefore, a certain statistical procedure is needed to summarize the Fisher information matrix into a scalar. Current studies have explored optimality-based approach such as A- and D-optimality that will be discussed in details in a later section.

Although the FI item selection method is popular in practice, several issues need to be addressed. Veerkamp and Berger (1997) pointed out several problems with the FI, both of which are due to the fact that FI uses the maximum likelihood estimate (MLE) of the ability, \( \hat{\Theta}_{MLE} \). One criticism toward maximum information is that as a point estimation it does not take into account the uncertainty in the estimation at each step. Another problem is that likelihood function does not have a finite maximum when no items are answered correctly or all items are answered correctly. An arbitrary extreme value on the ability scale is used in this case. One potential solution is to use EAP as the ability estimation procedure is (Veerkamp & Berger, 1997). Using an alternative item selection criteria, that does not use ability estimates at each step of a CAT, might avoid these two problems (Veerkamp & Berger, 1997). In addition, two other major issues
for FI are the attenuation paradox issue and the multi-peak issue and these are discussed in the next sections.

**Attenuation paradox issue of fisher information.**

The attenuation paradox issue is that FI tends to show unstable performance at the early stage of CAT (Lord & Novick, 1968). One assumption of the FI-related approach is that the intermediate ability estimates are close to true ability, and this is often violated with only a few items being administered at the beginning of CAT (e.g. Wang & Chang, 2011). Hence, when using FI information at the early stage of CAT, FI tends to have the optimal properties, such as the largest information, at a highly biased estimated theta. Furthermore, since items with larger discrimination parameters are more informative with respect to the ability parameter both in UIRT and MIRT circumstances (Lima Passos, Berger, & Tan, 2007; Mulder & van der Linden, 2010), high-discriminating items tend to be selected by the FI criterion. Thus, at the initial stage of CAT when ability estimates have relatively low precision, selecting high-discriminating items to match biased theta estimators actually provides little information toward latent trait estimation. This causes early stage CAT instability. Correspondingly, the sequential convergence of the ability estimator to the true ability value would be delayed considerably (Lima Passos, Berger, & Tan, 2007), and this lowers the efficiency of estimating examinees' ability in CAT. When the test length is relatively short, the accuracy of ability estimation is negatively affected in that the test might terminate at pre-determined test length while the ability estimator has not converged. Another side-effect of FI is that the item exposure rate becomes uneven. High-discriminating items become overexposed while low-discriminating items tend to be underexposed. These issues have been pointed out and solutions investigated in a number of studies (Chang & Ying, 1996; Chang & Ying, 1999; Chang & van der Linden, 2003; Lima Passos et al., 2007; Segall,
This attenuation paradox also happens in polytomous UIRT (e.g. Lima Passos et al., 2007) and in dichotomous MCAT (Wang, Chang, & Boughton, 2011; Mulder & van der Linden, 2009). No published studies investigating the attenuation paradox issue in polytomous MCAT circumstances to date have been reported.

One effective non-statistical solution is a multistage a-stratified item selection approach for UCAT proposed by Chang and Ying (1999). The criterion successfully balanced item exposure rate by controlling the item exposure rate of high-discriminating items and improves the utilization of low-discriminating items. A series of studies extended a-stratified approach to various test settings (e.g. Chang, Qian, & Ying, 2001; Cheng, Chang, Douglas, & Guo, 2009). Chang & Ying (2008) also showed why selecting high-discriminating items at the early CAT stage has negative impact on estimating latent traits from a theoretical perspective.

In addition to the a-stratified approach, several alternative item selection criteria have been proposed that weaken the FI assumption and alleviate the attenuation paradox issue. These criteria include KL information as the global information (Chang & Ying, 1996), Bayesian-based approaches (Berger & Veerkamp, 1996; van der Linden, 1998), FI-based maximum interval information (MII) or maximum likelihood weighted information (MLWI; Veerkamp & Berger, 1997), A- and D-optimality in polytomous IRT models (Lima Passos, Berger, & Tan, 2007a, 2007b), mutual information (Mulder & van der Linden, 2010), entropy-based method (Wang & Chang, 2011). All of these are discussed in detail in the following sections. The basic idea behind several of these item selection criteria, such as KL information, MII and MLWI, is to use the interval information around the provisional theta estimate instead of using the point information at the provisional estimate \( \Theta^{1st} \) (Lima Passos, Berger, & Tan, 2007). Lima Passos, Berger, and Tan (2007) then proposed A-optimality and D-optimality for polytomous UIRT models, using the
idea of interval information. Notice that these two optimality approaches in polytomous UIRT are different from those in the dichotomous MCAT, that are discussed in details below. In addition, Choi and Swartz (2009) compared several item selection methods for polytomous items and suggested that using interval information through summarizing information for a range of theta values, instead of point information at one estimated theta, might be an effective way to improve the performance in terms of item selection.

**Multi-peak issue of fisher information.**

The other FI-related issue is the multi-peak problem of the information functions. For dichotomous items in UCAT, multiple maximum values of the likelihood equations were found by Samejima (1973) under the U3PL model. Lord (1980) found that multiple solutions did not exist when test length was larger than 20. However, when the test length is short, this could be a problem.

For polytomous items, Muraki (1993) showed that item information functions of the GPCM are not necessarily single peaked and the maximum number of peaks is the number of item categories. Akkermans and Muraki (1997) pointed out that when the difference between the second and first item category parameters of three-category GPCM items is larger than a certain level (i.e. $a_1(b_{i2} - b_{i1}) \geq 4 \ln 2$), then the item information function (IIF) is bimodal. It is possible to select nonoptimal items if using a single point of the information and this could lead to inaccurate estimates of theta and instability of CAT (van Rijin et al., 2002). This issue could also be a problem for polytomous MCAT (PMCAT). One proposal to remedy the multi-peak issue is the use of interval information as is done by MII and MLWI (Berger & Veerkamp, 1997).

MII selects the item with the highest mean value of the information function in a confidence interval of the theta,
where the mean value is obtained by dividing the above formula by the length of the confidence interval.

MLWI uses the likelihood function as a weighted function and defined as the area under a function that is a product of the likelihood function and the information function (Veerkamp & Berger, 1997),

\[
\max_{i \in I_n} \int_{\theta} L_n(\theta; x_n) I_i(\theta) d\theta . \tag{2.13}
\]

In contrast, FI is classified as the point information criterion with ML estimation (Veerkamp & Berger, 1997).

Veerkam and Berger (1997) found that MLWI and FI with EAP estimation are good alternatives to FI under the dichotomous U3PL model. Van Rijn, Eggen, Hemker, and Sanders (2002) compared FI and MII for polytomous items by using the UGPCM model. The study did find negligible differences between the two methods, but MII sometimes performed worse than FI. This finding is similar to what Berger and Veerkamp (1997) found for dichotomous items. One possible reason is that MII also relies largely on the \( a_i \), the discrimination parameter. Thus, MII selects similar items as FI, especially those items with high \( a_i \) parameters. As a result, MII does not differ from FI very much in terms of accuracy and precision of the ability estimate. Van Rijn et al. (2002) suggested, but did not investigated, that one approach of overcoming multi-peak issue for polytomous items might be KL because it depends less on the discrimination parameter \( a_i \). Another example is that Choi and Swartz (2009) who used the unidimensional graded response model (UGRM) model and investigated various item selection criteria: FI, MLWI, and some Bayesian-based selection methods, that are discussed in a later section. Choi and Swartz
(2009) found that FI combined with EAP estimation is one of the best item selection method even though it is simple. The MLWI, however, had poor performance under certain conditions, such as for lower-ability examinees whose theta was in range of (-2, 0).

Both the attenuation paradox and the multi-peak issues seem to be caused or partially caused by problems of relying on point or local information and being affected by items' discrimination parameters. No studies have discussed the multi-peak issue for MCAT using polytomous items. This could be a research direction in the future. In MIRT, one feature similar to UIRT is that items with larger discrimination parameters are more informative with respect to the ability parameter (Mulder & van der Linden, 2010; Reckase & Mckinley, 1991). Determines whether such a characteristic exists in polytomous MCAT environment requires more studies.

**Optimality-based fisher information item selection methods.**

Since the optimal design studies have been applied to educational testing (Berger & Wong, 2005), various optimality criteria are also investigated in the adaptive testing (e.g. Mulder & van der Linden, 2009). Note that the studies of optimality-based item selection criteria using polytomous UIRT model take a different approach than those in MCAT. Optimality-based approach is to summarize the information matrix in MCAT while using the concept of interval information in UCAT. Fisher Information in MCAT is a matrix instead of a scalar. Hence, in MCAT, certain statistical concepts and statistical approaches have to be used to summarize the information matrix so that the item selection could be based on a scalar value. In contrast, the Fisher Information in UIRT models is a scalar instead of a matrix. Therefore, the optimality-oriented approaches for polytomous UIRT models are alternative procedures to the FI point information with the idea of using the interval information (Lima Passos et al., 2007).
Optimality-based fisher information in UCAT.

Both A- and D-optimality in polytomous UIRT context are the objective function \( g(\cdot) \) of an information measure (Lima Passos, Berger, & Tan, 2007). The function \( g(\cdot) \) should be maximized along a latent trait interval \( \Theta = [\theta_L, \theta_U] \) with \( \theta_t \in \Theta \), and \( t = 1, 2, \ldots, T \), \( \theta_L \) (t=1) and \( \theta_U \) (t=T) are lower and upper boundaries of the latent trait interval \( \Theta \). Berger and Veerkamp (1996) summarized that A-optimality corresponds to the arithmetic mean and D-optimality is to the geometric mean.

The A-optimality adopted by Lima Passos, Berger, and Tan (2007) in polytomous CAT is

\[
g_A(FI_i) = \sum_{t=1}^{T} FI_i . \tag{2.14}
\]

A-optimality can be considered as the weighted sum of Fisher Information measure, and the weights are given by the frequency distribution of \( \theta_t \in \Theta \) (Lima Passos, Berger, & Tan, 2007). Geometrically, the A-optimality criterion is a discrete version of the area under the curve of the item information function over ability range. Hence, the A-optimality is equivalent to MII (Veerkamp & Berger, 1997) when the ability range is a confidence interval of the theta estimate.

The adoption of D-optimality for polytomous UIRT models uses the following formula.

\[
g_D(FI_i) = \prod_{t=1}^{T} FI_i . \tag{2.15}
\]

It has the same notation as A-optimality above. The D-optimality is inversely proportional to the volume of estimated theta’s confidence ellipsoid and is invariant under any linear transformation of the independent variable scale. Hence, the D-optimality design remains the same irrespective of the scale used to measure the latent trait variable for test optimal design (Lima Passo, 2007).
Berger and van der Linden (1995) first utilized D-optimality in the optimal test assembly of fixed-form group-based test. Lima Passos, Berger, and Tan (2007a, 2007b) applied D-optimality item selection methods in the Nominal Response model (NRM) for both fixed-form and adaptive tests, respectively. Item selection strategies, the A-optimality and D-optimality criteria and the KL criterion were compared for polytomous items described by the Graded Response Model (GRM) (Lima Passos et al., 2007b) and by the NRM (Lima Passos et al., 2007a). According to Lima Passos, Berger, and Tan (2007a, 2007b), under NRM and GRM and under two different item pool compositions, D-optimality choose a relatively low discriminating item at the outset for both pools but the KL criterion ended up similar at the end although it selected more discriminating items at the beginning. Both D-optimality and KL information have a general robustness against the instability at the initial stage of CAT. The global information, KL is more robust against early stage instability than a local information criterion (Lima Passos, 2007b). To sum up, for certain polytomous UIRT models, KL information and D-optimality do not suffer from an attenuation paradox.

Note that D-optimality is reported to be sensitive toward the changes on underlying trait distribution based on the studies by Berger and van der Linden (1995) and Lima Passos (2005). Therefore, whether such findings generalized to the MCAT case where latent traits become vectors needs further exploration.

**Optimality-based fisher information in MCAT.**

Several optimality criteria based on the fisher information matrix in MIRT, including A-, D-, E-optimality, are discussed and compared here.

A-optimality minimizes the sum of the asymptotic sampling variances of the maximum likelihood estimation (MLE) of the abilities. Equivalently, it minimizes the trace of the inverse of
the information matrix (Mulder & van der Linden, 2009),

$$\arg \min_{i_k \in R_k} \text{trace}(I_{s_{k-1}} + I_{i_k}) = \arg \max_{i_k \in R_k} \frac{\det(I_{s_{k-1}} + I_{i_k})}{\sum_{j=1}^{3} \det(I_{s_{k-1},i_k} + I_{i_k}[f,j])}. \quad (2.16)$$

Using A-optimality selects different items than D-optimality because A-optimality only focuses on the variances of the ability estimators (Mulder & van der Linden, 2009).

The D-optimality algorithm maximizes the determinant of the fisher information matrix (Mulder & van der Linden, 2009),

$$\arg \max_{i_k \in R_k} \det(I_{s_{k-1}}(\hat{\theta}^H_{k-1}) + I_{i_k}(\hat{\theta}^H_{k-1})). \quad (2.17)$$

Under the MCAT case, it is to minimize the confidence ellipsoid of the ability estimate. Thus, it is equivalent to minimizing the generalized variance to yield the smallest confidence region for the ability parameters (Mulder & van der Linden, 2009).

E-optimality maximizes the smallest eigenvalue of the information matrix. Equivalently, it maximizes the generalized variance of the estimated thetas along the largest dimension. The disadvantage of the E-optimality is the lack of robustness in applications with sparse data. E-optimality is unfavorable for MCAT item selection because E-optimality is unstable and showed occasional erratic performance in item selection (Mulder & van der Linden, 2009).

Mulder and van der Linden (2009) also discussed several optimality-based criteria, such as $A_S$-optimality, $D_S$-optimality, $c(\lambda_1)$-optimality, $c(\lambda_2)$-optimality, and detailed discussions could be found in the study by Mulder and van der Linden (2009).

Which criterion would be used depends on the goal of the testing. For example, under the condition of all intentional abilities, A-optimality and D-optimality yield the most accurate estimates (Mulder & van der Linden, 2009).

According to Wang and Chang (2011), a major difference between the item selection in
UCAT and MCAT is caused by the nature of the optimization. The item selection method is a single-objective optimization problem in UCAT but is a multi-objective optimization problem in MCAT. The former measures one latent trait while the latter measures several traits simultaneously (Wang & Chang, 2011). For multi-objective case, two typical ways exists to deal with the optimization issue. One approach is to form a single aggregate objective function and the other is to take a sequential method to meet different objectives consecutively (Wang & Chang, 2011). Note that D-optimality and A-optimality criteria in MCAT are taking the approach of forming a single aggregate objective function to deal with the multi-objective optimization problem in MCAT (Wang & Chang, 2011).

In multidimensional case, there are several studies investigating A- and D-optimality for dichotomous MIRT models. Segall (1996), Luecht (1996), and Mulder and van der Linden (2009) used the D-optimality. van der Linden (1999) studied A-optimality.

In MCAT, there are several advantageous properties of the FI-based D-optimality criterion. therefore, D-optimality is commonly used (Atkinson & Donev, 1992; Berger & Veerkamp, 1996; Passo, 2007). First, through maximizing the determinant of the Fisher information matrix, D-optimality is believed as the most precise based on a Euclidean distance criterion, with the Euclidean distance being the distance between the true and estimated ability parameters (Luecht, 1996; Miller, Reckase, Spray, Luecht, & Davey, 1996). Second, D-optimality is suggested in the educational testing context since D-optimality selects a set of items from a bank with the smallest generalized variance of the ability estimators for a population of examinees. Items selected through D-optimality have the smallest confidence region for the ability parameters (Mulder & van der Linden, 2009). One non-statistical advantage of Fisher information is that it is simple to compute and requires less CPU time when picking the
succeeding items.

**Kullback-Leibler Information**

Generally, KL information is to measure the non-symmetric discrepancy between two probability distributions over the same parameter space (Cover & Thomas, 1991; Lehmann & Casella, 1998). The milestone article written by Chang and Ying (1996) introduced KL information into UCAT.

The KL item information is defined as

$$KL_i (\theta^H; \theta) = E_{\theta} \left[ \log \frac{L_i (\theta; U_i)}{L_i (\theta^H; U_i)} \right],$$

(2.18)

where $L_i (\theta; U_i) = P_i (\theta) Q^{1-U_i} (\theta)$. 

The item KL information for item j can also be expressed for dichotomous items

$$KL_i (\theta^H \| \theta) = P_j (\theta) \log \left[ \frac{P_j (\theta)}{P_j (\theta^H)} \right] + [1 - P_j (\theta)] \log \left[ \frac{1 - P_j (\theta)}{1 - P_j (\theta^H)} \right].$$

(2.19)

Correspondingly, for a polytomous IRT model, the item KL information for item j is

$$KL_i (\theta^H \| \theta) = \sum_{i=0}^{m} P_i (\theta) \log \left[ \frac{P_i (\theta)}{P_i (\theta^H)} \right].$$

(2.20)

The KL test information is defined as

$$KL(n) (\theta^H \| \theta) = E_{\theta} [l_n (\theta) - l_n (\theta^H)] = \sum_{j=1}^{n} KL_j (\theta^H \| \theta).$$

(2.21)

Since KL information is a function instead of a value, the index has to be generated so that KL information approach can be used as an item selection procedure. In UCAT, Chang and Ying (1996) proposed a single index, KL information index (KI) that integrates KL information over an interval including
\[ KI(\theta_n^{\text{Hat}}) = \int_{\theta_n^{\text{Hat}} - \delta_n}^{\theta_n^{\text{Hat}} + \delta_n} KL(\theta_n^{\text{Hat}} \parallel \theta) d\theta, \quad (2.22) \]

where \(\delta_n\) determines the size of the region over which the average is computed. \(\delta_n = \frac{d}{n^{1/2}}\).

The region is large at the early stage of CAT so as to contain the true ability value in the region as much as possible. However, the region becomes smaller as the test progresses since more information has been obtained to pinpoint the true ability. Geometrically, KI is the area under the KL function from \((\theta_n^{\text{Hat}} - \delta_n, \theta_n^{\text{Hat}} + \delta_n)\). The maximum area is equivalent to the maximum curvature and thus the maximum value of FI.

KL information has also been discussed under the MCAT context following MIRT models with the latent trait \(\theta\) in KL becoming a vector instead of a scalar in UIRT. Wang, Chang, and Boughton (2011) explore whether the properties of FI and KL established in UIRT extend to MIRT. One of the properties extended to MIRT is the connection between FI and KL information. According to Chang and Ying (1996), FI at \(\theta\) is the second derivative of KL at the same true latent trait value \(\theta\) in UIRT. Geometrically, FI is the curvature of the KL curve on the plane at \(\theta\) in UIRT. In the multi-dimensional case, Wang, Chang, and Boughton (2011) justified that the FI matrix is the second partial derivative, or the Hessian matrix, of the KL, which is

\[ I_{ij}(\theta) = \frac{\partial^2}{\partial \theta_i^{\text{Hat}} \partial \theta_j^{\text{Hat}}} KL(\theta^{\text{Hat}} \parallel \theta). \quad (2.23) \]

Several KL-based Bayesian methods are proposed recently in MCAT that are discussed in detail in the section on Bayesian item selection methods.

**Fisher Information versus Kullback-Leibler Information**

Fisher information and Kullback-Leibler Information are two main approaches of item selection and there are differences and connections between them. KL information has similar
features to Fisher information. Analogous to the additive feature of FI, KL test information is equal to the sum of all items’ KL item information.

Some features of KL information distinct from those of FI. First, FI is a function of \( \theta \) only and provides the discriminating power for two traits that are close to each other. In contrast, KL is a function of two trait levels, where one trait is the true ability \( \theta \) while the other trait \( \theta^{Hat} \) could vary over the whole range of theta. Furthermore, it does not require that \( \theta^{Hat} \) should be close to \( \theta \). Chang and Ying (1996) argued that FI represents a local information function while KL information serves as the global information. KL criterion provides information when the estimator is not close to the true ability and the item selection procedure is based on average global information. FI, however, quantifies the discrimination power around \( \theta \) (Chang & Ying, 1996; Hambleton & Swaminathan, 1985). Local information should be used when the test length \( n \) is large while global information used when \( n \) is small. KL information does not suffer from the attenuation paradox and instability issue. KL leads to more stable, efficient, and precise ability estimates, especially at early CAT or CAT with short test length. Veldkamp and van der Linden (2006) showed that KL-based item selection approach performs better than FI-based item selection method. Second, statistically, the KL information could be viewed as the likelihood ratio test since \( L_i(\theta;U_i) \) is the likelihood function for true theta \( \theta \) and \( L_i(\theta^{Hat};U_i) \) is the likelihood function for ability estimator. KL is the expectation of the log-likelihood ratio. It is known that the likelihood ratio test is the best test to distinguish \( \theta \) from \( \theta^{Hat} \) according to the Neyman-Pearson theory (Lehmann, 1986). Third, KL is a function of two levels, \( \theta^{Hat} \) and \( \theta \) while Fisher information is a fixed number (Chang & Ying, 1996). Thus, KL could be referred to as the relative entropy when using \( L_i(\theta;U_i) \) instead of using \( L_i(\theta;U_i) \). Fourth, one appealing feature of KL information is that no matter how many dimensions there are, the KL information
is always a scalar. Hence, KL immediately generalizes from a unidimensional test to a multidimensional test. Fisher information, however, is a matrix in MCAT and has to be further summarized by using certain mathematical approaches.

There are connections between FI and KL information. In UCAT case, Chang and Ying (1996) showed that FI at the true ability $\theta$ is the second derivative of KL at $\theta$,

$$\frac{\partial^2}{\partial(\theta_{Hat})^2} KL(\theta_{Hat} \parallel \theta) \big|_{\theta} = I(\theta).$$ (2.24)

KL quantifies how powerful or efficient the statistical test is and represents the discrimination power of the item when distinguishing $\theta$ from $\theta_{Hat}$. When $\theta_{Hat}$ is approaching $\theta$ or varying around $\theta$, the KL equals to the FI. Geometrically, when KL is a curve on the plane, FI is the curvature of the KL curve at the point of true theta $\theta$ (Chang & Ying, 1996; Wang, Chang, & Boughton, 2011). For DUCAT, maximizing KL is eventually equivalent to maximizing FI and simulation showed KL was as well as or better than FI (Chang & Ying, 1996). Wang, Chang and Boughton (2011) extended this connection between the item KL index (KI) and the FI matrix under a two-dimensional MIRT model adopting dichotomous M3PL model. FI matrix is shown to be the same as the second partial derivative matrix or Hessian matrix of the KL, with the mathematical expression as follows:

$$I_{ij}(\theta) = \frac{\partial^2}{\partial \theta_i^{Hat} \partial \theta_j^{Hat}} KL(\theta_{Hat} \parallel \theta).$$ (2.25)

The geometric relationship aforementioned was also presented in Wang, Chang and Boughton (2011). Wang and Chang (2011) further extend the above conclusion to a p-dimensional case, where $p>2$.

Wang, Chang, and Boughton (2011) showed that for two-dimensional M2PL model, when test length $L$ goes to infinity, the magnitude of KL index (KI) is proportional to the trace of
the Fisher information matrix and also proportional to the square of the item multidimensional discrimination (MDISC, Reckase, & McKinley, 1991),

\[
MDISC_i = \sqrt{\sum_{j=1}^{p} a_{ij}^2}.
\]  

(2.26)

Wang and Chang further extend this relationship to p-dimension and where \( p > 2 \).

\[
KI \approx \frac{2^{p-1}}{3} \left( \frac{r}{\sqrt{n}} \right)^{p+2} \frac{Q_i(\theta)[P_i(\theta) - c_i]^2}{P_i(\theta)(1 - c_i)} \sum_i a_{ij}^2.
\]  

(2.27)

The relationship shows that items highly discriminated on multiple dimensions are favored by KI.

In UCAT, Chang and Ying (1996) showed that maximizing KI eventually is equivalent to maximizing the Fisher information. In MCAT when the number of dimensions \( p \geq 2 \), however, this relationship does not exist (Wang, Chang, & Boughton, 2011; Wang & Chang, 2011). D-optimality maximizes the determinant of the Fisher information matrix. KI maximizes the trace or the summation of the eigenvalues of the Fisher information matrix. Hence, KI will never be equivalent to D-optimality even when the test length \( L \) is large.

D-optimality, however, favors items with high single discrimination parameters throughout the test (Mulder & van der Linden, 2009). Thus, when test length \( L \) is small, the ability estimation error will be large. KI, however, is more robust with shorter test because KL information is integrated over the whole range of theta levels so that it circumvent choosing high-discriminating items at low-precision ability estimates at the beginning of the test.

**Bayesian Item Selection Methods**

Owen (1975) introduced the Bayesian approach into the adaptive testing, and his approach was named approximate Bayesian criterion or the restricted Bayesian updating (van der Linden, 1998; Mulder & van der Linden, 2009). Owen (1975) used sequential Bayesian procedures so that the previous posterior distribution is used as the new prior distribution of the
unknown parameter. Note that the true posterior distribution is not a normal distribution since the likelihood function does not have a normal family as the class of conjugate distributions while the prior distribution is normal (van der Linden & Pashley, 2010). Owen (1975) proposed to assume a normal prior distribution and use a normal approximation to replace the true posterior by using the same mean and variance of the true distribution to avoid unresolvable computational difficulties at that time (van der Linden, 1998). This approach was later proved statistically or theoretically acceptable because Chang and Stout (1993) showed that the posterior distribution is asymptotically normal with a mean equal to theta under mild nonparametric assumptions.

In Owen's (1975) approach, the kth item is selected such that

\[ |b_i - E(\theta | u_i, \ldots, u_{i-1})| < \delta, \]

where \( \delta \) is a small value and \( \delta \leq 0 \), \( E(\theta | u_i, \ldots, u_{i-1}) \) is the Expected A Posterior (EAP) estimator, \( b_i \) is the item difficulty parameter. After administering the kth item, the likelihood is updated and combined with the previous posterior to calculate a new posterior and then a new item is selected (Owen, 1975).

Due to the rapid development of technology, the numerical complexity is not a problem and approaches of using full posterior were proposed into the item selection procedures. Van der Linden (1998) proposed several Bayesian-based item selection criteria that are classified into fully Bayesian approach or one of modern criteria by van der Linden and Pashley (2010).

The maximum posterior weighted information criterion (MPWI) generalizes maximum information concept in a Bayesian way. The idea is to choose the appropriate information measure and take the expectation across the posterior distribution (van der Linden, 1998). This criterion uses the posterior distribution to weight the information function and puts more weight.
on items with information around the location of the posterior distribution. Let

\[ J_{u_{i_1},...,u_{i_k}}(\theta) = I_{U_{i_1},...,U_{i_k}}(\theta), \]

and the next item selected maximizes the expected value of the observed information \( J(\theta) \) over the posterior distribution of \( \theta \) that is

\[ i_k = \text{max}_j \{ \int J_{U_j}(\theta)g(\theta | u_{i_1},...,u_{i_{k-1}})d\theta; j \in R_k \}, \]

where \( g(\theta | u_{i_1},...,u_{i_{k-1}}) \) is the posterior update as follows:

\[ g(\theta | u_{i_1},...,u_{i_{k-1}}) = \frac{L(\theta | u_{i_1},...,u_{i_{k-1}})g(\theta)}{\int L(\theta | u_{i_1},...,u_{i_{k-1}})g(\theta)d\theta}. \quad (2.28) \]

It is also possible to combine the criterion with KL measure (van der Linden & Pashley, 2010).

According to van der Linden (1998), in general, the likelihood weighted information (Veerkamp & Berger, 1997) is a superior method to the interval information (Veerkamp & Berger, 1997). However, for the first few items, the likelihood function is still flat and high-discriminating items might be over used again. One solution to improve the likelihood weighted information is to use posterior distribution of the ability parameter to weigh the information.

The following three criteria, the maximum expected information (MEI), the minimum expected posterior variance (MEPV), and the maximum expected posterior weighted information (MEPWI), are preposterior analysis. After \( k-1 \) items are administered, the response distributions on the remaining items in the item pool \( i \in R_k \) are predicted first, then the next item is chosen based on the update of a posterior quantity of these distribution. The predictive posterior distribution for the response on item \( i \) has the probability function as follows,

\[ p_j(U_j = u_j | u_{i_1},...,u_{i_{k-1}}) = \int p_j(U_j = u_j | \theta)g(\theta | u_{i_1},...,u_{i_{k-1}})d\theta. \quad (2.29) \]

The maximum expected information criterion (MEI) is to maximize observed information over the predicted responses on the \( k \)th item.
\[
  i_k = \max_j \left\{ p_j(U_j = 0 \mid u_i, \ldots, u_{i+k-1}) I_{u_i, \ldots, u_{i+k-1}, U_j = 0} \left( \hat{\theta}_{u_i, \ldots, u_{i+k-1}, U_j = 0} \right) + p_j(U_j = 1 \mid u_i, \ldots, u_{i+k-1}) I_{u_i, \ldots, u_{i+k-1}, U_j = 1} \left( \hat{\theta}_{u_i, \ldots, u_{i+k-1}, U_j = 1} \right); j \in R_k \right\}. \tag{2.30}
\]

The minimum expected posterior variance (MEPV) criterion replaces the observed information in MEI by the posterior variance of theta (van der Linden, 1998).

\[
  i_k = \min_j \left\{ p_j(U_j = 0 \mid u_i, \ldots, u_{i+k-1}) \text{Var}(\theta \mid u_i, \ldots, u_{i+k-1}, U_j = 0) + p_j(U_j = 1 \mid u_i, \ldots, u_{i+k-1}) \text{Var}(\theta \mid u_i, \ldots, u_{i+k-1}, U_j = 1); j \in R_k \right\}. \tag{2.31}
\]

The MEPV is preposterior risk related to a quadratic loss function for the estimator, and Owen (1975) used the approximate approach to simplify this numerically complicated version at the time. In addition, this criterion is a small-sample alternative to MEI since the reciprocal of the information measure is only a large-sample approximation to the true variance of the posterior (van der Linden, 1998). Wang and Chang (2011) pointed out that, in MCAT, if the variance in MEPV is replaced by the volume of ellipsoid around the point estimate, this MEPV criterion would become an extended Bayesian version of D-optimality.

The maximum expected posterior weighted-information (MEPW) is to weigh observed information using the posterior distribution of \( \theta \) and then take the expectation over the full predicted posteriors.

\[
  i_k = \arg \max_j \left\{ p_j(U_j = 0 \mid u_i, \ldots, u_{i+k-1}) \cdot \int J_{u_i, \ldots, u_{i+k-1}, U_j = 0}(\theta) g(\theta \mid u_i, \ldots, u_{i+k-1}, U_j = 0) d\theta + p_j(U_j = 1 \mid u_i, \ldots, u_{i+k-1}) \cdot \int J_{u_i, \ldots, u_{i+k-1}, U_j = 1}(\theta) g(\theta \mid u_i, \ldots, u_{i+k-1}, U_j = 1) d\theta; j \in R_k \right\}. \tag{2.32}
\]

Several studies have applied Bayesian-based item selection approaches into CAT adopting polytomous model. Penfield (2006) investigated MEI and MPWI adopting partial credit model in UIRT (UPCM) and claimed that these two methods have similar performance, both of which have superior efficiency of ability estimation compared with the FI approach. In addition, Penfield (2006) also pointed out that these two methods lead to more efficient ability estimation.
than FI. Choi and Swartz (2009) used the UGRM model and investigated various item selection
criteria: FI, MLWI, MPWI, MEI, MEPV, and MEPWI. Choi and Swartz (2009) claimed that the
advantages of Bayesian-based item selection methods in dichotomous UIRT models might be
masked for polytomous UIRT models in practice unless the item bank is large and item
information function covers a narrow range. Their findings also showed that for polytomous
items using the unidimensional GRM (UGRM) model, FI combined with EAP ability estimation
performs very well and other complex and computing-intensive item selection methods are not
competitive.

The Bayesian item selection method has been also applied to MCAT. One Bayesian
approach to item selection in MCAT (Segall, 1996) is to select the next item to maximize the
decrement in the volume of the posterior credibility ellipsoid, equivalently, to maximize the
determinant of the posterior variance-covariance matrix,

$$| \Sigma_k^{-1} | = | I(\theta, \hat{\theta}_k^{\text{hat}}) + I(\theta, \hat{u}_k) + \Phi^{-1} |,$$

where $\Sigma_k^{-1}$ is the covariance matrix of the posterior distribution after administering k items, $\Phi^{-1}$
is the covariance matrix of the prior distribution of abilities. Thus, the maximum likelihood item
selection criterion and the bayesian item selection method proposed by Segall (1996) differs only
by the term of the inverse of the covariance matrix of the prior distribution of abilities $\Phi^{-1}$. One
advantage of this Bayesian approach is to put prior knowledge about the relationships between
the ability variables into the item selection procedure (van der Linden, 1999).

Segall (2001) also proposed a Bayesian adaptive item selection algorithm that minimizes
the expected posterior variance. The posterior distribution is approximated by a multivariate
normal density based on the curvature at the mode, so that the mean equal to the posterior mode
$\theta_k^{\text{hat}}$, and the covariance matrix $\Sigma_{\theta_k^{\text{hat}}}^{-1}$ equals the inverse of the posterior information matrix at
the mode $\hat{\theta}_{k-1}$, that equals $\Sigma_{i \in S_{k-1}} = [\gamma_{i \in S_{k-1}}]^{-1}$. The information matrix is

$$
\gamma_{i \in S_{k-1}} = W_i + I + \sum_{j \in S_{k-1}} W_j,
$$

where $W_i = D^2 a_i a_i' \frac{Q_i(\theta) - P_i(\theta) - c_i}{P_i(\theta) - 1 - c_i}$ for M3PL model (Segall, 2001).

**Posterior expected KL information ($K^B$).**

Veldkamp and van der Linden (2002) proposed a Bayesian version of KL information combined with the shadow test approach adopting M2PL model. The criterion is called posterior expected KL information, denoted as $K^B$,

$$
\text{arg max}_{\theta} K^B_{i \in \Theta} (\theta_{Hat_{k-1}}) = \text{arg max}_{\theta} \int K_{i \in \Theta} (\theta_{Hat_{k-1}}; \theta) f(\theta | u_{k-1}) d\theta,
$$

(2.33)

where $u_{k-1}$ is the vector of item responses of the first $k-1$ administered items, and $\theta_{k-1}$ is the EAP estimate of the ability.

$$
\theta_{Hat_{k-1}} = \int \theta f(\theta | u_{k-1}) d\theta.
$$

This means that the selection of an item that maximally discriminates between the EAP estimate and the other abilities in the multidimensional ability space covered by the current posterior (Mulder & van der Linden, 2010). It is the distance between the response distributions at a current estimate of the ability vector $\theta$, and the true theta vector $\theta$ integrated over the current posterior distribution of the latter. This criterion was shown to outperform the FI counterpart at the beginning of tests and maintains the feature of additivity in the items in the shadow test. As Wang, Chang, and Boughton (2011) pointed out, one positive aspect of this index is that it combines both local and global information so that it makes appropriate usage of information throughout the entire test.
KL distance between subsequent posteriors \((K^P)\).

Mulder and van der Linden (2010) proposed the criterion of the expected KL distance across the response distribution, denoted as \(K^P\):

\[
\arg \max \, K^P_{i_k} [f(\theta | u_{k-1})] = \arg \max \, \sum_{i_k \in R_k} f(u_{i_k} | u_{k-1}) K(f(\theta | u_{k-1}), f(\theta | u_{k-1}, u_{i_k})),
\]

(2.34)

where \(K(f(\theta | u_{k-1}), f(\theta | u_{k-1}, u_{i_k}))\) is the KL distance between the present and updated posterior densities. Item with the largest expected distance between the current and new posterior distributions of \(\theta\) should be selected. Wang and Chang (2010) proposed the KL information with Bayesian update method (KLB), which is similar to the \(K^P\) algorithm but with information gain interpretation. It is the expected distance between the prior distribution of \(\theta\) and its posterior distribution after the administration of the candidate item.

\(K^B\) and \(K^P\) only differs in the definitions of the probabilities of a correct and incorrect response (Mulder & van der Linden, 2010). \(K^B\) uses the posterior distribution to calculate an update of the response probabilities, \(K^P\) takes the uncertainty in the ability estimate into account and is more robust with respect to ability estimation. \(K^P\) is the KL distance between the joint and product distributions of \(\theta\) and the response on the candidate item.

**Mutual Information**

Mutual information (MI) was first applied into adaptive testing by Weissman (2007) and Mulder and van der Linden (2010) generalized it to an MCAT case from a theoretical perspective. Wang and Chang (2011) further compared the MI with other item selection in MCAT context adopting M3PL model both from theoretical perspectives and simulation results.

For two continuous random variables \(X\) and \(Y\), mutual information is defined as
\[ MI(X, Y) = \int \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \, dx \, dy. \] (2.35)

MI is a measure of the amount of information X provides about Y, and also a measure of the amount of information in X about Y (Weissman, 2007; Mulder & van der Linden, 2010).

MI is a special case of KL information since MI can be considered as the KL between the joint distribution \( f(X, Y) \) and the product of marginal distribution \( f(X) \) and \( f(Y) \). When \( X \) and \( Y \) are independent, \( X \) carries no information about \( Y \) and vice versa. Statistically, if \( X \) and \( Y \) are independent, then \( f(X, Y) = f(X)f(Y) \), and the mutual information equals to zero.

There are several properties of mutual information. First, \( MI \geq 0 \), and the expectation is over the entire space of the joint distribution, both \( X \) and \( Y \). In contrast, the KL information is only taking expectation over \( X \). Second, KL information is not symmetric so that \( KL(\hat{\theta} \parallel \theta) \neq KL(\theta \parallel \hat{\theta}) \) except at the point of true theta \( \theta \). MI is symmetric, a desirable feature in sequential testing (Weissman, 2007). In addition, Weissman (2007) also pointed that the only restriction of mutual information is that the joint distribution \( f(X, Y) \) is a valid probability distribution function. Thus, MI exists under conditions of non-continuous distribution for \( f(X, Y) \) distribution, multidimensional latent trait case, both dichotomous and polytomous item responses, and no requirement for the parametric form of \( f(X, Y) \).

In the CAT case, let the current posterior distribution based on \( k-1 \) items administered be the \( Y \) in original formula, and let the predictive response distribution on candidate item conditioning on the previous responses be the \( X \) in original formula. Hence, the mutual information item selection criterion in CAT is to maximize the mutual information between the present posterior and predictive response distribution on the candidate item (Mulder & van der Linden, 2010).
Note that \( f(\theta, u_{i_k} \mid u_{k-1}) = f(u_{i_k} \mid \theta) f(\theta \mid u_{k-1}) \) so that the criterion can be simplified as

\[
\arg\max_{i_k \in R_k} MI(\theta; u_{i_k}) = \arg\max_{i_k \in R_k} \sum_{\mu_{i_k} = 0}^m \int_\theta f(\theta, u_{i_k} \mid u_{k-1}) \log \frac{f(\theta, u_{i_k} \mid u_{k-1})}{f(\theta \mid u_{k-1}) f(u_{i_k} \mid u_{k-1})} d\theta.
\] (2.36)

The item maximizing mutual information between the test taker's current posterior distribution and the response distribution on the candidate item should be selected because these items are the closest to \( \theta \) according to the posterior information in the previous items (Weissman, 2007). It is a symmetric version of the KL distance between the current posterior distribution of \( \theta \) and the response distribution on the candidate item.

Mutual information has an important interpretation related to the concept of conditional entropy. Conditional entropy is expressed as \( H(Y \mid X) = \sum_x p(x)H(Y \mid X = x) \). Hence, it is straightforward that \( H(Y \mid X) \) is smaller than \( H(Y) \), and the discrepancy between these two shows the decrement of the uncertainty through adding the information carried by \( X \). Therefore, the MI between \( X \) and \( Y \) indicates the difference in entropies \( I(X; Y) = H(Y) - H(Y \mid X) \) (Cover & Thomas, 1991; Weissman, 2007; Mulder & van der Linden, 2011). This expression in testing circumstances is presented as follows.

\[
MI(\theta; \mu_{i_k}) = H(\theta) - H(\theta \mid \mu_{i_k}),
\]

where

\[
H(\theta) = -\sum_{u_{i_k} = 0}^m \int_\theta f(\theta, \mu_{i_k}) \log f(\theta) d\theta,
\]

the information in \( \theta \),

and
\[ H(\theta | \mu_{u_k}) = -\sum_{u_{ki} = 0}^{m} f(\theta, \mu_{u_k}) \log f(\theta | \mu_{u_k}) d\theta, \] the information in \( \theta \) based on the observation \( u_{ki} \).

Mutual information is the average KL distance between the new and current posteriors, and \( K^p \) is the average KL distance between the current and new posteriors. These two are not the same since the measure is not symmetric. \( K^p \) is believed as a better item selection criterion than \( K^b \), and MI is more robust with respect to error in the ability estimate than \( K^p \) (Mulder & van der Linden, 2010). These three criteria are specially for the goal of all abilities intentional. Mulder & van der Linden (2010) also explored the modification of \( K^b \), \( K^p \), and MI in the cases of some abilities nuisances or a linear combination of intentional abilities.

**Entropy Method**

The concept of entropy is to measure the uncertainty of the distribution of a random variable so that entropy-based item selection methods select items to directly monitor the entropy of the posterior distribution of \( \theta^{tail} \) (Wang & Chang, 2011). Shannon entropy (Shannon, 1948) is applied to the adaptive testing content and is called continuous entropy or differential entropy when the random variable X follows a continuous distribution.

\[ H(P) = -\int p_i \log(p_i) dp, \]

When \( P(X) \) is most concentrated, \( H(P) = 0 \). (e.g. \( P(X = x_j) = 1 \) for a certain \( j \) but \( P(X = x_i) = 0 \) for all \( i \neq j \)); when \( P(X) \) has a uniform distribution \( \sim U(a, b) \), \( H(P) \) has maximum value.

The posterior continuous entropy is

\[ H(\pi_{k-1}(\theta | u_{k-1})) = \int \pi_{k-1}(\theta | u_{k-1}) \log \left( \frac{1}{\pi_{k-1}(\theta | u_{k-1})} \right) d\theta. \]

The expectation over the kth response should be taken since the response to the kth item is
unknown. The expected posterior continuous entropy after administering item j as the kth item is

\[ E_{ik}(H(\pi_k(\theta | u_{k-1}, u_{i_k})) = \sum_{U=0}^{m} H(\pi_k(\theta | u_{k-1}, u_{i_k} = U))P(u_{i_k} = U | u_{k-1}) \]

\[ = \sum_{U=0}^{m} \left[ \int \pi_k(\theta | u_{k-1}, u_{i_k} = U) \log \left( \frac{1}{\pi_k(\theta | u_{k-1}, u_{i_k} = U)} \right) d\theta \right] \left[ \int P(u_{i_k} = U | \theta)\pi_{k-1}(\theta | u_{k-1}) d\theta \right]. \]  \hspace{1cm} (2.37)

The expected continuous entropy should be minimized to minimize the uncertainty. Items satisfying the following rule should be selected,

\[ i \equiv \arg \min_i \{E_{i_k}(H(\pi_k(\theta | u_{k-1}, u_{i_k}))); i \in R_k \}. \]

The entropy-based algorithm was introduced into CD-CAT (Xu, Chang, & Douglas, 2005; Cheng, 2009). Wang and Chang (2010) applied it into MCAT adopting M3PL dichotomous model. When the entropy is used, items are selected to directly monitor the entropy of the posterior distribution of interim ability estimates (\( \theta^{Hat} \)) and the stopping rule of adaptive testing is that the uncertainty of the ability estimates is below a certain tolerance level. There are several advantages of entropy approach. One strength is that it does not require the interim ability estimate should be close to the true ability, which caused problems in FI-based approach. Another desirable feature is that it quantifies the uncertainty with respect to the entire posterior distribution rather than the point estimate so that it could avoid the error introduced by the interim estimation (Wang & Chang, 2011).

The CEM could be interpreted as information (Wang & Chang, 2011). As Renyi (1961) claimed, the entropy of a probability distribution can be interpreted as a measure of information in addition to the interpretation of a measure of certainty. Wang and Chang (2011) analytically showed that CEM and mutual information (MI) has similar approach but use different entropy as the baseline.

Wang and Chang (2011) noticed that CEM is similar to MEPV (van der Linden, 1998).
The difference between them is that MEPV uses posterior variance while CEM uses the continuous entropy of the posterior distribution.

Note that which item selection criterion would be the best is directly associated with the goal of testing, including whether abilities measured by the test are all intentional or include nuisance abilities and whether the interest is in scoring separate abilities or a composite of abilities (Mulder & van der Linden, 2009, 2010; van der Linden, 1996). Nuisance abilities in educational testing means that test items are sensitive to abilities related to their format besides the primary abilities tested (Mulder & van der Linden, 2009). An example of nuisance abilities is a mathematics test depending on verbal abilities required to understand items (van der Linden, 1999). A different optimal design criterion to each case for MCAT item selection is more appropriate (Mulder & van der Linden, 2009). Thus, some studies specifically proposed certain algorithm to for specific goals of testing. Van der Linden (1999) discussed composite ability case, a linear combination of the abilities incorporating a vector of nonnegative weights. Veldkamp and van der Linden (2002) investigated five cases, all abilities intentional, intentional and nuisance abilities, one composite ability as an explicit linear combination of theta, simple multidimensional ability structure, and simple unidimensional ability structure. Mulder and van der Linden (2009) examined KL-based item selection in the case of a weighted combination of ability parameters.

**Item Bank Construction**

The item pool structure affects the robustness and efficiency of item selection criteria and influences the quality of theta estimation in CAT (van der Linden, 1998; Lima Passos, Berger, & Tan, 2007b; Dodd et al., 1995). The elements of the item pool structure affecting CAT performance include the item bank size, composition of the items in the bank, and the individual
item characteristics. For example, van Rijin, Eggen, Hemker, and Sanders (2002) explored the number of items in the item bank, the number of categories of the items in the item bank, and the distribution of the item category parameters.

The rule of thumb is that dichotomous item pool size should be around 12 times the CAT test length (Stocking, 1994; Way, 1998). Chang and Zhang (2002), however, suggested larger ratios considering item pooling effect. For polytomous item banks, Dodd et al. (1995) proposed smaller sizes and claimed that item banks with 30 items maybe be sufficient to achieve the desirable theta estimation accuracy level because polytomous items provide more information than dichotomous ones. Nevertheless, Dodd et al. (1995) did not specifically point out the ratio of the item bank size and the test length. Existing studies on polytomous items, however, assume ideal item bank size. In the polytomous UCAT studies, van Rijin, Eggen, Hemker, and Sanders (2002) used item bank size 150 and 500 items fitting UGPCM model. Lima Passos et al. (2007a, 2007b) used item banks of 300 and 600 for UGRM and UNRM model for the study of CAT during the early stage and the test length of 15. In current dichotomous MCAT studies, the item pool size containing items from M2PL or M3PL is about 420-500 (Wang & Chang, 2011; Segall, 2001; Van der Linden, 1999; etc.). Mulder and van der Linden (2009), however, set pool size at 200 for items from a M3PL model.

In UCAT context, van Rijin, Eggen, Hemker, and Sanders (2002) claimed that item banks with trinary items, with three categories, gave the best performance when considering the total maximum score and more categories did not produce better results. There have been no studies exploring the categories of polytomous MIRT model in MCAT.

Van Rijin, Eggen, Hemker, and Sanders (2002) investigated three item category distributions for UGPCM model, \( N(-1, 1) \), \( N(0, 1) \), \( N(1, 1) \) when the discrimination distribution
is set as \( \log a \sim N(0,0.5) \). Van der Linden (1998) used \( a \sim U(0.5, 1.5) \) and \( b \sim U(-4, 4) \) for U2PL model. Van der Linden (1999) simulated \( a_1, a_2 \sim U(0, 1.3), d \sim U(-1.3, 1.3) \) and these distributions roughly correspond to the parameter values in a two-dimensional ACT Assessment Program Mathematics Item pool. Wang and Chang (2011) also used the same simulated distribution. Mulder and van der Linden (2009) generated \( a_1, a_2 \sim N(1, 0.3), b \sim N(0, 3) \).

Furthermore, a larger item bank might be required by concerns and issues such as content validity, item exposure, and test security for high stakes testing (Dodd et al., 1995).

**Latent Trait Estimation**

Two main categories of latent trait estimation are the maximum likelihood estimation method (MLE) and Bayesian methods. The MLE is that the estimate of examinee’s location is the theta-vector that results in the highest probability for the observed string of item scores (Reckase, 2009). The limitation of MLE is no finite values as the maximum exist if all responses are correct or wrong (Hambleton et al., 1991). Van der Linden (1999) pointed out that the MLE approach of theta is considerably biased and inefficient, even for linear combination of sub-scale abilities in MCAT. When MLE yields infinite values, extreme values that are typically observed in practice, such as 4 or -4, are assigned (Reckase, 2009).

A number of statistics can be used to estimate theta through using the Bayesian method. The expected a posteriori (EAP) and the maximum a posteriori (MAP) are two methods commonly discussed. Several polytomous CAT studies have adopted Bayesian methods (De Ayala, 1992; Chen, Hou, Fitzpatrick, & Dodd, 1995). EAP is also commonly used in MCAT (e.g. Wang, Chang, & Boughton, 2011; Wang & Chang, 2011; Segall, 1996; Luecht, 1996). EAP approach is often preferred in MCAT with several reasons. First, EAP approach is available for all response patterns, including null or perfect response (Bock & Mislevy, 1982; Luecht, 1996).
Second, EAP's mean squared error across all theta levels is smaller than that of MLE (Bock & Mislevy, 1982). The Bayesian approach could incorporate the information of thetas into the prior distribution of the abilities. "It is well known that, provided the information leads to a location of the prior distribution at the true values of the parameters, the result is a posterior that tends to be more informative than the sampling distribution of the ML estimators" (van der Linden, 1999, p.410). One limitation is that EAP also tends to overestimate the true correlations between the multivariate latent traits (Luecht & Miller, 1992; Segall, 1996). Reckase (2009) pointed out the regression effect of the Bayesian approach that EAP estimator is closer to the location of the mean vector for the prior distribution than is the MLE. The regression effect leads to statistically biased estimates of theta vectors but it is a reasonable trade-off for insuring finite estimates of thetas (Reckase, 2009). Bayesian procedure (Segall, 1996) is suggested when the test length is short in an MCAT. As the number of items increase, MLE and MAP achieve similar estimates (Reckase, 2009).

When the Bayesian method is used, a prior distribution of latent trait needs to be choosen. Reckase (2009) recommended taking a prior distribution either based on previous analyses of test data or from general knowledge about the form of distributions typically found in educational or psychological settings. The standard multivariate normal distribution with the identity matrix for the variance/covariance matrix is commonly taken as the prior distribution for Bayesian approaches when there is little empirical information about the distribution of thetas.

Other methods include the proposal by Mislevy (1984) that using pseudo-counts computed over the joint posterior for all examinees directly obtain better estimates of the population correlation or variance-covariance matrices. Segall (1996) also proposed the solutions for multidimensional Bayesian modal estimators that could be generalized to orthogonal or
oblique traits for a multivariate normal prior probability density. When the number of dimensions is more than two or three, estimating the modes of the posterior distribution might be easier in practice than estimating EAPs (Segall, 1996).

**Stopping Rules**

Several stopping rules have been discussed for polytomous CATs. One type of rule is to specify some minimum/maximum static stopping criterion such as fixed test length (Dodd et al., 1995). Another type of rule is dynamic stopping criterion that includes the minimum information stopping rule and achieving a pre-specified value for the standard error associated with the current theta estimate (Dodd et al., 1995). In the K-12 setting, it is hard to explain to examinees about the CAT rationale, and students will question the fairness of the testing if different test lengths are used. Hence, the fixed test length will be used in the test design; therefore, the stopping rule used is this study will be attaining specified test length(s).

The test length, however, is an important consideration. The initial stage of a CAT with a long test length or a CAT with a short test length is one of foci in this study. When the test length of a CAT is long (L>20), the differences among a variety of item selection criteria are negligible because the theta estimation will end up with an accurate and precise value (van der Linden, 1998; Lima Passos, Berger, & Tan, 2007). A variety of studies have shown such results (van der Linden & Pashley, 2000; van Rijin, Eggen, Hemker, & Sanders, 2002; Chen, Ankenmann, & Chang, 2000; etc.). When the test length is short (L<20), the attenuation paradox (Lord, 1968) and multiple maximum values (Samijima, 1973) will affect latent trait estimation (Lima Passos, Berger, & Tan, 2007).

According to van der Linden and Pashley (2000), three stages of a CAT are initial, interim, and final trait estimation, and unique requirements and problems exist in each stage. The
problems of the early stage have already been discussed, the attenuation paradox. The relevant asymptotic properties of estimators, including normally distributed and variance equaling the inverse of Fisher Information, do not hold early in a CAT (van der Linden & Pashley, 2000). A variety of studies have investigated the early stage issues in dichotomous UCAT (Chang & Ying, 1996; Chang & Ying, 1999; Chang & Ying, 2008; etc.) and polytomous UCAT (Lima Passos, Berger, & Tan, 2007a; Lima Passos, Berger, & Tan, 2007b; etc.). However, no studies in the literature have investigated this issue in polytomous MCAT.

With respect to the normality of estimated theta, a test of 10 to 20 items has satisfactory normality (Samejima, 1977; Chen et al., 2000; Lima Passos, Berger, & Tan, 2007a). The variance when administering 10 or 20 items still has a larger standard error than the ideal lower bound (Chen et al., 2000; Lima Passos, Berger, & Tan, 2007a). The bias and MSE were found to be high for the first 10 items in previous studies (Chen et al., 2000; Chen & Liou, 2000; van der Linden & Pashley, 2000); therefore, Lima Passos, Berger, and Tan (2007a) set 15 items as the upper boundary during early CAT in the case of polytomous UCAT.

Various test length such as 25 (Wang & Chang, 2011), 30 (Mulder & van der Linden, 2009), and 50 (van der Linden, 1999) have been discussed in recent MCAT studies of dichotomous items. In UCAT studies of polytomous items, test lengths such as 30 (van Rijn, Eggen, Hemker, & Sanders, 2002) and 15 for early stage CAT (Lima Passos, Berger, & Tan, 2007a, 2007b) have been set up. Some studies investigated the performance of item selection methods at different stages of CAT. For example, van der Linden (1998) explored test lengths of L=5, 10, 20, 30 for U2PL model for exploring the effect of Bayesian-based models at different stages of UCAT.
CHAPTER 3

METHODOLOGY

A simulation study is conducted assuming that the latent variable space is two-dimensional, and that item parameters in item banks have been accurately calibrated and are considered as the true item parameters.

Study Design

The main focus of this study interest is item selection methods in MCAT context under various conditions. Four item selection methods are investigated in the simulation study: (1) D-Optimality, (2) KI, (3) CEM, and (4) MI. These four methods were selected due to either the high level of popularity or outstanding performance in previous studies. In addition, Wang and Chang (2011) explored these four methods adopting multidimensional three-parameter logistic model (M3PL) for dichotomous items. Although all item selection methods could be interpreted as information-based approaches, variations exist among methods. The D-optimality and the KI use the local and global information concept, respectively, whereas, the CEM employs an entropy concept and MI could be categorized as a Bayesian method. MI is selected in the Bayesian methods because MI has higher robustness than $K^p$ method in terms of error in ability estimation (Mulder & van der Linden, 2010; Wang & Chang, 2011). Comparing these four methods helps to understand the effect of local information, global information, Bayesian approach, and entropy control for item selection.

In the simulation, the integration is taken over the range of ability [-3, 3].

Item Selection Methods

The formulas of four item selection methods compared in the simulation, including FI-based D-Optimality, KL information, MI, and CEM, are presented as follows.
FI-based D-optimality.

For the M3PL model, FI matrix is

\[
I_i(\theta) = \frac{Q_i(\theta)[P_i(\theta) - c_i]^2}{P_i(\theta)(1 - c_i)} \begin{bmatrix}
    a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{ip} \\
    a_{i1}a_{i2} & a_{i2}^2 & \cdots & a_{i2}a_{ip} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{i1}a_{ip} & a_{i2}a_{ip} & \cdots & a_{ip}^2
\end{bmatrix},
\]

(3.1)

For the MGPCM model, FI matrix is

\[
I_i(\theta) = \left( \sum_{u=0}^{m_i} u^2 P_{iu}(\theta) - E_i^2 \right) \begin{bmatrix}
    a_{i1}^2 & a_{i1}a_{i2} & \cdots & a_{i1}a_{ip} \\
    a_{i1}a_{i2} & a_{i2}^2 & \cdots & a_{i2}a_{ip} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{i1}a_{ip} & a_{i2}a_{ip} & \cdots & a_{ip}^2
\end{bmatrix},
\]

(3.2)

where

\[
E_i = \sum_{u=0}^{m_i} u P_{iu}(\theta) = \frac{\sum_{u=0}^{m_i} u \exp(u a_i^T \theta - \sum_{t=0}^{u} \beta_{it})}{\sum_{u=0}^{m_i} \exp(u a_i^T \theta - \sum_{t=0}^{u} \beta_{it})}.
\]

D-optimality maximizes the determinant of FI matrix,

\[
\arg \max_{i \in R_k} \det \left( I_{i_k} \left( \theta_{Hat}^{i_k} \right) + I_{i_k} \left( \theta_{Hat}^{k-1} \right) \right).
\]

(3.3)

Kullback-Leibler information index.

For the M3PL model, the item KL information for item i is

\[
KL_i(\theta_{Hat} \parallel \theta) = P_i(\theta) \log \left( \frac{P_i(\theta)}{P_i(\theta_{Hat})} \right) + [1 - P_i(\theta)] \log \left( \frac{1 - P_i(\theta)}{1 - P_i(\theta_{Hat})} \right).
\]

(3.4)

For MGPCM model, the item KL information for item i is

\[
KL_i(\theta_{Hat} \parallel \theta) = \sum_{t=0}^{m_i} P_i(\theta) \log \left( \frac{P_i(\theta)}{P_i(\theta_{Hat})} \right),
\]

(3.5)

where \( L_i(\theta; U_i) \) represents the true probability distribution of the observed data and \( L_i(\theta_{Hat}; U_i) \)
usually refers to the estimated probability distribution of \( L_i(\theta; U_i) \), \( i=0, 1, \ldots, m_i \) refers to the score category of polytomous items where score equals to 0, 1, \( \ldots, m_i \) respectively, and the number of score categories of polytomous items is \( m_i + 1 \).

The KL information index (KI) is used to serve as a criterion for item selection.

\[
KI(\hat{\theta}_n) = \int_{\mu_{a\theta} - \delta_n}^{\mu_{a\theta} + \delta_n} KL(\hat{\theta}_n \parallel \theta)d\theta,
\]  
where \( \delta_n \) determines the size of the region over which the average is computed. \( \delta_n = \frac{d}{n^{1/2}} \), \( d=3 \) and \( n \) refers to number of items which have been administered.

**Mutual information.**

Mutual information is as follows:

\[
\arg \max_{i \in \mathcal{R}} MI(\theta; \mu_i) = \arg \max_{i \in \mathcal{R}} \frac{\log f(\mu_i | \mu_{k-1})}{f(\mu_i | \mu_{k-1})} \int_0^1 f(\theta; \mu_i | \mu_{k-1}) \log \frac{f(\mu_i | \theta)}{f(\mu_i | \mu_{k-1})} d\theta,
\]  
where \( \mu_k \) is the response to the \( i \)th candidate item in the candidate item pool when test length \( L=k \).

**Continuous entropy method.**

The formula of CEM is as follows:

\[
E_i(\pi_k(\theta | \mathbf{u}_{k-1}, u_i)) = \sum_{U=0}^{m_i} H(\pi_k(\theta | \mathbf{u}_{k-1}, u_i = U))P(u_i = U | \mathbf{u}_{k-1})
\]  
\[
= \sum_{U=0}^{m_i} \left[ \pi_k(\theta | \mathbf{u}_{k-1}, u_i = U) \log \left( \frac{1}{\pi_k(\theta | \mathbf{u}_{k-1}, u_i = U)} \right) \int P(u_i = U | \theta) \pi_{k-1}(\theta | \mathbf{u}_{k-1}) d\theta \right].
\]  
and the selection criterion is as follows,

\[
i = \arg \min_{i \in \mathcal{R}} \left\{ E_i(\pi_k(\theta | \mathbf{u}_{k-1}, u_i)) ; i \in \mathcal{R} \right\},
\]  
where \( \mathbf{u}_{k-1} \) is the vector with the responses to the first \( k-1 \) items, \( \mu_{i_1} \) is the response to the \( i \)th
candidate item in the candidate item pool when test length $L=k$.

**Test Formats**

By introducing polytomous items into tests, the comparison of different test formats is an important study interest. A number of studies have been conducted to discuss a variety of item types and test formats (Weiner & Thissen, 1993; Qualls, 1995; Hardy, 1995; Ackerman & Smith, 1988; Birenbaum & Tatsuoka, 1987; Nandakumar, Yu, Li, & Stout, 1998; Johnson & Carlson, 1994). Each item type and each test format have their advantages and disadvantages. For example, some attractive features of multiple-choice items include the reliable and objective scoring and lower cost (Weiner & Thissen, 1993; Hardy, 1995) but have several undesirable features include emphasizing memorization and allowing test preparation. Polytomously-scored items have strengths such as providing diagnostic information and more systematic validity while their weaknesses include the difficulty to score objectively and reliably and the higher cost (Weiner & Thissen, 1993; Hardy, 1995). Using different types of items in one test might allow the concatenation of the advantages of each item type while compensating for disadvantages (Weiner & Thissen, 1993). Therefore, the test format with mixed types of items has been used extensively, including in NAEP, K-12 state assessment, and licensure exams (Yao & Schwarz, 2006; Nandakumar, Yu, Li, & Stout, 1998; Johnson & Carlson, 1995).

There are several different test formats. The first type of format is the uni-type test format that administers all items with the same number of response categories throughout the test, called uni-type test format in the study. When the number of response categories is two, it is a test administering dichotomously-scored items only, named DICHTYPE test in the study. When the number of response categories is larger than two, it is a test delivering polytomously-scored items only, named POLYTYPE test in the study. Items with the same number of categories might
belong to different item types. For instance, items with two response categories could be multiple choice items or true-or-false items. In this study, items with the same number of categories are simulated from the same MIRT model. The second type is to administer items with different number of response categories in a test, called mix-type test format in the study. Each item type is simulated from one MIRT model. In this study, the mixed test format contains dichotomous items with two response categories and polytomous items with three response categories. The study explores two categories for the mix-type test format. One category is the DPMIX test format which administers dichotomous items before selecting polytomous items. The other type is the PDMIX test format which delivers polytomous items at the beginning of the test and then administers dichotomous items at the later stage of the test. The proportion of dichotomous items is 75% and the proportion of polytomous items is 25%. It is based on the proportion in state assessments and NAEP test (Yao & Schwarz, 2006; Nandakumar, Yu, Li, & Stout, 1998). In the current study, given that test length is set to 25, the number of multiple-choice items is set as 18 and the number of polytomous items is set as 7.

Among the types of test format discussed in the study, the mix-type test formats have been used in assessments extensively. The POLYTYPE test format is discussed in some studies for comparison purpose by converting the standardized tests into a version with polytomous items only (Birenbaum & Tatsuoka, 1987; Ackerman & Smith, 1988). As discussed previously, with the importance of diagnostic purpose and improving instruction in current assessments, the POLYTYPE test format will become a major test format in formative assessment. Thus, it is imperative to explore POLYTYPE test format and mix-type test format. Under the mix-type test format, the exploration of PDMIX and DPMIX test formats is to study the impact of ordering different types of items in the mix-type test format.
For polytomous items discussed in this study, including polytomous items in POLYTYPE, DPMIX, and PDMIX test formats, the MGPCM (Reckase, 2009) is used for polytomously-scored items. MGPCM is selected because GPCM is fitting for data ordered to represent varying degrees of $\theta$ (Dodd, De Ayala, & Koch, 1995) and commonly used in high-stakes testings (Yao & Schwarz, 2006).

$$P(U_{ij} = u | \theta_j) = \frac{\exp(u\mathbf{a}_j^T\theta_j - \sum_{\ell=0}^{u - 1} \beta_{\ell})}{\sum_{v=0}^{m_i} \exp(v\mathbf{a}_j^T\theta_j - \sum_{\ell=0}^{v - 1} \beta_{\ell})},$$  \hspace{1cm} (3.14)

where $u = 0, 1, \ldots, m_i$, the score given to examinee $j$ on the item $i$, and $m_i$ is the highest score for item $i$. $\beta_{\ell}$ is the threshold parameter for score category $u$, $\beta_0 = 0$, $\theta$ and $\mathbf{a}$ are vectors of latent trait and discrimination respectively in MIRT models.

For dichotomous items discussed in this study, including dichotomous items in DPMIX and PDMIX test formats, the M3PL model is adopted for dichotomous items with two score categories (e.g. multiple-choice items).

$$P(U_{ij} = 1 | \theta_j, \mathbf{a}_i, c_i, d_i) = c_i + (1 - c_i) \frac{\exp(\mathbf{a}_i^T\theta_j + d_i)}{1 + \exp(\mathbf{a}_i^T\theta_j + d_i)},$$  \hspace{1cm} (3.15)

where $d_i$ is the intercept parameter, $\theta_j$ and $\mathbf{a}_i$ are vectors of latent trait parameter and discrimination parameter in MIRT models. The study, however, takes $c_i = 0$ so that the model can be considered as M2PL model.

**Simulating Examinees**

Similar to previous MCAT studies (Wang & Chang, 2011; van der Linden, 1999; Finkelman, Nering, & Roussos, 2009), examinee responses are simulated with true abilities on a two-dimensional grid spanning the square $\theta_1$ and $\theta_2 = -3.0, -2.5, \ldots, 2.5, 3.0$. Crossing 13
discrete points on each of two dimensions generates a grid of 169 $\theta$ vector points. Each theta vector point corresponds to each simulee. Therefore, the simulated examinee population is 169. The reason of simulating 169 discrete theta points instead of simulating examinees from a standard bivariate normal distribution is that one of main study interest is to investigate the conditional estimation accuracy of four item selection methods.

The study compares POLYTYPE test format, DPMIX test format, and PDMIX test format with one research interest of this study of exploring the estimation accuracy under different test formats and under different order of item types in the mix-type format. 500 replications were run at each theta point. Therefore, for each test format, including POLYTYPE test format, DPMIX test format, PDMIX test format, 84500 replications were conducted, which is equal to 500 times multiply by 169 theta points.

**Item Pool Structure**

When the POLYTYPE test format is explored, items with three response categories are explored for polytomous items. Van Rijn et al (2002) found the optimal number of categories of a polytomous item is three considering the maximum score of an item. The item discrimination parameters are to mimic distributions that roughly follow the distribution of the real data in a two-dimensional ACT Assessment Program Mathematics Item pool (van der Linden, 1996, 1999). In the study, the distribution of item discrimination parameters are as follows:

$$a_{i1}, a_{i2} \sim U(0, 1.3).$$

For threshold parameters of polytomous items, the distribution is revised based on previous studies (van Rijn et al., 2002) to generate polytomous item threshold parameters. $\beta_{i0}$ is arbitrarily set as zero. $\beta_{i0} = 0$, $\beta_{i1}$ follows truncated N(-1, 1). N(-1, 1) is truncated at 0, and any values greater than or equal to 0 is not used. $\beta_{i2}$ follows truncated N(1, 1). N(1, 1) is also truncated at 0, and any values less than 0 is not used. The polytomous item pool
size is set as 400 for the MCAT simulation using the POLYTYPE test format.

Table 3.1.

*Polytomous Item Bank Statistics under Uni-Type Test Format (n=400)*

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.6473</td>
<td>0.6640</td>
<td>0</td>
<td>-1.2788</td>
<td>1.2787</td>
</tr>
<tr>
<td>SD</td>
<td>0.3817</td>
<td>0.3693</td>
<td>0</td>
<td>0.7967</td>
<td>0.8139</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.0070</td>
<td>0.0135</td>
<td>0</td>
<td>-3.8946</td>
<td>0.0035</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.2978</td>
<td>1.2940</td>
<td>0</td>
<td>-0.0022</td>
<td>3.6323</td>
</tr>
</tbody>
</table>

Table 3.2.

*Polytomous Item Bank Statistics in Mix-Type Test Format (n=100)*

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.6377</td>
<td>0.6392</td>
<td>0</td>
<td>-1.2387</td>
<td>1.2076</td>
</tr>
<tr>
<td>SD</td>
<td>0.3692</td>
<td>0.3722</td>
<td>0</td>
<td>0.7495</td>
<td>0.8342</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.0070</td>
<td>0.0135</td>
<td>0</td>
<td>-3.3868</td>
<td>0.0481</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.2933</td>
<td>1.2877</td>
<td>0</td>
<td>-0.0297</td>
<td>3.4159</td>
</tr>
</tbody>
</table>

Dichotomous items in mix-type test format are simulated from M2PL model. The distribution of dichotomous item parameters are $a_{i1}, a_{i2} \sim U(0,1.3)$ and $b_{i1} \sim U(-1.3,1.3)$. These item parameters are simulated following the study design by Wang and Chang (2011). The dichotomous item pool size for the mixtype test format is set as 300. The polytomous item pool
size is set as 100. 100 polytomous items are randomly drawn from the polytomous item pool for the POLYTYPE item pool for meeting the comparison purpose.

Table 3.3.

*Dichotomous Item Bank Statistics under Uni-Type Test Format (n=400)*

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>d</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>0.6473</td>
<td>0.6568</td>
<td>-0.0541</td>
<td>0</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.3854</td>
<td>0.3454</td>
<td>0.7356</td>
<td>0</td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>0.0022</td>
<td>0.0142</td>
<td>-1.2971</td>
<td>0</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>1.2907</td>
<td>1.2902</td>
<td>1.2853</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.4.

*Dichotomous Item Bank Statistics under Mix-Type Test Format (n=300)*

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>d</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>0.6583</td>
<td>0.6529</td>
<td>-0.0742</td>
<td>0</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.3909</td>
<td>0.3432</td>
<td>0.7333</td>
<td>0</td>
</tr>
<tr>
<td><strong>Minimum</strong></td>
<td>0.0022</td>
<td>0.0142</td>
<td>-1.2971</td>
<td>0</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>1.2907</td>
<td>1.2901</td>
<td>1.2853</td>
<td>0</td>
</tr>
</tbody>
</table>

In the DPMIX test format, the first 18 items are selected from the dichotomous item pool, and the 19th item to the 25th item are selected from the polytomous item pool. In the PDMIX test format, the first 7 items are selected from the polytomous item pool, and the 8th item to the 25th
item are selected from the dichotomous item pool.

**Latent Trait Estimation**

EAP is used as the latent trait estimation approach when the test is ongoing. The prior distribution is to use the standard bivariate normal distribution with an identity matrix for the variance/covariance matrix according to the suggestion by Reckase (2009) when there is little empirical information about the form of distribution of thetas. This suggestion is based on years of experiences in practice and studies (Reckase, 2009).

\[
f(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left( \frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y} \right) \right), \tag{3.16}
\]

where \( \rho \) is the correlation between \( X \) and \( Y \), \( \mu = \begin{pmatrix} u_x \\ u_y \end{pmatrix} \), \( \sum = \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} \), and \( u = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \), \( \sum = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \) in the study.

**Stopping Rules**

This is a fixed-length MCAT and the test length is set to 25. The estimation accuracy of each item selection method as a function of the test progression is recorded to monitor the performance of selection algorithms at various stages of the tests. The study also examines the performance with respects to estimation accuracy and the item overlap rate among methods at initial, early, middle, and final stage of PMCAT, corresponding to test lengths of 2, 7, 18, and 25. The purpose of examining the initial stage is to investigate the attenuation paradox issue. Because the DPMIX format changes from dichotomous items to polytomous items at the 7th item, and because the PDMIX format switches from polytomous items to dichotomous items at the 18th item, 7th and 18th is set as the end of early stage and middle stage. In this way, when comparing DPMIX and PDMIX, there are only one type of items for each test format in the early
stage and final stage, making them comparable. Specifically, there will be polytomous items only
in early stage of DPMIX while dichotomous items only in PDMIX format; and there will be
polytomous items only in the final stage of DPMIX while dichotomous items only in PDMIX
format. Moreover, the first 7 items are 25% of the total test and the first 18 items are 75% of the
total test, which are reasonable length for the early and middle stage.

**Evaluation Criteria**

The simulation study is designed to evaluate and compare the ability estimation accuracy
and conditional estimation accuracy when using each of the four selected item selection methods
under different test formats. The conditional comparison is investigated conditioning on
examinee's ability level.

In the following evaluation criteria, n refers to the size of the item pool; N refers to the
size of the examinee population; L is the length of the adaptive test; i is the index of the items in
the item pool, and \(i = 1, \ldots, n\); j is the index of the examinees taking the tests, and \(j = 1, \ldots, N\).

The evaluation criteria of estimation accuracy at kth sub-dimension are as follows.

\[
Bias_k = \frac{1}{N} \sum_{j=1}^{N} (\hat{\theta}_{jk}^{Hat} - \theta_{jk}), \quad k = 1, 2. \quad (3.17)
\]

\[
MSE_k = \frac{1}{N} \sum_{j=1}^{N} (\hat{\theta}_{jk}^{Hat} - \theta_{jk})^2, \quad k = 1, 2. \quad (3.18)
\]

Euclidean distance as a global index of psychometric precision when the number of
dimension is 2:

\[
ED_j = \left[ (\hat{\theta}_{j_1}^{Hat} - \theta_{j_1})^2 + (\hat{\theta}_{j_2}^{Hat} - \theta_{j_2})^2 \right]^{1/2}.
\]

(3.19)

The Euclidean distance is a summary measure the estimation accuracy of the methods at
each ability point and it is used as evaluation criterion in several studies (Segall, 1996;
Finkelman et al., 2009; Wang & Chang, 2011). To measure the estimation accuracy of item
selection methods at each ability point, the average Euclidean distance (AED) is calculated. The conditional AED toward each item selection methods is calibrated and presented by the contour plots. Conditional AED at selected theta points are also provided. To present the overall performance of each item selection method under different test lengths, a prior weighted average euclidean distance (PAED, Finkelman et al., 2009; Wang & Chang, 2011), prior weighted mean square error in dimension one and dimension two (PMSE1, PMSE2), and prior weighted bias in dimension one and dimension two (PBIAS1, PBIAS2) are adopted as the evaluation criterion. The AED, MSE1, MSE2, BIAS1, and BIAS2 are averaged over the prior distribution of the ability. In this study, it is assumed that the prior distribution is the standard bivariate normal distribution. The PAED is calibrated as a function of test length.

The overlap of selected items among different item selection methods is calculated targeting at investigating the similarity between selection methods (Wang & Chang, 2011). Overlap rate is the proportion of common items between two selection methods. In the study, the overlap rate is compared toward the following six combinations: (1) D-Optimality vs. KI, (2) D-Optimality vs. MI, (3) D-Optimality vs. CEM, (4) KI vs. MI, (5) KI vs. CEM, (6) MI vs. CEM. The overlap rate is examined at early, middle and final stages, corresponding to test lengths of 7, 18, 25.

\[
\text{overlap}_{Method1, Method2} = \frac{1}{N} \sum_{i=1}^{N} \frac{\#(Method1 \cap Method2)}{L},
\]

where L is the test length, and N is the examinee population size, which is 169 in the study.

In order to compare the item selection pattern among item selection methods, the item discrimination index (MDISC) throughout the test progression is recorded. Geometrically, MDISC is the steepest slope on the item response surface (Reckase & McKinley, 1991).
where \( p \) is the number of ability dimensions, and \( a_{ik} \) is the discrimination parameter in \( k \)th dimension.

Wang, Chang, and Boughton (2011) claimed that MDISC is the primary criterion controlling item selection. MDISC is an overall measure of the capability of an item to distinguish between individual examinees that in different locations in the ability space.
CHAPTER 4  
RESULTS, FINDINGS, AND DISCUSSIONS

The discussion mainly consists of three parts: ability estimation accuracy, conditional ability estimation accuracy, and item selection pattern. Both figures and tables are used to present results. The figures show the general comparison. To distinguish differences of similar performance among item selection methods and test formats, tables provide numerical numbers to conduct accurate comparisons.

Ability Estimation Accuracy

The ability estimation accuracy is compared and discussed among item selection methods and among test formats, which will be examined in this section. In this study, ability estimation refers to unconditional ability estimation. PAED, PMSE1, PMSE2, PBIAS1, and PBIAS2 are used as the evaluation criteria. In contrast, the conditional ability estimation will be discussed in the following section. Conditional AED, MSE1, MSE2, BIAS1, and BIAS2 will be adopted as the evaluation criteria.

Comparing ability estimation accuracy among item selection methods are conducted under the POLYTYPE, DPMIX, and PDMIX test format. Figure 4.1, 4.2, and 4.3 shows the comparison among four item selection methods throughout the MCAT test under each test format. There are three plots in each of these three figures. In each plot, the X coordinate is the test length from L=1 to 25. In the first plot, the Y coordinate is the PAED; in the second right plot, the Y coordinate is PMSE1; and in the third plot, the Y coordinate is PMSE2. Table 4.1, 4.2, and 4.3 contains the PAED, PMSE1, PMSE2, PBIAS1, and PBIAS2 at four stages of a MCAT test under three test formats.

As shown in Figure 4.1, 4.2, and 4.3, D-optimality, MI, and CEM provide very similar
performance throughout the test under the POLYTYPE, DPMIX, and PDMIX. The KI method, however, shows larger estimation error than the other three methods, especially at the later stage.

Figure 4.1. PAED, PMSE1, and PMSE2 throughout the MCAT under the POLYTYPE format.
Figure 4.2. PAED, PMSE1, and PMSE2 throughout the MCAT under the DPMIX format.
Figure 4.3. PAED, PMSE1, and PMSE2 throughout the MCAT under the PDMIX format.
Table 4.1, 4.2, and 4.3 give the accurate comparison in PAED, PMSE1, PMSE2, PBIAS1, and PBIAS2 of four stages of a MCAT test among selection methods.

Table 4.1.

*PAED at Initial (L=2), Early (L=7), Middle (L=18), and Final Stage (L=25)*

<table>
<thead>
<tr>
<th>Method</th>
<th>PAED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Stage</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>1.039</td>
</tr>
<tr>
<td>KI</td>
<td>1.068</td>
</tr>
<tr>
<td>MI</td>
<td>1.030</td>
</tr>
<tr>
<td>CEM</td>
<td>1.033</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>1.121</td>
</tr>
<tr>
<td>KI</td>
<td>1.122</td>
</tr>
<tr>
<td>MI</td>
<td>1.108</td>
</tr>
<tr>
<td>CEM</td>
<td>1.102</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>1.053</td>
</tr>
<tr>
<td>KI</td>
<td>1.053</td>
</tr>
<tr>
<td>MI</td>
<td>1.039</td>
</tr>
<tr>
<td>CEM</td>
<td>1.116</td>
</tr>
</tbody>
</table>
Table 4.2.

*MSE at Initial (L=2), Early (L=7), Middle (L=18), and Final Stage (L=25)*

<table>
<thead>
<tr>
<th>Method</th>
<th>PMSE ($\theta_1$)</th>
<th>PMSE ($\theta_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Early</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>0.690</td>
<td>0.348</td>
</tr>
<tr>
<td>KI</td>
<td>0.733</td>
<td>0.536</td>
</tr>
<tr>
<td>MI</td>
<td>0.705</td>
<td>0.353</td>
</tr>
<tr>
<td>CEM</td>
<td>0.703</td>
<td>0.356</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>0.813</td>
<td>0.487</td>
</tr>
<tr>
<td>KI</td>
<td>0.820</td>
<td>0.624</td>
</tr>
<tr>
<td>MI</td>
<td>0.782</td>
<td>0.515</td>
</tr>
<tr>
<td>CEM</td>
<td>0.777</td>
<td>0.511</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>0.684</td>
<td>0.367</td>
</tr>
<tr>
<td>KI</td>
<td>0.735</td>
<td>0.518</td>
</tr>
<tr>
<td>MI</td>
<td>0.716</td>
<td>0.387</td>
</tr>
<tr>
<td>CEM</td>
<td>0.817</td>
<td>0.410</td>
</tr>
</tbody>
</table>
Table 4.3.

*PBIAS at Initial (L=2), Early (L=7), Middle (L=18), and Final Stage (L=25)*

<table>
<thead>
<tr>
<th>Method</th>
<th>PBIAS ($\theta_1$)</th>
<th></th>
<th></th>
<th></th>
<th>PBIAS ($\theta_2$)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>Early</td>
<td>Middle</td>
<td>Final</td>
<td>Initial</td>
<td>Early</td>
<td>Middle</td>
<td>Final</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.003</td>
<td>-0.001</td>
<td>0.010</td>
<td>0.007</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td>KI</td>
<td>0.004</td>
<td>0.005</td>
<td>0.001</td>
<td>0.0002</td>
<td>0.003</td>
<td>0.0009</td>
<td>0.0006</td>
<td>0.002</td>
</tr>
<tr>
<td>MI</td>
<td>0.0001</td>
<td>0.002</td>
<td>0.0001</td>
<td>-0.001</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.001</td>
<td>0.0008</td>
</tr>
<tr>
<td>CEM</td>
<td>-0.0007</td>
<td>-0.001</td>
<td>-0.006</td>
<td>-0.005</td>
<td>-0.002</td>
<td>-0.005</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>-0.003</td>
<td>-0.004</td>
<td>-0.005</td>
<td>-0.002</td>
<td>0.0002</td>
<td>0.0008</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td>KI</td>
<td>0.004</td>
<td>0.003</td>
<td>-0.0009</td>
<td>-0.007</td>
<td>0.004</td>
<td>0.003</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>MI</td>
<td>0.0002</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.0004</td>
<td>0.001</td>
<td>-0.002</td>
<td>-0.0005</td>
<td>-0.002</td>
</tr>
<tr>
<td>CEM</td>
<td>-0.001</td>
<td>0.004</td>
<td>0.004</td>
<td>0.002</td>
<td>-0.003</td>
<td>-0.007</td>
<td>-0.006</td>
<td>-0.005</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D-opt</td>
<td>-0.001</td>
<td>0.0007</td>
<td>0.002</td>
<td>0.002</td>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.001</td>
<td>-0.002</td>
</tr>
<tr>
<td>KI</td>
<td>-0.003</td>
<td>-0.002</td>
<td>-0.001</td>
<td>-0.001</td>
<td>-0.002</td>
<td>-0.0009</td>
<td>0.0004</td>
<td>-0.0008</td>
</tr>
<tr>
<td>MI</td>
<td>-0.0007</td>
<td>-0.0006</td>
<td>-0.0009</td>
<td>0.001</td>
<td>0.0002</td>
<td>0.001</td>
<td>0.0001</td>
<td>-0.0004</td>
</tr>
<tr>
<td>CEM</td>
<td>-0.220</td>
<td>-0.077</td>
<td>-0.055</td>
<td>-0.044</td>
<td>-0.219</td>
<td>-0.111</td>
<td>-0.060</td>
<td>-0.045</td>
</tr>
</tbody>
</table>

In general, D-optimality provides the best performance, followed by MI, CEM, KI. Moreover, several findings are explored as follows.

First, the performance of the KI method is not as well as the other three methods under three test formats, and this is similar to the results found by Wang and Chang (2011). Wang and
Chang (2011) pointed out that items with a larger KL index do not necessarily have higher
discrimination power to discriminate true theta from interim theta based on their MCAT study
adopting the M3PL model. If items with the following feature, these items might have high
MDISC while they do not necessarily have high discrimination power:
\[
\sum_{k=1}^{P} a_{i_k} (\theta_{k}^{Hat} - \theta_k) = 0.
\]

In the discussion of the item selection pattern found later in this study, KI is found to favor
polytomous items with high MDISC index and high discrimination parameters in both
dimensions. However, these items might not have high discrimination power as pointed out by
Wang and Chang (2011). Furthermore, since polytomous items belong to compensatory models.
the unsatisfactory performance of KI for MCAT using polytomous items confirms another
finding by Wang and Chang (2011). If the MIRT models are compensatory models, then for
certain interim thetas, KL is always zero for certain points. Then in the integration, large
information at other points average out zero information (Wang & Chang, 2011). These are
several main reasons explaining why KI does not show good performance for MCAT test
adopting polytomous items. Nevertheless, potential causes might exist for the unsatisfactory
performance of KI in MCAT adopting polytomous items. More studies are needed.

Second, D-optimality shows unstable performance at the beginning of the test. D-
optimality has larger estimation error at the initial stage than the other two methods, MI and
CEM. However, at the early stage when test length is 7 (L=7), D-optimality uniformly provides
the best performance. When examining the figures and tables, the estimation error of D-
optimality becomes superior to errors of other methods when test length is around 4. This
phenomenon shows that D-optimality seems to become stable when test length is around 4. It
confirms that the D-optimality has the attenuation paradox issue at the beginning of the test.
under the PMCAT setting. However, the test length of the unstable stage under the PMCAT needs to be further explored by using different item pool structures and test designs.

Third, D-optimality, MI, and CEM yields close performance in ability estimation when item bank size is 400. However, note that the item pool size might have effect on the performance of these three methods. Notice that in POLYTYPE format, these three methods show very similar estimation accuracy from the beginning of the test. Under the PDMIX format, three methods show larger discrepancies in performance throughout the test. Recall that the polytomous item bank size is 400 for the POLYTYPE format and 100 for polytomous item bank of the mix-type format. Furthermore, the PDMIX polytomous item bank was randomly drawn from the POLYTYPE polytomous item bank, meaning that they have similar structure. In this case, at the beginning of POLYTYPE and PDMIX formats, the only difference is the size of polytomous items. Thus, it seems reasonable to derive that when the polytomous item bank size drops from 400 to 100, the performance of item selection methods is affected.

One change with different test formats is that the POLYTYPE format shows superior accuracy than the PDMIX formats at the end of the early stage for all item selection methods based on Table 4.1 and 4.2. It means that the larger the bank size is, the better the estimation accuracy is. Thus, with the same percentage of high quality item assumption, the absolute number of high quality items in the POLYTYPE polytomous item bank is larger than the absolute number of high quality items in the PDMIX polytomous item bank. Accordingly, the speed of exhausting high quality items of the 400-item bank is slower than the speed of the 100-item bank. In addition, the highest quality items in the 400-item bank might be better than the highest quality items in the 100-item bank. Therefore, whether the item bank size is large enough affect the performance.
Another observation in change is that at the beginning of the PDMIX format, three selection methods do not have similar performance pattern as they do under POLYTYPE format. Specifically, D-optimality, CEM, and MI show similar performance under POLYTYPE format, but show large differences among them under PDMIX format. It could be interpreted that a certain method is affected more than other methods by the change of bank size. For example, based on table 4.1, CEM changes its PAED at early stage from 0.737 under POLYTYPE to 0.812 under PDMIX format, and this change is larger than the change of other methods (e.g. D-optimality from 0.723 to 0.768, KI from 0.881 to 0.868, MI from 0.731 to 0.786). Hence, CEM seems to be affected more apparently by the shrink in item bank size.

**Ability Estimation Accuracy Comparison among Test Formats**

The estimation accuracy of four item selection methods are compared among three test formats, POLYTYPE, DPMIX, and PDMIX.

Estimation accuracy of D-optimality, KI, MI, and CEM methods under three test formats, POLYTYPE, DPMIX, and PDMIX are compared in Figure 4.4, 4.5, 4.6, and 4.7. There are three plots in each of these four figures. In each figure, the X coordinate is the test length from L=1 to 25. In the first plot, the Y coordinate is the PAED; in the second plot, the Y coordinate is PMSE1; and in the third plot, the Y coordinate is PMSE2.

Generally speaking, under all three test formats and for all four item selections, at most stages of the MCAT test, the POLYTYPE test format is performing the best, followed by the PDMIX and DPMIX test format. At the final stage of the test, the POLYTYPE format shows much higher estimation accuracy than the other two test formats for all item selection methods. However, several results need to be discussed.
Figure 4.4. PAED, PMSE1, and PMSE2 of item selection methods under D-optimality.
Figure 4.5. PAED, PMSE1, and PMSE2 of item selection methods under KI.
Figure 4.6. PAED, PMSE1, and PMSE2 of item selection methods under MI.
Figure 4.7. PAED, PMSE1, and PMSE2 of item selection methods under CEM.
First, the general performance of three test formats shows that polytomous items contribute more information to ability estimation than dichotomous items. The finding that the POLYTYPE test format is performing the best among test formats is one of proofs. Under the DPMIX test format in Figure 4.2, the estimation error drops dramatically when the test length is 18, and the slope becomes steeper after the 18th item. Under the PDMIX test format in Figure 4.3, the slopes become different when the test length is 7, and the slopes becomes flatter after the 7th item but the change is not as sudden as the DPMIX format. Note the first 18 items under the DPMIX format are dichotomous items, and the first 7 items under the PDMIX format are polytomous items. These two phenomena show that the polytomous items provide larger information than dichotomous items at the early or the middle stage of the MCAT. Under the DPMIX format, when starting to deliver polytomous items, the magnitude of shrinking estimation error becomes much larger and such a faster decrease stays similar for the rest of polytomous items because the slope keeps similarly steep in the polytomous part. It shows that polytomous items contribute more information to examinees' ability estimation than dichotomous items, even at the later stage of the test. Notice that the slope of the previous dichotomous part has already become flat right before the polytomous part. It means that dichotomous items have already located examinees' abilities to an accuracy level where dichotomous items could not improve too much for the rest of the test. The adoption of polytomous items under the DPMIX format considerably enhances the accuracy at the later stage.

Under the PDMIX format, when switching from polytomous items to dichotomous items, the magnitude of shrinking estimation error becomes smaller because the slopes become flatter since then. It means that dichotomous items do not locate examinees' abilities as accurate as polytomous items, even at the early stage of a test when ability estimation still has much space to
improve. However, the change of slope is not large, meaning that the following dichotomous items still do a good job. One benefit of the PDMIX format is that polytomous items detected relatively sufficient information at the beginning, which might help to select high quality dichotomous items in the following part. A relatively large portion of the test (75%) consists of dichotomous items. The early polytomous items make up the dichotomous item's weakness of providing less information. Therefore, the final results yield similar accuracy between the DPMIX and the PDMIX format.

Second, when using mix-type test formats, which item type, dichotomous or polytomous items, being delivered first depends on the setting of the test. This study finds that D-optimality, MI, CEM provide close performance under both DPMIX and PDMIX test formats at the final stage of the test. In this study, when the test length is set as 25, which item type being delivered first does not affect the final estimation considerably.

However, when the test length is short or when the test design is different, which item type being delivered first might matter. One of the observations is that at the final stage of the tests, all DPMIX formats either have similar estimation accuracy as all PDMIX formats or outperforms the PDMIX format. Based on figures and tables, D-optimality, KI, and MI under the DPMIX formats show smaller estimation error than their counterparts using PDMIX formats. For the CEM method, the performance of two formats are very similar. Thus, if the test length is longer than 25 or if the percentage of polytomous items is larger than 25%, it is possible that the DPMIX achieves higher estimation accuracy than the PDMIX format. For the DPMIX format, with decent information provided by administering dichotomous items in the first part of the test, polytomous items did a great job estimating ability even at the later stage of the test. The PDMIX format, however, shows better performance than the DPMIX format throughout the test except
the last two or three items. Under the PDMIX format, relatively more sufficient information than
the DPMIX format are detected by polytomous items at the beginning of the test. The
informative beginning could help to select higher quality dichotomous items in the following
stages and thus better locate examinees' abilities. However, if the test length is long enough or
the percentage of dichotomous items is high enough, the PDMIX format might not perform as
well as the DPMIX format. The observation that the PDMIX format outperforms the DPMIX
format except for the last couple items shows the possibility. Hence, if the test length is short or
if the percentage of dichotomous items is smaller, it is possible that the PDMIX format is a better
option. Note that at the initial stage, the PDMIX formats of certain item selections have larger
estimation error than the DPMIX format, such as the CEM method. However, it only occurs for
the first couple items and will not affect the whole estimation accuracy. The employment of
using different mixed test format in a short-length or long-length MCAT needs further
exploration in the future study.

The third finding is that the item bank size affects the performance of item selection
methods. For all four item selection methods, the POLYTYPE format outperforms the PDMIX
format throughout all test stages except the first couple items. At the early stage of these two
formats, polytomous items are administered. As discussed above in the section of estimation
accuracy among selection methods, the item bank size is one of reasons explaining why the
POLYTYPE format outperforms the PDMIX format at the early stage of the test.

The fourth discussion is whether D-optimality, MI and CEM perform similarly? First of
all, D-optimality has close performance with the other two methods. However, the D-optimality
performs slightly better throughout the test under each test format, which can be distinguished
from the figure presentation and data in tables. Second, MI and CEM have very close
performance with each other. When analyzing the comparison of MI and CEM under the POLYTYPE format or under the DPMIX format, their performance is difficult to tell in figures since it is hard for graphs to show small differences. From the tables, it shows that MI outperforms the CEM under the POLYTYPE and DPMIX and the differences in DPMIX are larger than the difference in POLYTYPE. However, when comparing CEM and MI under the PDMIX format, it is interesting to find that the MI shows better performance in terms of PAED, PMSE1, PMSE2 than the CEM throughout the test. The difference between the MI and the CEM could be distinguished from the figures as well. One of possible causes is the effect of the item bank size, which was discussed above. However, at the informative stage such as the middle stage of the DPMIX format, the CEM using polytomous items performs similarly to MI using polytomous items even the item bank size is still 100. It is possible that with decent information provided by previous part of the test, CEM and MI perform similarly even when the item bank size is not large. However, there might be other factors affecting the performance of item selection methods and item selection pattern, such as item pool structure. Further study needs to be conducted to explore possible factors influencing the performance of item selection methods.

The discussion regards comparing CEM and MI so far focused on the ability estimation accuracy. Whether these two methods select same items or not will be investigated in the section of the item selection pattern.

**Conditional Ability Estimation Accuracy**

The conditional ability estimation accuracy is examined both under the conditions of item selection method comparison and test format comparison. Figure 4.8, 4.9, and 4.10 are surface plots of four item selection methods under the POLYTYPE, DPMIX, and PDMIX test formats. Figure 4.11, 4.12, and 4.13 are contour plots of four item selection methods under the
POLYTYPE, DPMIX, and PDMIX test formats. For the surface plot, two coordinates on the bottom plane are sub-thetas in dimension one and two, $\theta_1$, $\theta_2$. The Z coordinate is the third-dimensional coordinate representing the PAED values. Therefore, the higher the surface is, the larger the estimation error of the corresponding two-dimensional theta points is. For the contour plot, each circle in the plot represents the same estimation accuracy level. In another word, all two-dimensional theta points on the same circle have the same PAED level. Three patterns of theta are compared in terms of conditional accuracy. Thetas in Pattern I are those with the same integer values in two dimensions, or $\theta_1 = \theta_2$. This pattern represents thetas with zero distance between two sub-thetas. Seven two-dimensional theta points are selected: (-3, -3), (-2, -2), (-1, -1), (0, 0), (1, 1), (2, 2), and (3, 3). Thetas in Pattern II are those having the same absolute integer values in both dimensions but with positive values in one dimension and negative values in the other dimension, or $\theta_1 = -\theta_2$. The second pattern refers to thetas with extreme values in both dimensions. Also, the distance between two sub-thetas is the largest for sub-thetas at the same extreme level. Six selected two-dimensional theta points include: (-3, 3), (3, -3), (-2, 2), (2, -2), (-1, 1), (1, -1). Thetas in Pattern III are those thetas with values in one dimension being zero and values in another dimension being integer values, which is $\theta_1 / \theta_2 = 0$ ($\theta_1 = 0$ or $\theta_2 = 0$). The third pattern has the medium position between Pattern I and Pattern II in terms of the distance between two sub-thetas. Here are twelve selected two-dimensional theta points in Pattern III: (-3, 0), (0, -3), (-2, 0), (0, -2), (-1, 0), (0, -1), (1, 0), (0, 1), (2, 0), (0, 2), (3, 0), and (0, 3).

Table 4.4 and 4.5 provide average conditional AED of Pattern I, II, and III at four stages. Table 4.6 presents average MSE1 and MSE2 of Pattern I, II, and III at final stage. Table 4.7 and 4.8 show conditional AED at selected theta points under Pattern I, II, and III at the initial, early, middle, and final stage with the POLYTYPE test format.
Table 4.4.

**Average Conditional AED of Three Test Formats at Initial and Early Stage**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>AED at Initial Stage</th>
<th>AED at Early Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1.857</td>
<td>1.638</td>
</tr>
<tr>
<td>II</td>
<td>2.336</td>
<td>2.787</td>
</tr>
<tr>
<td>III</td>
<td>1.598</td>
<td>1.630</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2.081</td>
<td>1.927</td>
</tr>
<tr>
<td>II</td>
<td>2.551</td>
<td>2.816</td>
</tr>
<tr>
<td>III</td>
<td>1.737</td>
<td>1.728</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1.795</td>
<td>1.672</td>
</tr>
<tr>
<td>II</td>
<td>2.494</td>
<td>2.786</td>
</tr>
<tr>
<td>III</td>
<td>1.616</td>
<td>1.620</td>
</tr>
</tbody>
</table>

Table 4.5.

**Average Conditional AED at Middle and Final Stage**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>PAED at Middle Stage</th>
<th>PAED at Final Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.723</td>
<td>0.641</td>
</tr>
<tr>
<td>II</td>
<td>1.051</td>
<td>2.197</td>
</tr>
<tr>
<td>III</td>
<td>0.679</td>
<td>1.089</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.990</td>
<td>1.298</td>
</tr>
</tbody>
</table>

* (table continues)*
Table 4.5. continued

<table>
<thead>
<tr>
<th>Pattern</th>
<th>PAED at Middle Stage</th>
<th>PAED at Final Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>1.448</td>
<td>2.045</td>
</tr>
<tr>
<td>III</td>
<td>0.904</td>
<td>1.331</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.875</td>
<td>0.917</td>
</tr>
<tr>
<td>II</td>
<td>1.271</td>
<td>1.997</td>
</tr>
<tr>
<td>III</td>
<td>0.823</td>
<td>1.214</td>
</tr>
</tbody>
</table>

Table 4.6.

*Average Conditional MSE1 and MSE2 at Final Stage*

<table>
<thead>
<tr>
<th>Pattern</th>
<th>MSE1 at Final Stage</th>
<th>MSE2 at Final Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>POLYTYPE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.259</td>
<td>0.228</td>
</tr>
<tr>
<td>II</td>
<td>0.535</td>
<td>2.433</td>
</tr>
<tr>
<td>III</td>
<td>0.228</td>
<td>0.551</td>
</tr>
<tr>
<td>DPMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.429</td>
<td>0.486</td>
</tr>
<tr>
<td>II</td>
<td>0.852</td>
<td>1.899</td>
</tr>
<tr>
<td>III</td>
<td>0.348</td>
<td>0.738</td>
</tr>
<tr>
<td>PDMIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0.435</td>
<td>0.537</td>
</tr>
<tr>
<td>II</td>
<td>0.833</td>
<td>1.901</td>
</tr>
<tr>
<td>III</td>
<td>0.356</td>
<td>0.818</td>
</tr>
</tbody>
</table>
Table 4.7.

*Conditional AED with POLYTYPE Format at Initial Stage (L=2) and Early Stage (L=7)*

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Initial Stage AED (L=2)</th>
<th>Early Stage AED (L=7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td><strong>Pattern I</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>-3</td>
<td>3.268</td>
<td>2.905</td>
</tr>
<tr>
<td>-2</td>
<td>-2</td>
<td>1.969</td>
<td>1.628</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>0.943</td>
<td>0.813</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.613</td>
<td>0.404</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.950</td>
<td>0.735</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.977</td>
<td>1.823</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3.277</td>
<td>3.161</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>1.857</td>
<td>1.638</td>
</tr>
<tr>
<td><strong>Pattern II</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>3</td>
<td>3.595</td>
<td>4.136</td>
</tr>
<tr>
<td>3</td>
<td>-3</td>
<td>3.579</td>
<td>4.146</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td>2.310</td>
<td>2.774</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>2.293</td>
<td>2.776</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>1.104</td>
<td>1.441</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1.136</td>
<td>1.453</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td>2.336</td>
<td>2.787</td>
</tr>
<tr>
<td><strong>Pattern III</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>0</td>
<td>2.447</td>
<td>2.384</td>
</tr>
<tr>
<td>0</td>
<td>-3</td>
<td>2.445</td>
<td>2.378</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>1.549</td>
<td>1.575</td>
</tr>
<tr>
<td>0</td>
<td>-2</td>
<td>1.557</td>
<td>1.596</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0.807</td>
<td>0.907</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>0.814</td>
<td>0.932</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.849</td>
<td>0.877</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 4.7. continued

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Initial Stage AED (L=2)</th>
<th>Early Stage AED (L=7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>Pattern III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.841</td>
<td>0.889</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.524</td>
<td>1.595</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1.536</td>
<td>1.596</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2.391</td>
<td>2.441</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>2.418</td>
<td>2.394</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>1.598</td>
<td>1.630</td>
</tr>
</tbody>
</table>

Table 4.8.

*Conditional AED with POLYTYPE Format at Middle Stage (L=18) and Final Stage (L=25)*

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Middle Stage (L=18)</th>
<th>Final Stage (L=25)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>Pattern I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>-3</td>
<td>1.203</td>
<td>1.057</td>
</tr>
<tr>
<td>-2</td>
<td>-2</td>
<td>0.655</td>
<td>0.516</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>0.482</td>
<td>0.404</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.460</td>
<td>0.436</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.480</td>
<td>0.383</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.644</td>
<td>0.538</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.137</td>
<td>1.152</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.723</td>
<td>0.641</td>
</tr>
<tr>
<td>Pattern II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>3</td>
<td>1.723</td>
<td>3.361</td>
</tr>
<tr>
<td>3</td>
<td>-3</td>
<td>1.681</td>
<td>3.3363</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
<td>0.936</td>
<td>2.208</td>
</tr>
</tbody>
</table>

*(table continues)*
<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>Middle Stage (L=18)</th>
<th>Final Stage (L=25)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt</td>
<td>KI</td>
</tr>
<tr>
<td>$\text{Pattern II}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>0.933</td>
<td>2.105</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>0.504</td>
<td>1.140</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0.529</td>
<td>1.034</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>1.051</td>
<td>2.197</td>
</tr>
<tr>
<td>$\text{Pattern III}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>0</td>
<td>0.944</td>
<td>1.413</td>
</tr>
<tr>
<td>0</td>
<td>-3</td>
<td>0.927</td>
<td>1.505</td>
</tr>
<tr>
<td>-2</td>
<td>0</td>
<td>0.646</td>
<td>0.978</td>
</tr>
<tr>
<td>0</td>
<td>-2</td>
<td>0.629</td>
<td>1.196</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0.501</td>
<td>0.628</td>
</tr>
<tr>
<td>0</td>
<td>-1</td>
<td>0.503</td>
<td>0.660</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.482</td>
<td>0.623</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.495</td>
<td>0.657</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.597</td>
<td>1.203</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.593</td>
<td>1.005</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.906</td>
<td>1.702</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0.925</td>
<td>1.506</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.679</td>
<td>1.089</td>
</tr>
</tbody>
</table>

Table 4.9 presents conditional MSE1, MSE2 at selected theta points under Pattern I, II, and III at the final stage under the POLYTYPE test format.
Table 4.9.

*Conditional MSE with POLYTYPE Test Format at Final Stage (L=25)*

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>MSE ($\theta_1$)</th>
<th>MSE ($\theta_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt  KI  MI  CEM</td>
<td>D-Opt  KI  MI  CEM</td>
</tr>
<tr>
<td>Pattern I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3  -3</td>
<td>0.535  0.424</td>
<td>0.525  0.534</td>
<td>0.467  0.367</td>
</tr>
<tr>
<td>-2  -2</td>
<td>0.189  0.160</td>
<td>0.183  0.205</td>
<td>0.204  0.122</td>
</tr>
<tr>
<td>-1  -1</td>
<td>0.133  0.126</td>
<td>0.134  0.130</td>
<td>0.117  0.125</td>
</tr>
<tr>
<td>0   0</td>
<td>0.124  0.146</td>
<td>0.111  0.121</td>
<td>0.111  0.158</td>
</tr>
<tr>
<td>1   1</td>
<td>0.125  0.113</td>
<td>0.127  0.122</td>
<td>0.123  0.119</td>
</tr>
<tr>
<td>2   2</td>
<td>0.179  0.183</td>
<td>0.195  0.177</td>
<td>0.199  0.123</td>
</tr>
<tr>
<td>3   3</td>
<td>0.529  0.447</td>
<td>0.521  0.539</td>
<td>0.439  0.461</td>
</tr>
<tr>
<td>Average</td>
<td>0.259  0.228</td>
<td>0.256  0.261</td>
<td>0.237  0.211</td>
</tr>
<tr>
<td>Pattern II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3  3</td>
<td>1.128  4.696</td>
<td>1.254  1.251</td>
<td>1.064  4.393</td>
</tr>
<tr>
<td>3  -3</td>
<td>1.070  5.116</td>
<td>1.163  1.249</td>
<td>1.020  4.950</td>
</tr>
<tr>
<td>-2  2</td>
<td>0.358  1.889</td>
<td>0.371  0.405</td>
<td>0.352  1.898</td>
</tr>
<tr>
<td>2  -2</td>
<td>0.375  1.929</td>
<td>0.391  0.420</td>
<td>0.355  1.835</td>
</tr>
<tr>
<td>-1  1</td>
<td>0.137  0.513</td>
<td>0.146  0.152</td>
<td>0.139  0.545</td>
</tr>
<tr>
<td>1  -1</td>
<td>0.142  0.455</td>
<td>0.134  0.135</td>
<td>0.127  0.475</td>
</tr>
<tr>
<td>Average</td>
<td>0.535  2.433</td>
<td>0.576  0.602</td>
<td>0.509  2.349</td>
</tr>
<tr>
<td>Pattern III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-3  0</td>
<td>0.553  1.220</td>
<td>0.525  0.654</td>
<td>0.171  0.497</td>
</tr>
<tr>
<td>0  -3</td>
<td>0.215  0.545</td>
<td>0.221  0.212</td>
<td>0.555  1.203</td>
</tr>
<tr>
<td>-2  0</td>
<td>0.244  0.452</td>
<td>0.236  0.236</td>
<td>0.156  0.355</td>
</tr>
<tr>
<td>0  -2</td>
<td>0.157  0.454</td>
<td>0.140  0.150</td>
<td>0.241  0.635</td>
</tr>
<tr>
<td>-1  0</td>
<td>0.130  0.205</td>
<td>0.139  0.130</td>
<td>0.126  0.237</td>
</tr>
<tr>
<td>0  -1</td>
<td>0.121  0.250</td>
<td>0.126  0.136</td>
<td>0.130  0.258</td>
</tr>
<tr>
<td>1  0</td>
<td>0.127  0.211</td>
<td>0.123  0.141</td>
<td>0.115  0.181</td>
</tr>
</tbody>
</table>

*(table continues)*
Table 4.9 continued

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>MSE ($\theta_1$)</th>
<th>MSE ($\theta_2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D-Opt KI MI CEM</td>
<td>D-Opt KI MI CEM</td>
</tr>
<tr>
<td>Pattern III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.122 0.258 0.125 0.141 0.135 0.235 0.115 0.128</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.222 0.646 0.241 0.226 0.133 0.443 0.142 0.167</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.138 0.447 0.149 0.152 0.224 0.492 0.221 0.195</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0.541 1.387 0.569 0.583 0.164 0.641 0.185 0.180</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0.166 0.532 0.161 0.198 0.512 1.004 0.507 0.574</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.228 0.551 0.229 0.246 0.222 0.515 0.248 0.235</td>
<td></td>
</tr>
</tbody>
</table>

First, estimation patterns of four item selection methods could be roughly concluded based on the surface plot. There are several common features shared by four item selection methods. One common pattern is that all four methods have similar locations of the surface bottom. For the surface plot, the lowest height of the surface occurs at the bottom of the surface plots. Because the lowest height refers to the smallest estimation error, the two-dimensional theta points corresponding to the location of the surface bottom have the highest estimation accuracy. For all four item selection methods under three test formats, the lowest heights of the surface are located in the central area of the theta plane, which is a small range surrounding the center, where $\theta_1 = 0$ and $\theta_2 = 0$. Based on data in the tables, it also proves that when thetas are equal to the center or close to the center of the theta place, ability estimation achieves the best precision. Another common feature is that for all four item selection methods under three test formats, the highest heights occurs at $(3, -3)$ and $(-3, 3)$. It shows that when theta in one dimension is an extreme value at either positive end or negative end, and when theta in the other dimension is an extreme value at the other end, the estimation error is the largest.

The estimation pattern also differs between the KI method and the rest three methods.
Under all three test formats, the surface plots of D-optimality, MI, and CEM methods have similar shapes. Although the values are different, these three methods show similar estimation patterns for two-dimensional theta points on the theta plane. The KI method, however, has a different surface plot, meaning that the KI method has a different estimation patterns for two-dimensional theta points. Under the POLYTYPE test format, the KI method performs the best for Pattern I defined above, or $\theta_1 = \theta_2$. In contrast, the other three methods increase estimation error when theta values become extreme at both the negative and positive ends for $\theta_1 = \theta_2$ on the theta plane. But the estimation error is not as large as the pattern of $\theta_1 = -\theta_2$.

![Surface plots of four item selection methods under the POLYTYPE test format.](image)

*Figure 4.8. The surface plot of four item selection methods under the POLYTYPE test format.*
Figure 4.9. The surface plot of four item selection methods under the DPMIX test format.

Figure 4.10. The surface plot of four item selection methods under the PDMIX test format.
For the contour plots, under all three test formats, all four item selection methods provide lower AED values in the middle of the theta coordinates. The less extreme both theta points are, the more accurate estimation yields. D-optimality, MI, and CEM have similar shapes and area of circles, confirming that they have similar patterns. KI, however, has relatively smaller region of the lower AED level, and the angle is right on the $\theta_1 = \theta_2$ pattern of the theta surface. It also confirms the previous finding that the KI method generally have more estimation error for patterns other than $\theta_1 = \theta_2$, but yields the smallest estimation error for the pattern of $\theta_1 = \theta_2$.

Figure 4.11. The contour plot of item selection methods under the POLYTYPE test format.
Figure 4.12. The contour plot of item selection methods under the DPMIX test format.

Figure 4.13. The contour plot of item selection methods under the PDMIX test format.
Conditional Comparison among Test Formats

There are common patterns in terms of conditional accuracy among all four methods when comparing different test formats. The POLYTYPE surface plots of all methods generally are flatter than DPMIX and PDMIX. Therefore, the POLYTYPE does a better job than the mix-type test formats to improve estimation precision for thetas which are not in the central area of theta surface, or thetas with relatively extreme values. Moreover, by using POLYTYPE, more two-dimensional theta values could be estimated at a relatively higher precision level since the area of flat and low space is larger. DPMIX and the PDMIX, however, do not provide such good performance. In addition, DPMIX and the PDMIX have very similar shape except the slight changes in values. Hence, which item type being delivered first does not considerably affect the conditional estimation accuracy.

When the comparison is based on the test formats, KI shows different performance among test formats. In contrast, D-optimality, MI, and CEM do not change the shape of the surface plots dramatically. It means that their estimation patterns do not change a lot when the mix-type test formats are used. However, the surface plot of KI under POLYTYPE is different from the plot under PDMIX and DPMIX. Under POLYTYPE, KI achieves the best precision level among four selection methods. The estimation error for the pattern of $\theta_1 = \theta_2$ is the lowest. Even at two ends of $\theta_1 = \theta_2$, the estimation error is very small. This finding does not apply to KI under DPMIX and PDMIX. Under DPMIX and PDMIX, KI increases estimation error when theta values become extreme for the pattern of $\theta_1 = \theta_2$. Since PDMIX and DPMIX administers dichotomous items, it is possible that KI has larger estimation error for the case of $\theta_1 = \theta_2$ for dichotomous items. It is an interesting direction to explore what factors affect KI's estimation accuracy. Possible factors include theta pattern, item type, item structure, or item bank size.
**Figure 4.14.** The surface plot of D-optimality among test formats.

**Figure 4.15.** The surface plot of KI among test formats.
Figure 4.16. The surface plot of MI among test formats.

Figure 4.17. The surface plot of CEM among test formats.
From the contour plots, the shape and area of lower AED regions of stays similar among three formats. The region of the lower AED of the POLYTYPE format is larger than others. It is reasonable to confirm that the POLYTYPE format yields better estimation performance for thetas with relatively extreme values in both dimensions and in both ends. It also confirms that which item type being delivered first does not greatly affect conditional estimation precision.
Figure 4.18. The contour plot of D-optimality among test formats.
Figure 4.19. The contour plot of KI among test formats.
Figure 4.20. The contour plot of MI among test formats.
Figure 4.21. The contour plot of CEM among test formats.
**Item Selection Pattern**

The item selection pattern is investigated under the POLYTYPE test format. In order to investigate the item selection pattern, the MDISC index and discrimination parameters in dimension one and dimension two are compared among all four item selection methods. Figure 4.22 presents the characteristic of selected items at each possible test length throughout the MCAT test for all four item selection methods. In Figure 4.22, there are three plots reporting the MDISC index and discrimination parameters in dimension one and dimension two respectively. The X coordinate of each plot is the test length from 1 to 25. The Y coordinate in the left plot is the MDISC index; the Y coordinate in the upper right plot is the discrimination parameters in dimension one; and the Y coordinate in the bottom right plot is the discrimination parameters in dimension two. Figure 4.23 is the bubble plot to investigate the item selection patterns. The X coordinate is discrimination parameters in dimension one and the Y coordinate is discrimination parameters in dimension two. Thus, each bubble corresponds to one item in the item bank. The area of each bubble is proportional to the administration times of this bubble-represented item. The larger the bubble is, the higher frequency the item is selected.
Figure 4.22. Discrimination characteristics of selected items under the POLYTYPE test format.
Figure 4.23. The bubble plot of four item selection methods.

Table 4.10 and 4.11 also provide the MDISC index and discrimination parameters in both dimensions at four stages of the test.

Table 4.10.

<table>
<thead>
<tr>
<th>Method</th>
<th>Initial Stage</th>
<th>Early Stage</th>
<th>Middle Stage</th>
<th>Final Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-opt</td>
<td>0.913</td>
<td>1.136</td>
<td>1.128</td>
<td>1.109</td>
</tr>
<tr>
<td>KI</td>
<td>1.645</td>
<td>1.553</td>
<td>1.347</td>
<td>1.272</td>
</tr>
<tr>
<td>MI</td>
<td>1.708</td>
<td>1.203</td>
<td>1.129</td>
<td>1.122</td>
</tr>
<tr>
<td>CEM</td>
<td>1.672</td>
<td>1.168</td>
<td>1.107</td>
<td>1.100</td>
</tr>
</tbody>
</table>
Table 4.11.

Discrimination Parameters in Dimension One and Two under the POLYTYPE Test Format

<table>
<thead>
<tr>
<th>Method</th>
<th>Discrimination in Dimension 1</th>
<th>Discrimination in Dimension 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial Early Middle Final</td>
<td>Initial Early Middle Final</td>
</tr>
<tr>
<td>D-opt</td>
<td>0.674 0.647 0.767 0.763</td>
<td>0.615 0.933 0.827 0.806</td>
</tr>
<tr>
<td>KI</td>
<td>1.157 1.083 0.933 0.906</td>
<td>1.169 1.113 0.971 0.892</td>
</tr>
<tr>
<td>MI</td>
<td>1.181 0.882 0.767 0.775</td>
<td>1.234 0.817 0.828 0.810</td>
</tr>
<tr>
<td>CEM</td>
<td>1.161 0.859 0.754 0.763</td>
<td>1.203 0.791 0.811 0.792</td>
</tr>
</tbody>
</table>

In general, the KI method favors items with high MDISC index and high discrimination parameters in both dimensions throughout the MCAT test. Both Figure 4.22 and 4.23 confirm the conclusion. D-optimality, CEM, and MI favor items with high discrimination parameters either in both dimension or in one of two dimensions. The bubble plot shows such tendency. In addition, D-optimality and MI tend to select items with similar MDISC index, especially at the later stage of the test. CEM also select items with similar MDISC index as D-optimality and MI. However, CEM selection shows slightly lower discrimination power than D-optimality and MI. KI, CEM, and MI selected items which almost have the highest MDISC index when test length is around 3 or 4. The D-optimality method, however, does not select items with such high discrimination power.

The item overlap rates between two selection methods at early stage, middle stage, and final stage of the POLYTYPE MCAT test are examined. The result is reported in Table 4.12.
Table 4.12.

The Item Overlap Rate among Item Selection Methods

<table>
<thead>
<tr>
<th>Test Length</th>
<th>D-opt vs. KI</th>
<th>D-opt vs. CEM</th>
<th>D-opt vs. MI</th>
<th>KI vs. CEM</th>
<th>KI vs. MI</th>
<th>CEM vs. MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0.193</td>
<td>0.501</td>
<td>0.504</td>
<td>0.315</td>
<td>0.309</td>
<td>0.531</td>
</tr>
<tr>
<td>18</td>
<td>0.341</td>
<td>0.652</td>
<td>0.686</td>
<td>0.413</td>
<td>0.406</td>
<td>0.644</td>
</tr>
<tr>
<td>25</td>
<td>0.398</td>
<td>0.707</td>
<td>0.727</td>
<td>0.455</td>
<td>0.441</td>
<td>0.702</td>
</tr>
</tbody>
</table>

The larger the overlap rate is, the higher the similarity among item selection method is.

First, three pairs of comparison involving the KI method, including D-optimality vs. KI, MI vs. KI, CEM vs. KI, has relatively lower overlap rate. This means that the KI method differs in item selection pattern than the other three methods. Second, D-optimality, CEM, and MI have relatively higher overlap rate among each other. Specifically, they tend to select common items at the early stage of the test since the overlap rates among all these methods in the early stage exceed 0.5.

Because CEM and MI show similar ability estimation, it is interesting to understand whether CEM and MI are similar in item selection pattern. Based on the discrimination power investigation and on the overlap rate examination, CEM and MI do not always select the same items although they do tend to select similar items. With the fact that CEM produces larger estimation error when the item bank size is smaller, it is likely that when the item bank is large enough, though the two methods do not select same items, the item parameters are similar with slight difference so that their estimation accuracy and item selection pattern is similar. When the item bank size is not sufficient enough, MI is affected by the limited item bank less than the
CEM method. One possibility is that discrimination power is one of main factors that affect the MI selection methods so that MI favors higher quality items even when the item bank has limited size of items. CEM, however, might favors items with other features besides the discrimination power. When the item bank size is limited, the selection is affected by certain other features as well. It will be a promising research direction in the future study.
CHAPTER 5
LIMITATIONS AND FUTURE STUDIES

Summary of Findings

This study compares four item selection methods, D-optimality, KI, MI, and CEM under three test formats, POLYTYPE, DPMIX, and PDMIX. In general, D-optimality presents the best estimation performance, followed by MI, CEM, and KI. The KI method, however, shows the smallest estimation error for the theta pattern of $\theta_1 = \theta_2$ although it shows larger estimation error than the other three methods under other theta patterns and all explored test formats. D-optimality, MI, and CEM have similar estimation and item selection pattern, while KI differs than them. Polytomous items provide more information than dichotomous items so that the POLYTYPE test format yields the best conditional estimation accuracy and two-dimensional theta points with relatively extreme values could be better probed under the POLYTYPE format. This finding applies to all four item selection methods. Which item type, dichotomous or polytomous items, being administered first depends on the test designs and settings. When test length is normal such as around 25 and when the proportion of polytomous and dichotomous items is similar to the design of this study, which item type being administered first does not affect the estimation precision or conditional estimation precision. If the test length is very long or if the test length is very short, or if the proportion of dichotomous and polytomous items is different from the current design, it is likely that one type of mixed test format is better than the other. Further investigation is needed to make a conclusion. Another finding is that item bank size affects the estimation precision. When item bank size shrinks, estimation errors become larger for all item selection methods. CEM is affected apparently by the change of item bank size.
Limitations and Future Studies

Studies in the MCAT field have not been extensively conducted and a variety of studies could be done in the field.

The first direction is to study the MCAT test where items and examinees have complex structure. In the current two-dimensional study, items are assumed to be loaded on both dimensions and two dimensions have no correlation with each other. In addition, examinees' abilities in two dimensions are assumed to be uncorrelated. However, in practice, two dimensions of items and two dimensions of examinees might be correlated. It is important to investigate the MCAT test by using real data or assuming a real test setting.

Second, item exposure control, content constraint, and other non-statistical factors are not explored in this study. To make the MCAT become applicable in real world, it is necessary to discuss the MCAT test considering factors such as item exposure control, content constraints, etc.

In addition, explore item selection methods in the test formats that contain several types of polytomous items either with different response categories or from different MIRT models. In the real world practice, there are a variety of polytomous items examining student abilities in different fields. It is common that one test consists of several different item types to assess students' comprehensive abilities. These polytomous items might fit different MIRT models or have different numbers of response categories. Besides MGPCM model, various polytomous MIRT models need to be explored. Therefore, in the future study, the POLYTYPE test format or the mix-type test format containing polytomous items with mixed numbers of categories or from various MIRT models should be discussed.

Moreover, item selection methods with variable length MCAT test is an interesting direction. Both MCAT test delivery method and polytomous items provide more information
than UCAT method and dichotomous items, it is possible that students' abilities could be diagnosed when the test length is shorter than UCAT test using dichotomous items. Furthermore, when the purpose is to diagnose or to improve instruction, a variable-length adaptive test has advantages with the adoption of certain item selection methods providing sufficient information. One advantage is that test length could be shortened and the item exposure rate is thus controlled. Therefore, the items could last longer in the pool before they are retired and students will get lower testing burden with shorter test time and length.

Another promising direction is to study item selection methods in MCAT test that has intentional and nuisance abilities and one composite ability as an explicit linear combination of theta. This study finds that it is likely item selection methods excel in one certain theta pattern but produces larger estimation error in another certain theta pattern. Therefore, for different case of multi-dimensional thetas, item selection methods might show a variety of performance patterns and accuracy levels. The further exploration is needed.

Item pool structure is one of factors affecting estimation accuracy. For example, MI and CEM provide similar performance under different test formats, while MI is slightly better than CEM. Under different item bank structures and different conditions, MI and CEM might have different performance. More explorations related to item pool structure need to be conducted.

Conclusion

MCAT will become one of main test delivery approaches in the future testing thanks to its diagnostic feature. To facilitate the development of formative assessments and testing for diagnosis, studies in MCAT are getting increasingly important. Similarly, polytomous items should be applied into the test so that tests could take full advantage of polytomous items' strength in providing diagnostic information and testing complicated abilities and skills. As a key
element of a MCAT test, item selection methods play an important role in improving ability
estimation precision and fulfilling the diagnostic purpose. D-optimality is found to provide the
best performance in general in this study. The other three methods, however, possess their unique
advantages in a variety of fields. Moreover, under different item pool structures and test designs,
the performance of item selection methods varies accordingly. Research in MCAT test using
polytomous items should be investigated profoundly both from theoretical perspective and for
practical application.
REFERENCES


