Item Selection Methods in Multidimensional Computerized Adaptive Testing With Polytomously Scored Items

Applied Psychological Measurement
2018, Vol. 42(8) 677–694
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618762748
journals.sagepub.com/home/apm

Dongbo Tu¹, Yuting Han^{1,2}, Yan Cai¹ and Xuliang Gao¹

Abstract

Multidimensional computerized adaptive testing (MCAT) has been developed over the past decades, and most of them can only deal with dichotomously scored items. However, polytomously scored items have been broadly used in a variety of tests for their advantages of providing more information and testing complicated abilities and skills. The purpose of this study is to discuss the item selection algorithms used in MCAT with polytomously scored items (PMCAT). Several promising item selection algorithms used in MCAT are extended to PMCAT, and two new item selection methods are proposed to improve the existing selection strategies. Two simulation studies are conducted to demonstrate the feasibility of the extended and proposed methods. The simulation results show that most of the extended item selection methods for PMCAT are feasible and the new proposed item selection methods perform well. Combined with the security of the pool, when two dimensions are considered (Study 1), the proposed modified continuous entropy method (MCEM) is the ideal of all in that it gains the lowest item exposure rate and has a relatively high accuracy. As for high dimensions (Study 2), results show that mutual information (MUI) and MCEM keep relatively high estimation accuracy, and the item exposure rates decrease as the correlation increases.

Keywords

multidimensional computerized adaptive test, multidimensional graded response model, Fisher information, Kullback–Leibler information, polytomous items

Introduction

Multidimensional computerized adaptive testing (MCAT) is based on the multidimensional item response theory (MIRT) and computerized adaptive testing (CAT). It can not only assess the examinees' multidimensional ability synchronously but also improve the measurement accuracy and test efficiency of the assessment (Luecht, 1996; Segall, 1996). Currently, almost

¹School of Psychology, Jiangxi Normal University, Nanchang, China

²School of Psychology, Beijing Normal University, Beijing, China

Corresponding Author:

Yuting Han, School of Psychology, Beijing Normal University, No. 19, Xijiekouwai St., Haidian District, Beijing 100875, P. R. China.

Email: hanyuting716@gmail.com

all of the MCAT studies only focus on the dichotomously scored data assessment (e.g., Mulder & van der Linden, 2009; Reckase, 2009; Segall, 1996; van der Linden, 1999; Wang & Chang, 2011; Wang, Chang, & Boughton, 2011). However, polytomously scored items have been broadly used in a variety of tests. In psychological inventories, it is common to use the Likert-type items which are usually scored according to the number of response categories allowed. Polytomous items are also highly applied in the achievement context, which could include some special item types, such as performance tasks, selected responses, and fill-in-the-blank. It is believed that polytomously scored items provide more information than dichotomously scored items (Donoghue, 1994; Samejima, 1976). It could not only measure concepts and skills in greater depth than dichotomous items but also reduce the test length while achieving the same effects, particularly under the CAT context (De Ayala, 1992).

Item selection strategy constitutes the key component of CAT, which not only makes tests adaptive but also exerts a direct impact on the quality of test. The algorithms of the item selection methods derived for multidimensional three-parameter logistic (M3PL) or two-parameter logistic (M2PL) model focuses on multidimensional dichotomous items other than multidimensional polytomous items. Lin (2012) compared D-optimality, Kullback–Leibler information index (KI), mutual information (MUI), and continuous entropy method (CEM) in MCAT adopting polytomously scored items under multidimensional generalized partial credit model (MGPCM; Yao & Schwarz, 2006). According to Lin (2012), test format completely delivering polytomous items yields the best estimation accuracy and D-optimality, MUI, and CEM have similar estimation and item selection pattern, while KI differs from them. Based on the work of Lin (2012), this article will further focus on the item selection strategy of the polytomously scored MCAT (PMCAT), attempting to (a) explore the proper means to extend more conventionally dichotomous MCAT selection algorithms to polytomous MCAT (PMCAT) and (b) propose two new item selection methods to improve the existing selection strategies for PMCAT.

The remainder of this paper is organized as follows: Firstly, the employed polytomously-scored model, the multidimensional graded response model (MGRM) are described. Secondly, some item selection methods of PMCAT extended from the methods of MCAT are introduced. Next, two new item selection methods of PMCAT considering the posterior distribution are proposed, followed by two Monte Carlo simulation studies, to explore the statistical properties of each strategy and the feasibility of PMCAT, and how the correlation among attributes affects the accuracy of high dimensional PMCAT. At last, the full paper is summarized, and the future research are discussed, which would lay foundations for the further research.

MGRM

In psychological inventories, it is common to use the Likert-type items which assume that the difficulty parameters increase or decrease sequentially. The graded response model (GRM) is suitable for the analysis of this kind of Likert-type data and easier for the user to understand, especially in CAT. There are many instances of CAT system for personality scale by using the GRM (Flens et al., 2017; Gardner et al., 2004; Smits, Cuijpers, & van Straten, 2011). Therefore, the two-parameter MGRM (Muraki & Carlson, 1995) is employed under the PMCAT in this study, which is generalized from the GRM. The logistic form of MGRM can be defined as,

$$P_{jt}^* = \frac{1}{1 + \exp \left[-D \left(\mathbf{a}_j^T \boldsymbol{\theta}_i - b_{jt} \right) \right]} \quad (1)$$

$$P_{jt}(\boldsymbol{\theta}_i) = P(u_j = t | \boldsymbol{\theta}_i) = P_{jt}^* - P_{j,t+1}^*. \quad (2)$$

wherein $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \dots, \boldsymbol{\theta}_{ip})^T$ denotes a set of p coordinates, and it is the ability parameter for examinee i ; $\mathbf{a}_j^T$ is a vector of discrimination parameters for item j ; b_{jt} is the t th threshold parameter for item j , which satisfies $b_{j1} < b_{j2} < \dots < b_{jmf_j}$; mf_j represents the maximum score of item j . Furthermore, P_{jt}^* expresses the cumulative probability of examinee i , getting at least a score t on item j , and $P_{jt}(\boldsymbol{\theta}_i)$ is the probability of examinee i , responding to item j in a specific category score t . In addition, it assumes that $P_{j0}^* = 1$ and $P_{j,mf_j+1}^* = 0$.

For the convenience of presentation, the following notations will be used throughout this article. Assume there are N items in the item pool and n items in an adaptive test; $l \in \{1, \dots, p\}$ denote the component of ability vector $\boldsymbol{\theta}$; j_k denotes the item in the pool administered as the k th item in the test; S_{k-1} denotes the set of first $k-1$ administered items; $\mathbf{U}_{k-1}$ is the item response matrix of $k-1$ administered items; R represents the item pool; and R_k represents the set of remaining items in the pool after the $(k-1)$ th item is administered which equals to the set $R - S_{k-1}$.

The Extension of Item Selection Methods of MCAT to PMCAT

Under the framework of MCAT, some popular item selection methods have been proposed (e.g., Chang & Ying, 1996; Mulder & van der Linden, 2009, 2010; Segall, 1996; Veldkamp & van der Linden, 2002). From the profile of the information statistics, these methods could be classified into three types: One is the method based on Fisher information (FI), and the other two are based on Kullback–Leibler information (KL-based methods), which are the methods based on traditional KL information and the methods based on KL information between posteriors. As there are similarities and differences between MCAT and PMCAT, it attempts to investigate the feasibility of some of these methods under the framework of PMCAT. The details are discussed in the following.

The Extension of FI-Based Methods

Comparison of methods based on FI under MCAT can be found in Appendix A in the online supporting information. Under the framework of PMCAT, with MGRM model employed, the FI matrix for item j is defined as,

$$I_j^*(\boldsymbol{\theta}) = \sum_{t=0}^{mf_j} (P_{jt}^* - P_{j,t+1}^*) (1 - P_{jt}^* - P_{j,t+1}^*)^2 \begin{bmatrix} a_{j1}^2 & a_{j1}a_{j2} & \cdots & a_{j1}a_{jp} \\ a_{j1}a_{j2} & a_{j2}^2 & \cdots & a_{j2}a_{jp} \\ \vdots & \vdots & \ddots & \vdots \\ a_{j1}a_{jp} & a_{j2}a_{jp} & \cdots & a_{jp}^2 \end{bmatrix}. \quad (3)$$

where P_{jt}^* is defined by (1). Under the conditional independence assumption, the FI of a set of S_{k-1} items could be computed by

$$I_{S_{k-1}}^*(\boldsymbol{\theta}) = \sum_{j \in S_{k-1}} I_j^*(\boldsymbol{\theta}). \quad (4)$$

According to (4), the six methods could be expressed as follows.

D-optimality and Bayesian D-optimality methods. D-optimality method is to select the next item that maximizes the determinant of (4), which is to yield the smallest confidence region for the ability parameters, and its expression is

$$j_k = \arg \max_{j \in R_k} \left\{ \det \left[I_{S_{k-1}}^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right] \right\}. \quad (5)$$

With the consideration of the prior distribution of $\boldsymbol{\theta}$, the Bayesian D-optimality method is obtained:

$$j_k = \arg \max_{j \in R_k} \left\{ \det \left[I_{S_{k-1}} \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + \sum_0^{-1} \right] \right\}, \quad (6)$$

where $\hat{\boldsymbol{\theta}}_{k-1}$ is the ability estimator after first $k - 1$ items are administered, and Σ_0 is the prior covariance matrix of $\boldsymbol{\theta}$ (Segall, 1996).

A-optimality and bayesian A-optimality methods. A-optimality method is to select the item that minimizes the trace of the inverse of (4), which is to minimize the sum of the (asymptotic) sampling variances of the MLEs of the abilities. Its expression is

$$j_k = \arg \min_{j \in R_k} \left\{ \text{trace} \left[\left(I_{S_{k-1}}^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right)^{-1} \right] \right\} = \arg \max_{j \in R_k} \left\{ \frac{\det \left[I_{S_{k-1}}^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right]}{\sum_{l=1}^p \det \left(\left[I_{S_{k-1}}^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right]_{[l,l]} \right)} \right\}, \quad (7)$$

where $\left[I_{S_{k-1}} \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right]_{[l,l]}$ is the submatrix of $\left[I_{S_{k-1}} \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right]$ when omitting the l th row and column. Similarly, by adding Σ_0 to (7), the Bayesian A-optimality is obtained (Mulder & van der Linden, 2009).

E-optimality and bayesian E-optimality methods. The E-optimality method is to maximize the smallest eigenvalue of (4), which is equivalent to minimize the generalized variance of the ability estimators along their largest dimension; its expression is

$$i_k = \arg \max_{j \in R_k} \left\{ \min \left\{ \text{eigen} \left[I_{S_{k-1}}^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) + I_j^* \left(\hat{\boldsymbol{\theta}}_{k-1} \right) \right] \right\} \right\}. \quad (8)$$

By adding Σ_0 to (8), the Bayesian E-optimality is obtained (Mulder & van der Linden, 2009).

The Extension of KL-Based Methods

Four popular methods of MCAT based on KL information are discussed here (Chang & Ying, 1996; Mulder & van der Linden, 2010; Veldkamp & van der Linden, 2002; Wang & Chang, 2011), that is, KL index (KI) method, posterior expected KL information method (K^B), method of KL information between subsequent posteriors (KLP), and MUI. To be available in PMCAT, these methods will be modified.

Methods based on traditional KL information. For item j , the KL information under MGRM model is defined as

$$KL_j^*(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta}) = \sum_{t=0}^{mf_j} P_{jt}(\boldsymbol{\theta}) \log \left[\frac{P_{jt}(\boldsymbol{\theta})}{P_{jt}(\hat{\boldsymbol{\theta}})} \right]. \quad (9)$$

where $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ are estimated and unknown ability parameters, respectively, and $P_{jt}(\boldsymbol{\theta})$ is defined by (2).

Thus, the KI and K^B method are expressed as Equations 10 and 11, respectively.

$$j_k = \arg \max_{j \in R_k} \{ KL_j(\hat{\boldsymbol{\theta}}_{k-1}) \} = \arg \max_{j \in R_k} \left\{ \int_{\hat{\boldsymbol{\theta}}_{k-1} - \delta_{k-1}}^{\hat{\boldsymbol{\theta}}_{k-1} + \delta_{k-1}} KL_j^*(\hat{\boldsymbol{\theta}}_{k-1}||\boldsymbol{\theta}) \partial \boldsymbol{\theta} \right\}. \quad (10)$$

$$\begin{aligned} j_k &= \arg \max_{j \in R_k} K_j^B(\hat{\boldsymbol{\theta}}_{k-1}) = \arg \max_{j \in R_k} \int_{\boldsymbol{\theta}} KL_j^*(\hat{\boldsymbol{\theta}}_{k-1}||\boldsymbol{\theta}) \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1}) \partial \boldsymbol{\theta} \\ &= \arg \max_{j \in R_k} \int_{\boldsymbol{\theta}} \left\{ \sum_{u_{jk}=0}^{mf_j} P_j(u_{jk}|\hat{\boldsymbol{\theta}}_{k-1}) \log \left[\frac{P_j(u_{jk}|\hat{\boldsymbol{\theta}}_{k-1})}{P_j(u_{jk}|\hat{\boldsymbol{\theta}})} \right] \right\} \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1}) \partial \boldsymbol{\theta}, \end{aligned} \quad (11)$$

wherein $\delta_{k-1} = d/\sqrt{k-1}$ determining the size of the region over which the average is computed. $k-1$ refers to the number of items which have been administered, and d usually set to be 3, corresponding to the range of ability (Chang & Ying, 1996; Veldkamp & van der Linden, 2002). $\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})$ is the posterior distribution of $\boldsymbol{\theta}$ after administering $k-1$ items, $P_j(u_{jk}|\hat{\boldsymbol{\theta}})$ is the predicted probability when selecting the item j as the k th item of the test. From (11), it could be known that K^B method is a Bayesian version of the KI method by considering the posterior distribution of $\boldsymbol{\theta}$ (Veldkamp & van der Linden, 2002).

Methods based on KL information between posteriors. In CAT, the items which yield the largest change of the posterior distribution should be selected. The KL information can be used to formalize this argument. These methods differ profoundly from the traditional KL methods in that the former uses KL information to measure the divergences between the two consecutive posteriors of $\boldsymbol{\theta}$, whereas the latter uses KL information to measure the divergence between the response distributions generated by two different ability levels.

Method of KL information between subsequent posterior distributions. The KL information between subsequent posterior distributions $\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})$ and $\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})$ is equal to

$$KL(\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})||\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})) = \sum_{t=0}^{mf_j} \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1}) \log \left[\frac{\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})}{\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})} \right]. \quad (12)$$

Then, the KLP method, noted as K^P , is to select the next item by

$$j_k = \arg \max_{j \in R_k} K_j^P = \arg \max_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} \int_{\boldsymbol{\theta}} P(u_{jk}|\mathbf{U}_{k-1}) KL(\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})||\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})) \partial \boldsymbol{\theta} \quad (13)$$

MUI method. In PMCAT, with the analogy to the work of Mulder and van der Linden (2010) in MCAT, the MUI is defined as the KL information between the joint distribution and the product of the marginal distribution. Thus, the polytomous MUI method could be expressed as

$$\begin{aligned}
j_k &= \arg \max_{j \in R_k} \text{MUI}(\boldsymbol{\theta}; u_{jk}) = \arg \max_{j \in R_k} \int_{\boldsymbol{\theta}} KL[P(\boldsymbol{\theta}; u_{jk} | \mathbf{U}_{k-1}) || P(u_{jk} | \mathbf{U}_{k-1}) \pi_{k-1}(\boldsymbol{\theta} | \mathbf{U}_{k-1})] \partial \boldsymbol{\theta} \\
&= \arg \max_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}; u_{jk} | u_{k-1}) \log \frac{P(\boldsymbol{\theta}; u_{jk} | \mathbf{U}_{k-1})}{P(u_{jk} | \mathbf{U}_{k-1}) \pi_{k-1}(\boldsymbol{\theta} | \mathbf{U}_{k-1})} \partial \boldsymbol{\theta} \\
&= \arg \max_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}; u_{jk} | \mathbf{U}_{k-1}) \log \frac{P(u_{jk} | \boldsymbol{\theta})}{P(u_{jk} | \mathbf{U}_{k-1})} \partial \boldsymbol{\theta}
\end{aligned} \tag{14}$$

Furthermore, MUI is equivalent to the average KL information between the new and current posteriors $KL(\pi_k(\boldsymbol{\theta} | \mathbf{U}_{k-1}, u_{jk}) || \pi_{k-1}(\boldsymbol{\theta} | \mathbf{U}_{k-1}))$.

Two New Item Selection Methods of PMCAT

In the section “The Extension of Item Selection Methods of MCAT to PMCAT” based on the similarity between MCAT and PMCAT, some selection methods of MCAT to PMCAT have been extended. However, during the work of extension, some information could be incorporated into some of the former methods. Hence, two new methods are proposed in this segment, that is, the modified method of posterior expected KL information (simplified as MK^B method) and the modified method of CEM (simplified as modified continuous entropy method [MCEM] method), and the details are presented in the following.

MK^B Method

In K^B method, the criterion is computed based on the current estimate $\hat{\boldsymbol{\theta}}_{k-1}$ under the assumption that the point estimate is a good summary of the posterior distribution. It is also not a bad idea of considering the entire posterior distribution to substitute for the point estimate. The modified K^B (MK^B) method could be defined as

$$\begin{aligned}
\arg \max_{j \in R_k} MK_j^B &= \arg \max_{j \in R_k} \int_{\boldsymbol{\theta}_d} \left(\int_{\boldsymbol{\theta}_c} KL_j(\boldsymbol{\theta}_d || \boldsymbol{\theta}_c) \pi_{k-1}(\boldsymbol{\theta}_c | \mathbf{U}_{k-1}) \partial \boldsymbol{\theta}_c \right) \pi_{k-1}(\boldsymbol{\theta}_d | \mathbf{U}_{k-1}) \partial \boldsymbol{\theta}_d \\
&= \arg \max_{j \in R_k} \int_{\boldsymbol{\theta}_d} \left(\int_{\boldsymbol{\theta}_c} \left\{ \sum_{u_{jk}=0}^{mf_j} P_j(u_{jk} | \boldsymbol{\theta}_d) \log \left[\frac{P_j(u_{jk} | \boldsymbol{\theta}_d)}{P_j(u_{jk} | \boldsymbol{\theta}_c)} \right] \right\} \pi_{k-1}(\boldsymbol{\theta}_c | \mathbf{U}_{k-1}) \partial \boldsymbol{\theta}_c \right) \pi_{k-1}(\boldsymbol{\theta}_d | \mathbf{U}_{k-1}) \partial \boldsymbol{\theta}_d,
\end{aligned} \tag{15}$$

where $\boldsymbol{\theta}_c$ and $\boldsymbol{\theta}_d$ are integration variables that represent the true ability of examinee.

Compared with the K^B method, the numerator in Equation 14 considers all the possible ability vectors, and weights them accordingly. Because it does not require estimating the ability vector $\hat{\boldsymbol{\theta}}$, which may not be accurately described at the early stage of testing, and it might show more informative than the K^B method. Its superiority still needs to be verified.

MCEM Method

Suppose the random variable X follows a continuous distribution, the continuous entropy is defined as the Shannon entropy $H(X)$, which is computed by

$$H(X) = \int f(x) \log \frac{1}{f(x)} dx. \quad (16)$$

Then, the posterior continuous entropy after $k - 1$ items are administered is equal to

$$H(\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})) = \int_{\boldsymbol{\theta}} \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1}) \log \left[\frac{1}{\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})} \right] \partial \boldsymbol{\theta}. \quad (17)$$

For MCAT, Wang and Chang (2011) suggested a CEM to select the next item that has the smallest expected posterior continuous entropy; the criterion could be expressed as

$$\begin{aligned} j_k &= \arg \min_{j \in R_k} E_j \{ H[\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})] \} = \arg \min_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} H[\pi_k(\boldsymbol{\theta}|u_{jk}, \mathbf{U}_{k-1})] P(u_{jk}|\mathbf{U}_{k-1}) \\ &= \arg \min_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} \left\{ \int_{\boldsymbol{\theta}} \pi_k(\boldsymbol{\theta}|u_{jk}, \mathbf{U}_{k-1}) \log \left[\frac{1}{\pi_k(\boldsymbol{\theta}|u_{jk}, \mathbf{U}_{k-1})} \right] \partial \boldsymbol{\theta} \right\} \left[\int_{\boldsymbol{\theta}} P(u_{jk}|\boldsymbol{\theta}) \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1}) \partial \boldsymbol{\theta} \right]. \end{aligned} \quad (18)$$

When polytomous data are analyzed, and the MGRM model is used in Equation 18, the CEM method for PMCAT is obtained. In the method, the posterior distribution of $\boldsymbol{\theta}$ is

$$\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk}) = \frac{g(\boldsymbol{\theta})L(\mathbf{U}_{k-1}, u_{jk}|\boldsymbol{\theta})}{\int g(\boldsymbol{\theta})L(\mathbf{U}_{k-1}, u_{jk}|\boldsymbol{\theta}) \partial \boldsymbol{\theta}}. \quad (19)$$

In standard Bayesian methods, the prior distribution $g(\boldsymbol{\theta})$ is fixed before any data are observed. In contrast, in empirical Bayesian methods, $g(\boldsymbol{\theta})$ is estimated from the data (Robbins, 1985). Using the idea of empirical Bayes methods for reference, the prior distribution $g(\boldsymbol{\theta})$ in Equation 19 is replaced by the current posterior distribution $\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})$, which is updated after each item administered and uses all the information provided by the responses. Thus, the modified posterior distribution is expressed by

$$\pi_k'(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk}) = \frac{\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})L(\mathbf{U}_{k-1}, u_{jk}|\boldsymbol{\theta})}{\int \pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})L(\mathbf{U}_{k-1}, u_{jk}|\boldsymbol{\theta}) \partial \boldsymbol{\theta}}. \quad (20)$$

When $\pi_k'(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})$ is used in Equation 18 instead of $\pi_k(\boldsymbol{\theta}|u_{jk}, \mathbf{U}_{k-1})$, the MCEM method is obtained, which is modified as

$$j_k = \arg \min_{j \in R_k} \text{MCEM} = \arg \min_{j \in R_k} \sum_{u_{jk}=0}^{mf_j} H[\pi_k'(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})] P(u_{jk}|\mathbf{U}_{k-1}). \quad (21)$$

By using the updated prior distribution, the MCEM method could extract more information from the responses; therefore, it can be expected to lead a more accurate estimate than CEM at least in the early stage of PMCAT intuitively.

In addition, according to Wang and Chang (2011), the item selection rule of MUI could be rewritten as the difference between two entropy measures:

$$\arg \max_{j \in R_k} \langle E_{jk} \{ H[\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})] - H[\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})] \} \rangle. \quad (22)$$

It should be noticed that the first part of this equation $H[\pi_{k-1}(\boldsymbol{\theta}|\mathbf{U}_{k-1})]$ is invariant for any item in R_k , so the selection of the k th item is only decided by the second term $H[\pi_k(\boldsymbol{\theta}|\mathbf{U}_{k-1}, u_{jk})]$ in the equation, which is exactly the expected posterior continuous entropy after administering k th. Therefore, mathematically, MUI and CEM will choose the same items in MCAT. Thus, in this article, CEM will not be discussed.

Simulation Studies

Simulation studies are carried out in this research for two important purposes: (a) to detect the statistical properties of the methods proposed in sections “The Extension of Item Selection Methods of MCAT to PMCAT” and “Two New Item Selection Methods of PMCAT” and to demonstrate the feasibility of PMCAT; (b) to verify the performance of the polytomous item selection methods in high dimensional context, and to find out the influence of correlation among multiple abilities on the accuracy and security of PMCAT. To show support to both of these purposes, two separate simulation studies would be implemented.

Simulation Study I

Simulation conditions. A Monte Carlo simulation study is conducted to compare the performance of the 12 selection strategies, that is, D-optimality, A-optimality, E-optimality, and their Bayesian versions, KI, K^B , KLP, MUI, MK^B , and MCEM method. Among these criteria, MK^B and MCEM method are first introduced in this study.

In the first simulation study, the item pool consists of 450 items following MGRM. Suggested Wang and Chang (2011) and van der Linden (1999) and Lin (2012), the item discrimination parameters a_{j1} and a_{j2} are drawn from $U(0,1.3)$. Items with four response categories are explored for each item. To ensure that the threshold parameters of polytomous items cover the range of $\mathbf{a}_j^T \boldsymbol{\theta}_i$, b_{j1} , b_{j2} , and b_{j3} are drawn from $N(0, 2)$, and ordered as $b_{j1} < b_{j2} < b_{j3}$. Similar to previous MCAT studies (Finkelman, Nering, & Roussos, 2009; van der Linden, 1999; Wang & Chang, 2011), examinee responses are simulated with true abilities on a two-dimensional grid spanning the square $\theta_1, \theta_2 = -3.0 - 2.5 - 2.0, \dots, 3.0$. Crossing 13 discrete points on each of two dimensions generates a grid of 169 vector points. To balance the random error, 100 replications are run at each $\boldsymbol{\theta}$ point. Therefore, 16,900 replications are conducted, which is equal to 100 times multiply by 169 $\boldsymbol{\theta}$ points. The descriptive statistics of the item bank are shown in Table 1.

This is a fixed-length MCAT and the test length is set to 25. The first item is chosen by random. It is expected that a posteriori (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. In our examples, the Gauss–Hermite numerical integration formulas from Glas (1992) will be used and the integration is taken over the range of ability $[-3, 3]$.

Evaluation criteria. Euclidean distance is taken as a global index of psychometric precision when the number of dimension is 2 (Finkelman et al., 2009; Segall, 1996; Wang & Chang, 2011),

$$ED_i = \left[(\theta_{i1} - \hat{\theta}_{i1})^2 + (\theta_{i2} - \hat{\theta}_{i2})^2 \right]^{\frac{1}{2}}. \quad (23)$$

To measure the estimation accuracy of item selection methods at each ability point, the average Euclidean distance (AED) is calculated. And a prior weighted AED (PAED, Finkelman

Table 1. Descriptive Statistics of the Item Banks ($M = 450$) for Study I.

	a_1	a_2	b_1	b_2	b_3
M	0.6863	0.6388	-2.8645	0.1874	3.0191
SD	0.3683	0.3674	2.3827	2.7374	2.3937

et al., 2009; Wang & Chang, 2011) is adopted as the evaluation criterion of the overall performance of each item selection method under different test lengths

$$PAED = \sum_{q=1}^{169} AED(\boldsymbol{\theta}_q)A(\boldsymbol{\theta}_q), \quad (24)$$

wherein $\boldsymbol{\theta}_q$ represents the 169 ability vector points, $AED(\boldsymbol{\theta}_q)$ represents the AED among 100 replications on $\boldsymbol{\theta}_q$, and $A(\boldsymbol{\theta}_q)$ refers to the weight of $\boldsymbol{\theta}_q$ in the standard bivariate normal distribution, $\sum_{q=1}^{169} A(\boldsymbol{\theta}_q) = 1$.

The exposure rate (ER) and test overlap ration (TOR) index are used to measure the security of the item pool. $ER_j = f_j/N$. ER_j is the exposure rate of item j . f_j is the number of times that item j is selected. The smaller ER_j is, the lower the exposure rate is. The chi-square statistic is used to reflect the overall exposure of the item bank as

$$\chi^2 = \sum_{j=1}^M \frac{[ER_j - E(ER_j)]^2}{E(ER_j)}, \quad (25)$$

wherein $E(ER_j) = L/M$ is the expected exposure rate of item j , L represents the test length and M is number of items in item pool (Chang & Ying, 1999). The χ^2 index reflects the difference between the observed item exposure rate and expected exposure rate. The smaller the χ^2 is, the more evenly the calls of items are, which means that the item pool is more safety.

The test overlap ration (TOR) represents average between-test overlap, which is the arithmetic mean of the between-test overlaps across all possible pairwise comparisons. Therefore, the calculation of TOR is related to the item exposure rate, test length, and number of subjects. Chen, Ankenmann, and Spray (2003) gave its calculation formula as follows:

$$\bar{T} = \frac{N \times \sum_{j=1}^M (ER_j)^2}{(N-1) \times L} - \frac{1}{N-1}. \quad (26)$$

To compare the item selection pattern among item selection methods, the item discriminations throughout the test progression will also be recorded.

Results of Study I

Ability estimation accuracy. Table 2 shows the overall estimation accuracy (PAED) of each method against test length. As anticipated, the PAED decreases as the test length increases. Except KI method, all the proposed methods achieve relatively high estimate accuracy, which demonstrates their applicability for PMCAT. In details: (1) For the FI-based methods, A-optimality method earns the smallest PAED. (2) There are great differences between the KI method and the other two traditional KL-information-based methods. The KI method obviously

Table 2. Estimation Accuracy (PAED) of Each Method Against Test Length for Study I.

Test length	FI-based methods					Methods based on traditional KL information				Methods based on KL information between posteriors			
	A	D	E	Bayes		KI	K^B	MK^B	KLP	MUI	MCEM		
				-A	-D								
1	1.164	1.170	1.165	1.166	1.173	1.165	1.163	1.172	1.162	1.167	1.163		
3	0.885	0.866	0.918	0.906	0.950	0.917	0.879	0.884	0.882	0.887	0.875		
5	0.737	0.728	0.762	0.731	0.792	0.764	0.724	0.730	0.729	0.724	0.722		
7	0.644	0.639	0.670	0.638	0.684	0.666	0.648	0.647	0.646	0.643	0.645		
9	0.580	0.579	0.599	0.578	0.614	0.598	0.586	0.587	0.582	0.581	0.580		
11	0.531	0.532	0.547	0.528	0.561	0.552	0.537	0.540	0.536	0.533	0.532		
13	0.492	0.494	0.504	0.490	0.520	0.507	0.499	0.499	0.501	0.496	0.494		
15	0.459	0.461	0.473	0.458	0.487	0.472	0.467	0.467	0.466	0.465	0.463		
17	0.434	0.439	0.448	0.436	0.455	0.447	0.441	0.442	0.437	0.439	0.439		
19	0.414	0.417	0.427	0.413	0.433	0.425	0.421	0.424	0.418	0.420	0.417		
21	0.394	0.400	0.405	0.394	0.415	0.406	0.403	0.405	0.401	0.405	0.399		
23	0.380	0.386	0.392	0.380	0.397	0.390	0.387	0.388	0.385	0.389	0.383		
25	0.365	0.372	0.377	0.366	0.381	0.376	0.373	0.374	0.372	0.376	0.370		

Note. PAED = prior weighted AED; AED = average Euclidean distance. KL = Kullback–Leibler; KLP = KL information between subsequent posteriors; MUI = mutual information.

performs the worst, while it is hard to distinguish the PAED between K^B and MK^B method, with the former being slightly better than the latter. (3) As to the methods based on KL information between posteriors, MCEM method generates a little smaller PAED than the other two.

Conditional ability estimation accuracy. To show the estimation accuracy of the methods at each ability point, the Euclidean distance defined in Equation 23 is averaged over the 100 repeat subjects.

Estimation patterns of 12 item selection methods could be roughly concluded based on the surface and contour plots of the AED, seen in Figure 1 and Figure 1 of Appendix B in the online supporting information. For the surface plot, the lowest height refers to the smallest estimation error. Figure 1 shows that all methods yield lower height on the diagonal $\theta_1 = \theta_2$ than the other values. It means that the estimated accuracy is better when the two dimensions of θ are similar. This discrepancy differs in each method, while the surface of E-optimality and Bayesian E-optimality methods are relatively even, and the surface of KI method is more sloping and has distinctively lower AED at the $\theta_1 = \theta_2$ diagonal. The highest heights occur at (3, -3) and (-3, 3). It shows that the estimation error is the largest when both dimensions of theta are extreme values at different ends.

It could be found from the contour plot (Figure 1 of Appendix B) that all methods provide lower AED values in the middle of the theta coordinates. Furthermore, the regions of the lower AED level of E-optimality and Bayesian E-optimality methods are more concentrated in the center, while the shapes of other methods tend to have an angle on the $\theta_1 = \theta_2$ pattern of the theta surface.

The exposure rate. Exposure control is an important aspect of adaptive tests, especially for high-stake tests. The following indices are summarized in Table 3 to quantify the equalization of exposure rates: (a) the quartiles of the item exposure rates; (b) the number of overexposed items, which have exposure rates larger than 0.2; (c) the χ^2 statistic; and (d) the test overlap ration (TOR) index.

It can be easily inferred from Table 3 that the MCEM method has the lowest item exposure rates and gains the least number of overexposed items without sacrificing estimation accuracy. In addition, the distribution of item utilization rates produced by the MCEM method is a little more even, compared with the distribution of item utilization rates generated by other methods. The largest item exposure rate when the A-optimally method is applied is as high as 0.972 compared with the largest item exposure rate of 0.802 when the MCEM is applied. In short, the item exposure rates are relatively less evenly distributed when the methods based on FI matrix are applied compared with that when the methods based on traditional KL information and methods based on the KL information between posteriors are applied.

Item selection pattern. The bubble plot to investigate the item selection patterns can be found in Figure 2 of Appendix B in the online supporting information. In general, the items with low discrimination parameters in both dimensions are barely selected. E-optimality and Bayes E-optimality methods intend to select the items with large discrimination parameter in one dimension while the discrimination parameter in the other dimension is relatively small (see the green shadow). A-optimality and Bayes A-optimality also have this intendancy. KI method prefers the items with high discrimination parameters in both dimensions (see the red shadow); however, it cannot guarantee the best estimate accuracy. The other item selection methods favor the items with high discrimination parameters in either dimension or one of two dimensions. The investigation of item selection patterns can help us generate more simplified item selection methods based on item parameters. Ensuring a high degree of discrimination parameters in one dimension, and selecting the items with large difference between the discrimination parameters in the two dimensions may get good estimation accuracy. Combined with Figure 1, the items with high

Figure 1. Conditional AED for each method with surface plot for Study 1.
Note. AED = average Euclidean distance; KI = Kullback–Leibler information index; KLP = KL information between subsequent posteriors; MUI = mutual information.

discrimination parameters in both dimensions, seen in KI method and E-optimality, might influence the conditional ability estimation accuracy at the $\theta_1 = \theta_2$ diagonal.

Simulation Study 2

To verify the performance of the polytomous item selection methods in high dimensional context, we conducted the second study with a four-dimensional model. Study 2 also intends to investigate the influence of correlation among multiple abilities on the accuracy and security of PMCAT in high dimension.

Table 3. Exposure rate ($L = 25$) for Study I.

Methods	Select Method	Exposure rate						TOR	χ^2
		Minimum	Q1	Q2	Q3	Maximum	NO		
Based on FI	A	0.001	0.002	0.002	0.005	0.972	53	0.391	150.803
	D	0.001	0.002	0.002	0.011	0.966	53	0.343	129.148
	E	0.001	0.002	0.002	0.003	0.955	51	0.474	188.378
	Bayes-A	0.001	0.002	0.002	0.005	0.878	54	0.372	142.191
	Bayes-D	0.001	0.002	0.002	0.013	0.807	56	0.334	125.208
	Bayes-E	0.001	0.002	0.002	0.002	0.959	51	0.475	188.519
Based on traditional KL information	KI	0.001	0.002	0.002	0.007	0.824	50	0.369	140.955
	K ^B	0.001	0.002	0.002	0.016	0.812	54	0.328	122.808
	MK ^B	0.001	0.002	0.002	0.014	0.812	53	0.337	126.579
Based on KL information between posteriors	KLP	0.001	0.002	0.002	0.015	0.815	52	0.337	126.425
	MUI	0.001	0.002	0.002	0.015	0.805	53	0.335	125.542
	MCEM	0.001	0.002	0.003	0.031	0.802	49	0.300	110.163

Note. Q1, Q2, and Q3 represent the quartiles of the item exposure rates. Due to the first item in PMCAT is set to be random selected, the number of unexposed items of each select method are all zeroes, so they are omitted here. NO = number of overexposed items; TOR = the test overlap rate; FI = Fisher information; KL = Kullback-Leibler; KLP = KL information between subsequent posteriors; MUI = mutually scored MCAT; MCAT = polytomously scored MCAT; MCEM = multidimensional computerized adaptive testing; MCEM = modified continuous entropy method.

Figure 2. Average MSE and TOR of PMCAT with different correlation levels for Study 2.
 Note. MSC = mean squared error; TOR = test overlap ratio; PMCAT = polytomously scored MCAT.

Simulation conditions of Study 2. Eight target methods are considered in the simulation: A-optimality, Bayes-A, D-optimality, Bayes-D, KI, K^B , MUI, and MCEM. E-optimality, Bayes-E, MK^B , and KLP are not considered here in that (a) E-optimality is in lack of robustness in applications with sparse data and its use is not recommended (Mulder & van der Linden, 2009); (b) MK^B is very time-consuming when being applied to high dimensional data; and (c) MUI is more robust than KLP with respect to error in ability estimation (Mulder & van der Linden, 2010).

The correlations among multiple measured traits are set to three levels: 0.2, 0.5, and 0.8, which represent low, moderate, and high correlation among abilities, respectively. The examinees are generated by following a common multivariate normal distribution. Because three levels of trait correlation are involved in this study, three populations are used for sampling. They share the same mean vector (0, 0, 0, 0), but differ in covariance/correlation matrixes among abilities. Under each correlated condition, a sample size of 1,000 is chosen to obtain stable estimates. Item bank size is again set to be 450 as Study 1. There are only 15 item patterns ($2^4 - 1 = 15$) for four dimensions, and the 15 item patterns are replicated 30 times. If item j measures the l th ability, a_{jl} is drawn from $U(0.5, 1.5)$; else $a_{jl} = 0$. The threshold parameters b_{j1} and b_{j2} are drawn from $N(0, 2)$, and $b_{j1} < b_{j2}$. The first item is chosen by random, and the test length is set to be 20. EAP is used for ability estimation again, in which the prior density is chosen to match the multivariate normal distribution used to generate examinees' abilities. The Euclidean distance (ED), exposure rate (ER), and test overlap ration (TOR) index are used to measure the security of the item pool as Study 1 again. The AED calculates the average Euclidean distance among 1,000 examinees under each condition, and the mean squared error (MSE) is added to measure the estimation accuracy of each dimension,

$$MSE_l = \frac{1}{N} \sum_{i=1}^N (\theta_{il} - \hat{\theta}_{il})^2, \quad l = 1, 2, 3, 4, \quad (27)$$

Results of Study 2. The ability estimation accuracy and item exposure rate in the high dimensional case under each correlation levels are reported in Table 4, which indicate that (a) in the multidimensional situation, the estimation accuracy of KI method is still the worst. (b) The estimation accuracy of MUI and MCEM method remains relatively high, and the exposure rate of MCEM is lower than MCEM. (c) Except for KI method, the estimation accuracy of A-optimality is relatively poor, and the exposure rate is relatively low, which is contrary to the performance of A-optimality in the two-dimensional case. Figure 2 shows the influence of correlation

Table 4. MSE AED and Exposure Rate of PMCAT With Different Correlation Levels.

Select Method	MSE(θ_1)	MSE(θ_2)	MSE(θ_3)	MSE(θ_4)	Average MSE	AED	χ^2	TOR
<i>r</i> = 0.2								
A	0.1386	0.1431	0.1465	0.1382	0.1416	0.6996	80.9450	0.2235
D	0.1405	0.136	0.1384	0.1328	0.1369	0.6899	104.3726	0.2757
Bayes-A	0.1455	0.1358	0.1461	0.1267	0.1386	0.6846	101.3575	0.2690
Bayes-D	0.1403	0.1422	0.1349	0.1269	0.1361	0.6801	112.2605	0.2932
KI	0.1948	0.1956	0.1841	0.1658	0.1851	0.7875	125.9938	0.3238
K^B	0.1368	0.1441	0.1352	0.1211	0.1343	0.6764	122.3438	0.3156
MUI	0.1271	0.1364	0.1292	0.1276	0.1301	0.6630	133.6977	0.3409
MCEM	0.1343	0.1057	0.1418	0.1114	0.1233	0.6443	120.6732	0.3085
<i>r</i> = 0.5								
A	0.1499	0.1390	0.1299	0.1376	0.1391	0.6899	75.3229	0.2110
D	0.1493	0.1256	0.1282	0.1316	0.1337	0.6856	90.7243	0.2453
Bayes-A	0.1490	0.1290	0.1356	0.1369	0.1376	0.6722	89.4731	0.2425
Bayes-D	0.1493	0.1290	0.1343	0.1364	0.1373	0.6818	91.3529	0.2467
KI	0.1891	0.1621	0.1513	0.1535	0.1640	0.7433	112.1356	0.2929
K^B	0.1514	0.1251	0.1282	0.1315	0.1341	0.6719	100.2106	0.2664
MUI	0.1457	0.1255	0.1323	0.1300	0.1333	0.6708	111.5860	0.2917
MCEM	0.1527	0.1308	0.1301	0.1341	0.1354	0.6757	104.5333	0.2760
<i>r</i> = 0.8								
A	0.0985	0.0956	0.0931	0.1010	0.0971	0.5747	69.1168	0.1972
D	0.0966	0.0882	0.0871	0.0943	0.0916	0.5698	81.0363	0.2237
Bayes-A	0.104	0.0891	0.089	0.0956	0.0944	0.5607	81.1072	0.2239
Bayes-D	0.1086	0.0896	0.0941	0.1004	0.0982	0.5794	81.7891	0.2254
KI	0.1178	0.0966	0.0999	0.1059	0.1051	0.6012	107.1943	0.2819
K^B	0.0985	0.0852	0.0913	0.0944	0.0924	0.5621	86.1786	0.2352
MUI	0.0969	0.0842	0.0899	0.0955	0.0916	0.5607	93.2439	0.2509
MCEM	0.0975	0.0862	0.0867	0.0956	0.0915	0.5592	88.1030	0.2395

Note. *r* = the correlation among multidimensional abilities. MSC = mean squared error; AED = average Euclidean distance; PMCAT = polytomously scored MCAT; TOR = the test overlap rate; KL = Kullback–Leibler; MUI = mutual information; MCAT = multidimensional computerized adaptive testing; MCEM = modified continuous entropy method.

on estimation accuracy and item pool security. From the figure, we can clearly see that (a) for most strategies, the conditions of *r* = 0.2 and *r* = 0.5 have very similar results, while *r* = 0.8 yields lower MSE than the former two conditions. For the KI method, the MSE and AED decrease significantly with the increase of correlation. (b) The item exposure rates decrease as the correlation increases.

Summary and Discussion

The research of MCAT delivering polytomous items has both theoretical and practical significance. Several selection criteria in PMCAT were extended to fit polytomous MIRT model in this study, and two new item selection methods had been proposed. Two simulation studies were carried out to compare the performance of the PMCAT under different conditions. The PMCAT of two dimensional and higher dimensional would be discussed separately in the following.

When only two dimensions are considered, the proposed MK^B and MCEM methods are all rational and suitable for PMCAT. By using the modified posterior distribution, which replaces the fixed prior distribution by the updating posterior distribution, the MCEM method extracts more information from the responses, so as to lead a more accuracy estimate than KLP and MUI. Combined with the security of the pool, MCEM is the ideal one among all, for it earns the lowest item exposure rate and a relatively high accuracy.

From the profile of estimate accuracy, most of the extended selection criteria for PMCAT are feasible except KI method which shows relatively lower estimation precision. It is consistent with the previous study on dichotomous item selection methods (Wang & Chang, 2011) and polytomous item selection methods (Lin, 2012), compared with D-optimal, MUI, CEM, and KI method. It may be because that items with larger KI do not necessarily provide higher power for discriminating θ from $\hat{\theta}$. For any item j satisfies $\sum_{l=1}^p a_{jl}(\hat{\theta}_l - \theta_l) = 0$, it may has high KI but to provide no discrimination power with respect to θ and $\hat{\theta}$ as $KL_j(\hat{\theta}||\theta) = 0$, and hence, it should be avoided (Wang & Chang, 2011).

Mulder and van der Linden (2009) assumed that E-optimality was unfavorable for MCAT because of its instability. However, E-optimality and Bayes E-optimality performed well in simulation study 1. It may be because E-optimality is in lack of robustness in applications with sparse data (Mulder & van der Linden, 2009), but the FI matrix under two-dimensional case in this study is not a sparse one.

Furthermore, some conclusions can also be obtained under some given conditions. First, A-optimality, K^B , and MCEM earn the best estimation accuracy in each type of item selection method. By putting aside the KI method, the three types of item selection method have no significant differences in estimation accuracy. But the methods based on KL information between posteriors are more flexible, for they can be straightforwardly applicable to different model-based CAT settings (Wang & Chang, 2011). Second, most of the item selection methods of PMCAT yield much smaller estimation error for the pattern of $\theta_1 = \theta_2$, which is in accordance with the reality that most abilities are related. From the record of discrimination parameters, we can conclude that only selecting items with high discrimination parameters in both dimensions cannot guarantee the best estimate accuracy, like KI method.

Last, by utilizing high dimension matrix and parallel computing technique instead of loop statement in the programming of using MATLAB software, this research also achieved high computing speed. (When calculating the selection index, the remaining items are represented by a number of items $\times$ number of response categories $\times$ number of integral node matrix, and the loop between items is replaced by matrix multiplication. So, the item selection index of each remaining items can be calculated simultaneously). When the two dimensional study was implemented, under the operating environment of using an ordinary laptop computer (i5-3320M CPU@2.6GHz, RAM 4.00G), except the MK^B method (needs about 2 s), all the criteria cost less than 0.4 s of each replication finishing 25 items, which meets the time requirements of CAT.

Above all, Simulation Study 1 reflects that the polytomous methods in the study are all feasible to the two-dimensional PMCAT, except for KI method.

The findings of high-dimensional PMCAT with different correlation levels suggest that the multivariate KI method is not recommended in PMCAT for it could not satisfy the accuracy requirement in the four-dimensional case either. Compared with other methods, A-optimality method earns the relatively poor estimation accuracy in the four-dimensional case rather than the highest estimation accuracy in the two-dimensional case, which might indicate the unstable performance of A-optimality in high-dimensional context. Despite of the popularity of the optimality-based criteria in this study, they may behave unfavorably when being used for item selection in high-dimensional MCAT, because they involve taking the inverse or calculating the determinant of the FI matrix, which may run into numerical difficulty if the information matrix is singular or near singular. Besides, MUI and MCEM keep relatively high estimation accuracy, especially when $r = 0.2$ and $r = 0.8$, MCEM earns the highest estimation accuracy of all. Although their exposure rates are higher than the optimality-based criteria, the exposure rate of MCEM is still less than the MUI (the same to original CEM) method.

In addition, except for KI method, the correlativity has an impact on estimation accuracy only when the correlation among ability dimensions is higher than medium. And the estimation error of KI method drops dramatically while the correlation becomes larger. This suggests that the effect of attribute correlation is not linear. Moreover, the greater the correlation among ability dimensions is, the lower the item exposure rate would be.

Last, the time of computing in four-dimensional case gets longer than that in two-dimensional case, especially for the algorithms using integrals. The FI-based criteria cost about 0.5 s of each replication finishing 20 items, while the KI-based criteria (all contain integral operations) cost about half a minute. It is practical to identify simpler and equally efficient methods that can reduce the computational intensity in higher dimensional structure.

This represents an initial step in the study of item selection strategies for PMCAT. Based on the research, as a next step, it would be interesting to investigate item selection under various conditions. To make the PMCAT become applicable in real work, it is necessary to discuss the nonstatistical factors, such as item exposure control, content constraints, and so on. Moreover, the practical precondition of PMCAT is the item pool consisting of considerable number of items that have been well calibrated. The item parameter equating, item online calibration, and other issues, which may be faced in the building of item pool, must be further discussed. MCAT will become one of main test delivery approaches in the future testing thanks to its diagnostic feature. Furthermore, polytomous items should be applied into the test for their strength in providing more information and testing the complicated abilities and skills. Research in MCAT test using polytomous items should be investigated profoundly from both theoretical perspective and practical application. To facilitate information, the simulation study seems promising in how researchers and actual users could potentially choose item selection method when dealing with polytomous items in MCAT in the studied context, particularly when the MGRM is considered.

Acknowledgments

Useful suggestions given by the anonymous reviewers are acknowledged.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article is supported by National Natural Science Foundation of China (31660278 and 31760288).

Supplemental Material

Supplemental material is available for this article online.

References

- Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229. doi:10.1177/014662169602000303
- Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222. doi:10.1177/01466219922031338
- Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40, 129-145. doi:10.1111/j.1745-3984.2003.tb01100.x

- De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, 16, 327-343. doi:10.1177/014662169201600403
- Donoghue, J. R. (1994). An empirical examination of the IRT information function of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 31, 295-311.
- Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, 46, 84-103. doi:10.1111/j.1745-3984.2009.01070.x
- Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a Computer Adaptive Test for depression based on the Dutch-Flemish Version of the PROMIS Item Bank. *Evaluation & the Health Professions*, 40, 79-105. doi:10.1177/0163278716684168
- Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, 4, Article 13. doi:10.1186/1471-244X-4-13
- Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In *Objective measurement: Theory into practice* (Vol. 1, pp. 236-258). Norwood, NJ: Ablex.
- Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model* (Doctoral dissertation). University of Illinois at Urbana-Champaign.
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20, 389-404. doi:10.1177/014662169602000406
- Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, 74, 273-296. doi:10.1007/s11336-008-9097-5
- Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing, statistics for social and behavioral sciences* (pp. 77-101). doi:10.1007/978-0-387-85461-8
- Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, 19, 73-90. doi:10.1177/014662169501900109
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. doi:10.1007/978-0-387-89976-3
- Robbins, H. (1985). An empirical Bayes estimation problem. In T. L. Lai & D. Siegmund (Eds.), *Herbert Robbins selected papers* (pp. 72-73). New York, NY: Springer. doi:10.1007/978-1-4612-5110-1_6
- Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the first Conference on Computerized Adaptive Testing* (pp. 5-17). Washington, DC: U.S. Government Printing Office.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354. doi:10.1007/BF02294343
- Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188, 147-155. doi:10.1016/j.psychres.2010.12.001
- van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24, 398-412. doi:10.3102/10769986024004398
- Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67, 575-588. doi:10.1007/BF02295132
- Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika*, 76, 363-384. doi:10.1007/s11336-011-9215-7
- Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, 76, 13-39. doi:10.1007/s11336-010-9186-0
- Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30, 469-492. doi:10.1177/0146621605284537