A Comparison Study of Item Exposure Control Strategies in MCAT

Authors: Xiuzhen Mao; Burhanettin Ozdemir;Yating Wang; Tao Xin

Publication date:2016/10

Abstract

Four item selection indexes with and without exposure control are evaluated and compared in multidimensional computerized adaptive testing (CAT). The four item selection indices are D-optimality, Posterior expectation Kullback–Leibler information (KLP), the minimized error variance of the linear combination score with equal weight (V1), and the minimized error variance of the composite score with optimized weight (V2). The maximum priority index (MPI) method for unidimensional CAT and two item exposure control methods (the restrictive threshold (RT) method and restrictive progressive (RPG) method, originally proposed for cognitive diagnostic CAT) are adopted. The results show that: (1) KLP, D-optimality, and V1 perform well in recovering domain scores, and all outperform V2 in psychometric precision; (2) KLP, D-optimality, V1, and V2 produce an unbalanced distribution of item exposure rates, although V1 and V2 offer improved item pool usage rates; (3) all the exposure control strategies improve the exposure uniformity greatly and with very little loss in psychometric precision; (4) RPG and MPI perform similarly in exposure control, and are both better than RT.

Keywords: Multidimensional Item Response Theory; Computerized Adaptive Testing; Item

Selection Methods; Exposure Control Strategy; Psychometric Precision.

Introduction

The fact that test items are chosen sequentially and adaptively in computerized adaptive testing (CAT) has broken the traditional testing mode in which thousands of people respond to the same items at the same time. Nowadays, CAT is increasingly favored by test practitioners and researchers for its higher efficiency, shorter test time, and lower pressure than paper and pencil (P&P) testing. Another more fascinating characteristic of CAT is that different item response models can be applied, including unidimensional, multidimensional, and cognitive diagnostic models.

Multidimensional computer adaptive testing (MCAT) possesses the advantages of both multidimensional item response theory (MIRT) and CAT. On the one hand, a large number of studies based on different test conditions have arrived at the conclusion that MCAT provides higher efficiency than unidimensional CAT. For example, Segall (1996) employed simulated data based on nine adaptive power tests of the Armed Services Vocational Aptitude Battery (ASVAB) to show that MCAT reduced by about one-third the number of items required to generate equal or higher reliability with similar precision to unidimensional CAT. Luecht (1996) demonstrated that MCAT can reduce the number of items for tests with content constraints by 25–40%. Further, Wang and Chen (2004) illustrated the higher efficiency of MCAT compared with unidimensional

CAT under different latent trait correlations, latent numbers, and scoring levels. On the other hand, the fact that several ability profiles are estimated simultaneously indicates the ability of MCAT to offer detailed diagnostic information regarding domain scores and overall scores. The advantages of multi-dimensionality and high efficiency make MCAT better suited to real tests than unidimensional CAT. Hence, many studies on MCAT have considered real item pools, such as TerraNova (Yao, 2010), American College Testing (ACT) (Veldkamp & van der Linden, 2002), and ASVAB (Segall, 1996; Yao, 2012, 2014a).

Since Bloxom and Vale (1987) extended unidimensional CAT to MCAT, it has received increasing attention, and several breakthroughs have been reported in the last decade. Among the studies on ability estimation methods, the testing stopping rule, and item replenishing, item selection rules have become popular because of their important role in affecting the test quality and psychometric precision. Thus, most researchers focus on proposing new item selection indices to decrease errors in ability estimation. However, Yao (2014a) pointed out that most item selection methods tend to select a particular type of item, leading to the problem of unbalanced item utility. She also gave an example of the Kullback–Leibler index, which prefers items that have either a high discriminator at each dimension or significantly different discriminators among different dimensions. As another example, the D-optimality index tends to select items with a high discriminator in only one dimension (Wang, Chang, & Boughton, 2011). Nowadays, CAT is increasingly used in many kinds of tests. Hence, item exposure control is important in the application of MCAT, especially for its application to high-stakes tests. Furthermore, few studies have investigated this problem in MCAT. Hence, the goal of the present study is to evaluate the performance of some exposure control techniques in MCAT.

To date, many of the exposure control methods used in unidimensional CAT have been generalized to MCAT. For example, Finkelman, Nering and Roussos (2009) extended the Sympson–Hetter (S-H) (Sympson & Hetter, 1985) and Stocking–Lewis (S-L) (Stocking & Lewis, 1998) methods to MCAT. They found that all the S-H, generalized S-H, and generalized S-L methods do well in controlling the maximum item exposure rates. However, simulation experiments to create the exposure control parameters are time-consuming. Furthermore, there still exist some underexposed items. In addition, Yao (2014a) compared S-H with the fix-rate procedure. The fix-rate procedure is similar to the maximum priority index (MPI) method proposed by Cheng and Chang (2009) for unidimensional CAT. She showed that the S-H method performs better in terms of test precision, whereas the latter gives a higher item bank usage and controls the maximum item exposure rate well.

The $|a_{j1}-a_{j2}|$-stratification method (Lee, Ip, & Fuh, 2008) is based on the principle of the a-stratification method (Chang & Ying, 1999). The item pool is stratified according to the absolute value of $a_{j1}-a_{j2}$, where $a=(a_{j1},a_{j2})$ denotes the item discrimination vector of item j. It was reported that the $|a_{j1}-a_{j2}|$-stratification method is effective in combating overused items and increasing the item pool usage. However, this method cannot guarantee that no items are overexposed. Thus, Huebner, Wang, Quinlan, and Seubert (2015) combined $|a_{j1}-a_{j2}|$-stratification with the item eligibility method (van der Linden & Veldkamp, 2007) with the aim of enhancing the balance of item exposure. This combination method improves the exposure rates of underused items and suppresses the observed maximum item exposure rate.

However, these two methods are restricted to tests with two dimensions. Constructing a suitable functional of the discrimination parameter for tests with more than two dimensions remains an important research problem.

It is well known that the uniformity of item exposure rates is affected by the numbers of overexposed and underexposed items. Of the above mentioned exposure control methods used in MCAT, the S-H, generalized S-H, generalized S-L, fix-rate, and item eligibility methods perform well in suppressing the maximum item exposure rates, and the $|a_{j1}-a_{j2}|$-stratification method effectively improves the utility of underexposed items. Although the combination method used by Huebner, et al. (2015) performs well in both aspects, it is only suitable for tests with two dimensions.

The uniformity of item exposure rates and measurement precision are the two most important considerations during the application of MCAT to practical tests, especially for high-stakes tests. Because they always trade-off with one another, practitioners hope to find some item selection method that not only guarantees test precision, but also decreases the maximum item exposure rate while increasing the exposure rate of underexposed items. However, there are no methods that can effectively balance item exposure rates for tests with more than two dimensions. In addition, two exposure control methods have not been studied for MCAT: the restrictive threshold (RT) method and the restrictive progressive (RPG) method. It has been reported that they perform well in balancing the item exposure rate of cognitive diagnostic CAT (Wang, Chang, & Huebner, 2011). Therefore, the focus of the present study is whether RT and RPG can simultaneously suppress the maximum item exposure rates and increase the exposure rates of underexposed items without losing psychometric precision in MCAT. Further, their performance is compared with that of the MPI method.

In the remainder of this paper, we first introduce the MIRT model employed in this study and the ability estimation method. Then, some item selection indices and exposure control strategies are described. The performance of four item selection indices with and without each of the three exposure control strategies under different latent trait correlation levels are examined through a series of simulation experiments. The results, conclusions, and discussion are given in the final two sections of the paper.

MIRT model and ability estimation method

Multidimensional Two-Parameter Logistic (M-2PL) Model

MIRT models are usually classified as compensatory or non-compensatory based on whether a strong ability can compensate for other weak profiles. Bolt and Lall (2003) reported that both types are able to fit the data generated by non-compensatory models, but non-compensatory models cannot match the data generated from compensatory models. Thus, because of the advantages of compensatory models and the wide usage of MCAT in dealing with dichotomous items (van der Linden, 1999; Veldkamp & van der Linden, 2002; Mulder & van der Linden, 2010), the M-2PL model was adopted to simulate item parameters and generate item responses.

For some item j, M-2PL includes a scalar difficulty parameter b_j and discrimination vector $a_j = (a_{j1}, a_{j2}, ..., a_{jD})^T$ (McKinley & Reckase, 1982), where *T* denotes the transpose and *D* is the number of dimensions. For an examinee with ability $\theta = (\theta_1, \theta_2, ..., \theta_D)^T$, the item response function can then be described as:

$$P_j(\vec{\theta}) = P(x_j = 1 \mid \vec{\theta}, \vec{a}_j, b_j) = \frac{1}{1 + \exp[-(\vec{a}_j^T \cdot \vec{\theta} - b_j)]}. \qquad (1)$$

where $\vec{a}_j^T \cdot \vec{\theta} - b_j = \sum_{l=1}^{D} a_{jl} \cdot \theta_l - b_j$ denotes a straight line in *D*-dimensional space. The compensatory features of M-2PL originate from the fact that all examinees giving equal $\vec{a}_j^T \cdot \vec{\theta}$ possess the same response probability.

Ability Estimation Method: Maximum a Posteriori (MAP) Estimation

Yao (2014b) compared MAP, expected a posteriori (EAP), and maximum likelihood estimation (MLE) in a simulation experiment using item parameters estimated from the ASVAB Armed Forces Qualification Test. She pointed out that: (a) MLE generates smaller bias and larger root mean square error (RMSE), whereas MAP and EAP using strong prior information or standard normal priors produced higher precision in the recovery of ability, (b) EAP and MAP behave similarly, but EAP takes a longer time than MAP. Recently, Huebner, et al. (2015) compared EAP with MLE in MCAT, and proved that EAP always produces more stable results and lower mean square error in the ability estimators than MLE. MAP is adopted in this study for its competitive precision and easier computation compared with EAP in MIRT.

Let $f(\vec{\theta})$ denote the prior density function of $\vec{\theta}$. This is assumed to be a multivariate normal distribution with mean value $\vec{\mu}_0$ and variance-covariance matrix Σ_0. For convenience, the response to item j is indicated as x_j, and $\vec{X}_{k-1}$ represents the response vector of the first $k-1$ items administered. The posterior density function of $\vec{\theta}$ is denoted by $f(\vec{\theta} \mid \vec{X}_{k-1})$. Based on Bayes' theorem, $f(\vec{\theta} \mid \vec{X}_{k-1}) \propto L(\vec{X}_{k-1} \mid \vec{\theta}) \cdot f(\vec{\theta})$, where $L(\vec{X}_{k-1} \mid \vec{\theta})$ denotes the likelihood function. Hence, the goal of MAP is to find the mode that maximizes the posterior density function $f(\vec{\theta} \mid \vec{X}_{k-1})$. That is, the ability estimator $\hat{\vec{\theta}}^{MAP}$ is equivalent to the solution of $\frac{\partial \log f(\vec{\theta} \mid \vec{X}_{k-1})}{\partial \vec{\theta}_l} = 0 \quad (l = 1,2,\ldots,D)$. Furthermore, Newton-Raphson iteration can be used to solve this equation; for details, see Yao (2014b).

Item Selection Indices and Exposure Control Strategies

To simplify the description, we first introduce some notation. N represents the number of examinees, and L is the test length. Set R refers to the item bank, which has a capacity of M. Set $R_{k-1} = R \setminus \{i_1, i_2, \ldots i_{k-1}\}$ and $\hat{\vec{\theta}}^{k-1}$ express the remainder of the item bank and the temporary estimator after administering the first $k-1$ items, respectively.

Item Selection Indices

The following four indices are chosen as item selection criteria based on the consideration of computation complexity and running time.

D-optimality. The Fisher information of each item in MIRT is no longer a number, but a matrix. Specifically, the Fisher information for the *j*th item in M-2PL is

$$I_j(\vec{\theta}) = P_j(\vec{\theta}) \cdot (1 - P_j(\vec{\theta})) \cdot (\vec{a}_j^T \vec{a}_j). \tag{2}$$

After $k-1$ items have been administered, the estimators form an ellipse or sphere V_{k-1}. To decrease the size or volume of V_{k-1} as quickly as possible, Segall (1996) proposed that the *k*th item should maximize the determinant of the posterior test Fisher information matrix. Thus, the Bayesian item selection rule is expressed as

$$D_k = \max\{| I_{k-1}(\hat{\vec{\theta}}^{k-1}) + I_j(\hat{\vec{\theta}}^{k-1}) + \Sigma_0^{-1} |, \qquad j \in R_{k-1}\}. \tag{3}$$

where $I_{k-1}(\hat{\vec{\theta}}^{k-1})$ represents the test information of the first $k-1$ items already be administered calculated at the current estimated ability, and $I_j(\hat{\vec{\theta}}^{k-1})$ indicates the Fisher information of the *j*th $(j \in R_{k-1})$ candidate item. This method was called D-optimality by Mulder and van der Linden (2009), and the item with the largest D_k is chosen from the remainder pool.

Posterior Expected Kullback–Leibler Information (KLP). This method is obtained by weighting the KL information according to the posterior distribution of ability. That is, the *k*th item is selected according to

$$KLP_k = \max\{\int_{\vec{\theta}} KL_j(\hat{\vec{\theta}}^{k-1},\vec{\theta})\cdot f(\vec{\theta} \mid \vec{X}_{k-1})d\vec{\theta}, \qquad j \in R_{k-1}\}. \tag{4}$$

where

$$\begin{aligned} KL_j(\hat{\vec{\theta}}^{k-1},\vec{\theta}) &= E_{\vec{\theta}} \log[\frac{P_j(x_j \mid \vec{\theta},\vec{a}_j,b_j)}{P_j(x_j \mid \hat{\vec{\theta}}^{k-1},\vec{a}_j,b_j)}] \\ &= P_j(\vec{\theta})\log\frac{P_j(\vec{\theta})}{P_j(\hat{\vec{\theta}}^{k-1})} + (1-P_j(\vec{\theta}))\log\frac{(1-P_j(\vec{\theta}))}{(1-P_j(\hat{\vec{\theta}}^{k-1}))}. \end{aligned} \tag{5}$$

The integral interval is generally narrowed to simplify the computation, and (9) is replaced with

$$KLP_k = \max\{\int_{\theta_1^{k-1}-\gamma_j}^{\theta_1^{k-1}+\gamma_j} \ldots \int_{\theta_D^{k-1}-\gamma_j}^{\hat{\theta}_D^{k-1}+\gamma_j} KL_j(\hat{\vec{\theta}}^{k-1},\vec{\theta})\cdot f(\vec{\theta} \mid \vec{X}_{k-1})d\theta_1 \cdots d\theta_D, \qquad j \in R_{k-1}\}, \tag{6}$$

where γ_j usually takes a value of $3/\sqrt{j}$.

Minimum Error Variance of the Linear Combination Score with Equal Weight (V1). From the perspective of error variance, van der Linden (1999) suggested that the *k*th item should minimize the error variance of the composite score $\vec{\theta}_\alpha = \sum_{l=1}^{D}\theta_l \cdot w_l$. Let $SEM(\vec{\theta}_\alpha)$ denote the standard error of measurement (SEM) for composite score $\vec{\theta}_\alpha$. Yao (2012) derived the formula $SEM(\vec{\theta}_\alpha) = (V(\vec{\theta}_\alpha))^{1/2} = (\vec{w}V(\vec{\theta})\vec{w}^T)^{1/2}$, where $V(\vec{\theta})$ is usually approximated by

$I_{k-1}(\hat{\vec{\theta}}^{k\text{-}1})^{-1}$. Given equal weights $w = (1/D,\ 1/D, \ldots,\ 1/D)$ among the different dimensions, the item that minimizes $SEM(\vec{\theta}_{\alpha})$ will be selected by V1.

Minimum Error Variance of the Linear Combination score with Optimized Weight (V2). The weight that minimizes the SEM of the composite ability is named the optimal weight. Yao (2012) proved the existence of the optimized weight, and derived its formula as

$$w = \frac{1}{\sum_{o=1}^{D}\sum_{l=1}^{D} b_{ol}} \cdot [1,1,\ldots,1]_{1\times D} \cdot I_{k-1}(\vec{\theta}) \quad . \tag{7}$$

In this expression, b_{ol} denotes the element of $I_{k-1}(\vec{\theta})$ located on the oth row and lth column. The procedure of V2 involves finding the optimal weight vector, then calculating SEM for each candidate item according to the optimal weight. Finally, the item with the lowest SEM is selected from the remainder pool. Note that the optimal weight is updated after administering each item. Thus, the only difference between V2 and V1 is in the determination of the weight used to compute $SEM(\vec{\theta}_{\alpha})$.

Strategies of Item Exposure Control

The RT and RPG methods proposed by Wang, et al. (2011) are two exposure control methods used in cognitive diagnostic CAT. Both can be easily generalized to MCAT.

The RT method. In the RT method, a shadow item bank is constructed at the beginning of each test by removing all overexposed items from the original item bank. Each item is then selected at random from the candidate item set constructed beforehand. Let “Index” denote the value of the item selection indices. The candidate item set includes all items whose information values lie in $[\max(Index)-\delta, \max(Index)]$ for both D-optimality and KLP or $[\min(Index), \min(Index)+\delta]$ for V1 and V2. The constant δ is defined as $\delta = [\max(Index)-\min(Index)]\cdot(1-k/L)^{\beta}$. Larger values of β give a shorter information interval length. As a result, the measurement precision is improved by decreasing the uniformity of the item exposure distribution. In summary, β is used to balance the requirements of item exposure rate control and measurement precision. In this study, we use β = 0.5.

The RPG Method. The kth (k = 1, 2, …, L) item is selected according to (8) for D-optimality and KLP, and according to (9) for V1 and V2. They are

$$i_k = \max\{(1-er_j/r^{\max})\cdot[(1-k/L)u_j + Index_j \times \beta k/L], \qquad j \in S_{k-1}\} \tag{8}$$

$$and \qquad i_k = \max\{(1-er_j/r^{\max})\cdot[(1-k/L)R_j + (C-Index_j)\times \beta k/L], \qquad j \in S_{k-1}\}, \tag{9}$$

where er_j denotes the observed exposure rate of item j and $r^{\max}$ denotes the allowed maximum exposure rate. Let H^* be the maximum item information in S_{k-1}. Then, u_j is uniformly extracted from interval $(0, H^*)$. The parameter β plays the same role and takes the same value as in the RT method. The constant *C* should be greater than all the SEMs; in this study, we set *C* = 10000. Note that SEM is always very large for the first several items, and decreases rapidly to less than 1000. Thus, it is better to set *C* to be greater than 1000.

The maximum priority index method (MPI). According to Cheng and Chang (2009), the priority index (*PI*) of item j with the requirement of the maximum exposure rate is expressed as

$$PI_j = \frac{r^{\max} - n_j/N}{r^{\max}} \cdot Index_j, \tag{10}$$

where n_i represents the administration frequency of item j, and "*index*" refers to the D-optimality or KLP index. Finally, the task of the MPI method is to identify the item with the largest *PI*. The role of *C* is similar to that in RPG. For V1 and V2, PI_j should be changed accordingly, that is

$$PI_j = \frac{r^{\max} - n_j / M}{r^{\max}} \cdot (C - Index_j) . \quad (11)$$

Method

A simulation study was conducted to evaluate and compare the effectiveness of the above exposure control methods. Matlab (version7.10.0.499) was used to write MCAT codes and run the simulation experiments.

Design of Simulation Study

Item Bank Construction. Although Stocking (1994) suggested that the pool should contain at least 12 times as many items as the test length, many simulation studies on MCAT have used a more restrictive item pool. For example, the item pool used by van der Linden (1999) contained 500 items while the test length was 50; Lee, et al. (2008) used an item pool of 480 items with test lengths of 30 and 60; and the item pools described in Veldkamp and van der Linden (2002) and Mulder and van der Linden (2009) contained fewer than 200 items while the test length was greater than 30. Thus, it is reasonable to construct an item pool of 450 items for a test length of 30.

To simplify the experimental conditions, most simulation studies generate item parameters and item responses according to M-2PL or M-3PL with the assumption that there are two or three dimensions (van der Linden, 1999; Veldkamp & van der Linden, 2002; Lee et al., 2008; Mulder & van der Linden, 2009; Finkelman et al., 2009; Wang, Chang, & Boughton, 2013; Wang & Chang, 2011). Hence, without loss of generality, the items in our simulation contained three dimensions, and the item parameters of the M-2PL model were generated in a similar way to those of Yao and Richard (2006) and Wang and Chang (2011). Specifically, (a_{j1}, a_{j2}, a_{j3}) for item $j(j = 1,2,...450)$ were drawn from $\log N(0,\ 0.5)$ independently and $b_j(j = 1,2,...450)$ were drawn from $N(0,1)$.

Examinees and Item Responses. All 5000 examinees were simulated uniformly from a multivariate normal distribution, as in previous research (Wang & Chang, 2011; Yao, Pommerich, & Segall, 2014; Wang et al., 2013). Three levels of correlation were considered in the experiments. The mean ability was [0, 0, 0] and the variance-covariance matrix was

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} (\rho = 0.3, 0.6, 0.8).$$

Let P_{ij} and x_{ij} denote the correct response probability and actual response (0 or 1) corresponding to the *j*th $(j = 1,2,...,450)$ item and the *i*th $(i = 1,2,...,5000)$ examinee. P_{ij} was computed from the M-2PL model, and u_{ij} was selected uniformly from (0, 1). We set x_{ij} = 1 if $P_{ij} \geq u_{ij}$. Otherwise, if P_{ij} < u_{ij}, x_{ij} = 0.

Item Selection Methods. Four item selection indices with and without the three exposure control methods yields a total of 16 item selection strategies.

Estimation of Ability. The initial abilities were selected from the standard multivariate normal distribution. MAP was used to update the domain abilities during the test, and multivariate standardized normality was applied as the prior distribution.

Evaluation Criteria. The bias and mean square error (MSE) of each dimension were used to evaluate the precision of the ability estimators. They were computed as

$$Bias_l = \frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{\theta}_l - \theta_l) \qquad (l = 1,2,3), \tag{12}$$

and

$$MSE_l = \frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{\theta}_l - \theta_l)^2 \qquad (l = 1,2,3). \tag{13}$$

To assess the equalization of exposure rates, we used (a) the number of items never reached and the number of items with exposure rates greater than 0.2, (b) the χ^2 statistic, and (c) the test overlap rate. The χ^2 statistic was calculated as

$$\chi^2 = \sum_{i=1}^{N} \frac{(er_i - \overline{er})^2}{\overline{er}}. \tag{14}$$

Smaller values of χ^2 indicate smaller differences between the observed and expected item exposure rates. Finally, the test overlap rate was computed according to the expression proposed by Chen, Ankenmann, and Spray (2003):

$$\hat{T} = \frac{M}{L} S_{er}^2 + \frac{L}{M}. \tag{15}$$

In (15), S_{er}^2 denotes the variance of item exposure rates. Generally, smaller values of $\widehat{T}$ demonstrate more balanced item utility.

Results

Results of Ability Estimation

The differences in bias between two arbitrary dimensions of each method were so small that Figure 1 presents the mean bias of three dimensions. Figure 2 lists the MSEs of each dimension for the different item selection methods and correlation levels.

It is easy to summarize the following results: (a) the biases generated by D-optimality, V1, and V2 are similar and greater than the bias produced by KLP, and (b) for each dimension, KLP produces the smallest MSE, followed by D-optimality, V1, and V2. Generally, it is easy to sort the indices into descending order of KLP, D-optimality, V1, and V2 according to their measurement precision.

The effects of item exposure control methods on the psychometric precision were checked through three aspects. First, from Figure 1, the item exposure strategies have no significant effect on the bias, as the biases produced by the same item selection index using different exposure control methods are similar.

Second, the results of each item selection index with and without item exposure control can be compared. From Figure 2, all the item exposure strategies led to an increase in MSE except for V2. The MSE of V2 was larger than that of V2-RT in most of the cases. The decreased measurement precision may result from the characteristics of V2 in improving the item pool utility. Overall, using an exposure control strategy always decreases the measurement precision.

Furthermore, when the item exposure control methods were combined with D-optimality, KLP, or V2, their performance differed considerably in terms of the measurement precision. However, all the item exposure control methods yielded similar measurement precision when combined with V1. In addition, a higher level of ability correlation seems to narrow the gap in the precision generated by different exposure control methods when combined with the same item selection index.

Finally, we can compare the results of different item exposure control methods. RT always produced the lowest MSE values, thus giving higher measurement precision than RPG and MPI. RPG and MPI performed similarly, although their precision under different item selection indices varied to some degree. The performance of RT and RPG was in accordance with that reported by Wang et al. (2011). Overall, the general order of different exposure control methods sorted by decreasing measurement precision was RT, RPG, and MPI.

Results of Item Exposure Rates. The item exposure rates associated with each item selection index with and without exposure rate control are presented in Table 1 and Figures 3-4.

First, it is easy to infer that the exposure rates are distributed unevenly for D-optimality, KLP, V1, and V2. Taking D-optimality and KLP for illustration, they generate the lowest item bank usage rates and the largest overexposed item and test overlap rates. Although the number of never-reached items in V1 and V2 is close to 0, and the test overlap rates and χ^2 values are smaller than those of D-optimality and KLP, these exposure rate control methods still produce an unsatisfactory item exposure rate distribution. These characteristics can be clearly observed in Figure 4(a), where the exposure rates are depicted in ascending order for each of the four item selection indices. In addition, the results for V1 and V2 obtained from this study coincide with those reported by Yao (2014a).

Second, all the exposure control methods improved the uniformity of exposure rates significantly in terms of increasing item bank usage and lowering overexposed item rates, test overlap rates, and χ^2. According to Table 1, RPG outperformed the other methods in most cases, although MPI performed similarly. From Table 1, it is apparent that all the item exposure distributions follow the same pattern when different item selection indices are combined with the same exposure control method. Hence, Figure 4(b) only illustrates the exposure rate distributions of the exposure control strategies combined with KLP.

In addition, different characteristics of the item exposure rate distribution were observed in different item exposure control methods. From Figure 3, it can be seen that the item pool usage rate reaches 100% for all methods except KLP-MPI. In other words, all item exposure methods significantly improve the item pool usage. Checking the overexposed items, both RPG and MPI produced more overexposed items than RT under most test conditions. Generally, RT is able to control the item exposure rates to be lower than the allowable maximum value,, whereas both RPG and MPI result in some items with exposure rates greater than 0.2.

Further, it is worth pointing out some special findings when it comes to discussing certain exposure control methods. First, compared to D-MPI, V1-MPI, and V2-MPI, KLP-MPI generated a more unbalanced item exposure rate distribution. Second, when RPG was used with V1 or V2, there were always one or two items exposed to everybody. Checking the internal results of V1-RPG and V2-RPG revealed that many error variance values in Matlab were labeled “NaN” in the case of choosing the first or second item. In other words, it can be inferred that the overexposed items in V1-RPG and V2-RPG were mainly due to the non-distinctive item information matrix in V1 and V2. Furthermore, the test overlap rate and χ^2 of V1-RPG and V2-RPG were affected by the first one or two administered items accordingly.

Overall, although the item exposure control strategies produced different patterns of item exposure rates, they all considerably improve the balance of the item exposure distribution. This can be seen from comparing Figure 4(a) and 4(b). In addition, the trade-off between the measurement precision and the item exposure distribution is also displayed in the results.

Conclusions and Discussions

Many studies have acknowledged the advantages of CAT over P&P tests and computer-based tests, such as the decrease in test length, increase in measurement precision, and better model fits. Along with the obvious advantages of MCAT, choosing the most appropriate item selection rule is a vital step for a successful application (Wang & Chang, 2011). Although the proposed item selection methods yield good results in precision, they are vulnerable to the issue of dealing with overexposed items (those that are used too often) and underexposed items (used too rarely). As a solution to this problem, different item exposure control methods have been adopted and used together with different item selection methods.

This study has examined the performance of four item selection indices combined with different exposure control methods in MCAT.

Simulations showed that V2 outperforms D-optimality, KLP, and V1 with respect to higher item bank usage rates, fewer overexposed items, and lower test overlap rates. Generally, the results of all item selection indices without using item exposure control were unsatisfactory with respect to item exposure statistics. The results indicate that, without using item exposure control, the item selection indices can be sorted in order of psychometric precision as KLP, D-optimality, V1, and V2. In addition, when using item exposure control methods, the measurement precision tended to decrease in all item selection indices.

In comparing the item exposure rate distribution generated by different item exposure control methods, RPG outperformed the other methods in most cases, although MPI performed similarly. The RT method gave the worst performance. Furthermore, each item exposure control method yields the same exposure rate pattern under different item selection indices. When it comes to comparing the measurement precision, the performance of the different exposure control methods can be ordered as RT, RPG, and MPI. This kind of trade-off between measurement precision, utility of item pool, and evenness of item exposure rate has been observed in many studies (Chang & Twu, 1998). İn other words, the measurement precision needs to be sacrificed, to some extent, to keep the exposure rate at the desired value.

Both the present study and the work of Wang et al. (2011) showed that the measurement precision of the RT method was higher than that of the RPG method under the same test conditions, and the RT method performed slightly worse than RPG in the evenness of the item exposure distribution. In conclusion, among the three exposure control methods examined in this study, both RT and RPG offer balanced precision and item exposure control, whereas MPI performed well in controlling the item exposure rate with a noticeable loss in precision.

Several issues regarding item selection methods for MCAT deserve further investigation. First, although D-optimality, V1, and V2 are much faster than KLP, the run-time usually increases with the number of test dimensions. As a consequence, time-consuming methods can hinder the practice of MCAT in dealing with complex test conditions. In fact, the benefits of MCAT over unidimensional CAT mainly lie in the detailed cognitive information obtained based on multiple dimensions. Hence, there is a need for more work on algorithms that reduce the computation time of the item selection methods, or simplified and valid item selection methods based on existing rules, such as the two simplified KL indexes provided by Wang et al. (2011).

Second, the test measurement precision of each dimension can be guaranteed by most MCAT item selection methods automatically, but thousands of other constraints are encountered in real tests. Hence, it would be useful to research how to deal with nonstatistical constraints in MCAT.

Third, polytomous items such as opening responding items and construction items have now begun to appear in CAT (Bejar, 1991). There is no doubt that research on polytomous items will increase in popularity. However, most current research on MCAT deals with dichotomous items. Thus, it is important for researchers to propose item selection methods or extend methods for dichotomous items, such as the mutual information index, KL, and Shannon entropy, to deal with polytomous items.

References

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology*, 76, 522-532.

Bloxom, B. M., & Vale, C. D. (1987). Multidimensional adaptive testing: a procedure for sequential estimation of the posterior centriod and dispersion of theta. Paper presented at the meeting of the Psychometric society, Montreal, Canada.

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

Psychological Measurement, 27, 395-414.

Chang, S. W. & Twu, B. Y. (1998). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. ACT Research Report Series, 98-3.

Chang, H. H., & Ying, Z. L. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.

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.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British journal of mathematical and statistical psychology*, 62, 369–383.

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.

Huebner, A. R., Wang, C., Quinlan, K., & Seubert, L. (2015). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posterior estimation. *Behavior Research Methods*, DOI 10.3758/s13428-015-0659-z.

Lee, Y. H., Ip, E. H., & Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, 68, 215-232.

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

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

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria. *Psychometrika*, 74, 273-296.

Mulder, J. & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden AND c. A. W. Glas (eds.), Elements of Adaptive Testing, Statistics for Social and Behaviroal Sciences, Springer Science+Businesws Media, 2010.

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

Stocking, M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report No. 94–5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, 23, 57–75.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

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

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, 32, 398-418.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67, 575-588.

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

Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, 76, 13-39.

Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 37(2), 99-122.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48, 255-273.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychologica Measurement*, 28, 295–316.

Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement*, 47, 339-360.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: theory and applications. *Psychometrika*, 77, 495-523.

Yao, L. (2014a). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constrains. *Journal of Educational Measurement*, 51,18-38.

Yao, L. (2014b). Multidimensional item response theory for score reporting. In Cheng, Y., & Chang, H.-H. (Eds.), Advances in modern international testing:Transition from summative to formative assessment. Charlotte, NC: Information Age.

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using Multidimensional CAT to Administer a Short, Yet Precise, Screening Test. *Applied Psychological Measurement*, 38,614-631.

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*, 37, 3-23.

Table 1. Item exposure statistics of each method

Methods	*Overlap rate*	χ^2		Methods	*Overlap rate*	χ^2
D	0.408/0.23/0.23	*152.6/75.14/75.14*		V1	*0.253/0.241/0.237*	*83.5/78.78/76.29*
D-RPG	0.067/0.065/0.068	*3.78/2.53/3.97*		V1-RPG	*0.124/0.124/0.124*	*25.90/25.95/25.83*
D-RT	0.123/0.122/0.123	*25.63/24.89/24.86*		V1-RT	*0.099/0.101/0.098*	*14.76/14.72/14.84*
D-MPI	*0.075/0.073/0.069*	*0.97/0.974/0.96*		V1-MPI	*0.072/0.073/0.072*	*2.52/2.59/2.55*
KLP	*0.145/0.238/0.325*	*42.02/78.54/96.15*		V2	*0.114/0.113/0.113*	*21.37/20.83/20.81*
KLP-RPG	*0.078/0.074/0.074*	*7.23/3.40/3.45*		V2-RPG	*0.124/0.125/0.124*	*15.89/25.92/15.90*
KLP-RT	*0.121/0.119/0.118*	*24.45/23.47/23.10*		V2-RT	*0.092/0.086/0.093*	*11.64/8.61/11.88*
KLP-MPI	*0.087/0.098/0.098*	*10.35/14.29/14.19*		V2-MPI	*0.074/0.077/0.074*	*3.29/4.44/3.29*

Note: *In each cell, results represent correlation of 0.3/0.6/0.8.*

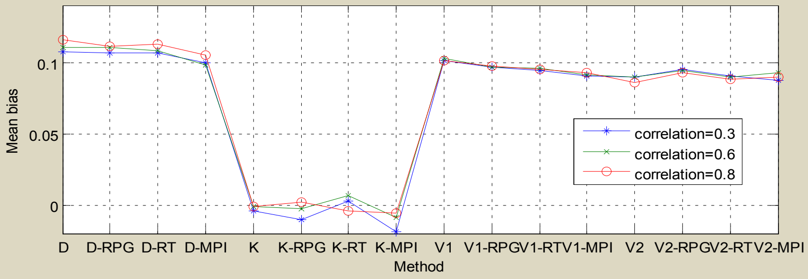

Figure 1. Mean bias of the three ability dimensions under each item selection method

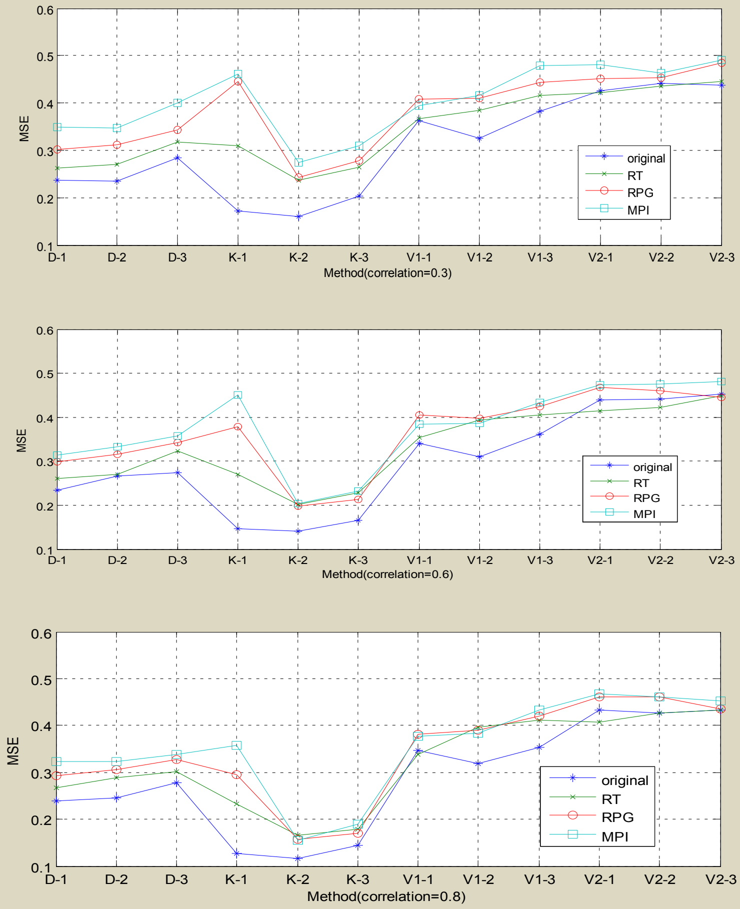

Figure 2. MSE of each ability dimension under each item selection method

Note: Original=Items Selection Index without using item exposure controlling strategies;

D=D-optimality; K=KLP; ‘-1’,’-2’, and ’-3’denote the first, second and third dimensions.

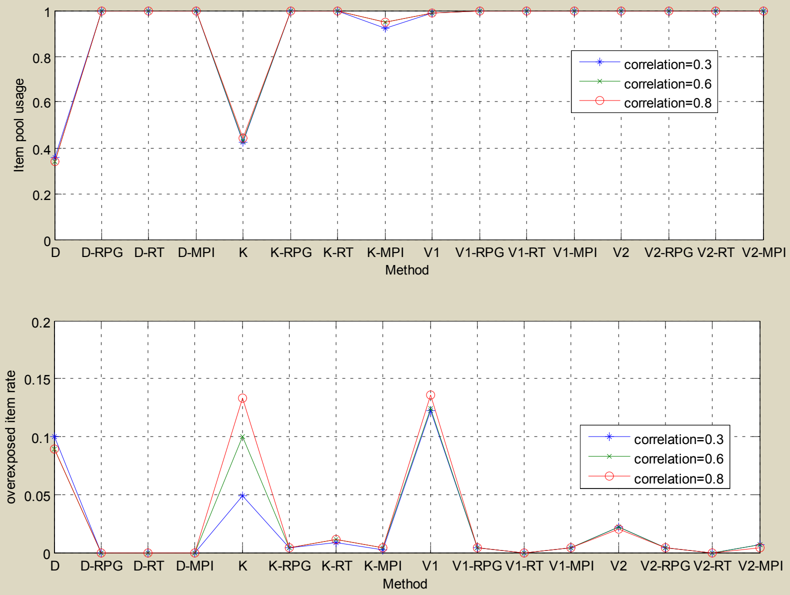

Figure 3. Item pool usage and overexposed item rates for each method under different correlations.

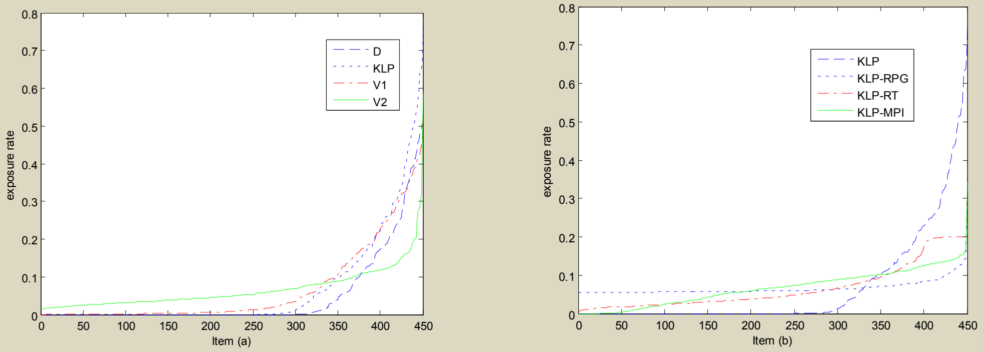

Figure 4. Item exposure rates of different methods under a correlation of 0.6 for (a) the four item selection indices without item exposure control, (b) the three item exposure control methods combined with KLP.