ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3);295-311

Investigating The Effect of Exposure-Control Strategies on Item Selection Methods in MCAT

Xiuzhen MAO * Burhanettin ÖZDEMİR ** Yating WANG*** Tao XIN****

Abstract
This study aims to investigate the effect of different item exposure controlling strategies on item selection methods in the context of multidimensional computerized adaptive testing (MCAT). Additionally, this study aims to examine to what extend the restrictive threshold (RT) and the restrictive progressive (RPG) exposure methods suppress the item exposure rates and increase the exposure rates of underexposed items without losing psychometric precision in MCAT. For this purpose, the performance of four item selection methods with and without exposure controls are evaluated and compared so as to determine how results differ when item exposure controlling strategies are applied with Monte-Carlo simulation method. The four item selection methods employed in this study are D-optimality, Kullback–Leibler information (KLP), the minimized error variance of linear combination score with equal weight (V1), the composite score with optimized weight (V2). On the other hand, the maximum priority index (MPI) method proposed for unidimensional CAT and two other item exposure control methods, that are RT and RPG methods proposed for cognitive diagnostic CAT, are adopted. The results show that: (1) KLP, D-optimality, and V1 performed better in recovering domain scores, and all outperformed V2 with respect to precision; (2) although V1 and V2 offer improved item bank usage rates, KLP, D-optimality, V1, and V2 produced an unbalanced distribution of item exposure rates; (3) all exposure control strategies improved the exposure uniformity greatly and with very little loss in psychometric precision; (4) RPG and MPI performed similarly in exposure control, and outperformed RT exposure control method.

Keywords: Multidimensional computerized adaptive testing, item selection methods, exposure control strategies.

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 compared to 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 declared 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)

* Assistant Prof. Dr., Sichuan Normal University, Sichuan-China, maomao_wanli@163.com, ORCID ID: 0000-0001-8245-3633
** Assistant Prof. Dr., Siirt University, Faculty of Education, Siirt-Turkey, b.ozdemir025@gmail.com, ORCID ID: 0000-0001-7716-2700
*** Prof. Dr., Sichuan Normal University School of Education, Sichuan-China, 1358178364@qq.com, ORCID ID: 0000-0001-9328-5380
**** Prof. Dr., Beijing Normal University, Institute of Educational Statistics and Measurement, Beijing-China, xintao@bnu.edu.cn, ORCID ID: 0000-0003-2297-2604

DOI: *10.21031/epod.407521*
Received: 21.03.2018
Accepted 30.08.2018

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 banks, such as Terra Nova (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 discrimination 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 examine the performance of some exposure control techniques along with item selection methods 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 bank 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 bank 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, there are two other exposure control methods that 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.

METHOD

A Monte Carlo 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 conditions.

Design of Simulation Study

Item bank construction: Although Stocking (1994) suggests 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 bank. For example, the item bank used by van der Linden (1999) contained 500 items while the test length was 50; Lee, et al. (2008) used an item bank of 480 items with test lengths of 30 and 60; and the item banks 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 bank 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 $N(0,1)$ and each condition is replicated for 100 times.

Examinees and item responses: All 5000 examinees were simulated uniformly from a multivariate normal distribution, as in previous researches (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:

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

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 methods with and without the three exposure control methods yields a total of 16 item selection methods.

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 estimations. The formula for bias and MSE are as follows:

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

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

To assess the effect of exposure rates, we used (a) the number of items never administered and the number of items with exposure rates greater than 0.2, (b) the χ^2 statistic, and (c) the test overlap rate. The formula χ^2 statistic is as follows:

$$\chi^2 = \sum_{i=1}^{N} \frac{(er_i - \overline{er})^2}{\overline{er}}. \tag{3}$$

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):

$$\hat{T} = \frac{M}{L} S_{er}^2 + \frac{L}{M}. \tag{4}$$

where S_{er}^2 denotes the variance of item exposure rates . Generally, smaller values of $\hat{T}$ demonstrate more balanced item utility.

In the following sections, 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.

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:

$$P_j(\vec{\theta}) = P(x_j = 1 | \vec{\theta}, \vec{a}_j, b_j) = \frac{1}{1 + \exp[-(\vec{a}_j^T \cdot \vec{\theta} - b_j)]}. \quad (5)$$

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

In this study, MAP is adopted for its competitive precision and easier computation compared to expected a posteriori (EAP) ability estimation method in MIRT. 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, while EAP estimation 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.

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 $f(\vec{\theta} | \vec{X}_{k-1})$. Based on Bayes' theorem, $f(\vec{\theta} | \vec{X}_{k-1}) \propto L(\vec{X}_{k-1} | \vec{\theta}) \cdot f(\vec{\theta})$, where $L(\vec{X}_{k-1} | \vec{\theta})$ denotes the likelihood function. Hence, the goal of MAP is to find the mode that maximizes the posterior density function $f(\vec{\theta} | \vec{X}_{k-1})$. That is, the ability estimator $\hat{\vec{\theta}}^{MAP}$ is equivalent to the solution of $\frac{\partial \log f(\vec{\theta} | \vec{X}_{k-1})}{\partial \vec{\theta}_l} = 0 \quad (l = 1,2,...,D)$. Furthermore, Newton-Raphson iteration can be used to solve this equation (for more details see, Yao, 2014b).

Item Selection Methods 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

$R_{k-1} = R \setminus \{i_1, i_2, ...i_{k-1}\}$ 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 Methods

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

$$I_j(\vec{\theta}) = P_j(\vec{\theta}) \cdot (1 - P_j(\vec{\theta})) \cdot (\vec{a}_j^T \vec{a}_j). \tag{6}$$

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

$$D_k = \max\{|\ I_{k-1}(\hat{\vec{\theta}}^{k-1}) + I_j(\hat{\vec{\theta}}^{k-1}) + \Sigma_0^{-1}|, \qquad j \in R_{k-1}\}. \tag{7}$$

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

$$KLP_k = \max\{\int_{\vec{\theta}} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) \cdot f(\vec{\theta} \mid \vec{X}_{k-1}) d\vec{\theta}, \qquad j \in R_{k-1}\}. \tag{8}$$

where

$$\begin{aligned} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) &= E_{\vec{\theta}} \log[\frac{P_j(x_j \mid \vec{\theta}, \vec{a}_j, b_j)}{P_j(x_j \mid \hat{\vec{\theta}}^{k-1}, \vec{a}_j, b_j)}] \\ &= P_j(\vec{\theta}) \log \frac{P_j(\vec{\theta})}{P_j(\hat{\vec{\theta}}^{k-1})} + (1 - P_j(\vec{\theta})) \log \frac{(1 - P_j(\vec{\theta}))}{(1 - P_j(\hat{\vec{\theta}}^{k-1}))}. \end{aligned} \tag{9}$$

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

$$KLP_k = \max\{\int_{\theta_1^{k-1} - \gamma_j}^{\theta_1^{k-1} + \gamma_j} ... \int_{\theta_D^{k-1} - \gamma_j}^{\theta_D^{k-1} + \gamma_j} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) \cdot f(\vec{\theta} \mid \vec{X}_{k-1}) d\theta_1 \cdots d\theta_D, \qquad j \in R_{k-1}\}, \tag{10}$$

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 $I_{k-1}(\hat{\vec{\theta}}^{k-1})^{-1}$.

Given equal weights $w = (1/D,\ 1/D,...,\ 1/D)$ 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:

$$w = \frac{1}{\sum_{o=1}^{D}\sum_{l=1}^{D} b_{ol}} \cdot [1,1,...,1]_{1\times D} \cdot I_{k-1}(\vec{\theta}) \quad . \tag{11}$$

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 $SEM(\vec{\theta}_\alpha)$.

Item Exposure Controlling Methods

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, β = 0.5 is favored.

The RPG method: The kth (k = 1, 2, …, L) item is selected according to formula (12) for D-optimality and KLP, and according to formula (13) for V1 and V2. These two formulas are as follows:

$$i_k = \max\{(1 - er_j / r^{\max}) \cdot [(1 - k/L)u_j + Index_j \times \beta k/L], \qquad j \in S_{k-1}\} \tag{12}$$

$$i_k = \max\{(1 - er_j / r^{\max}) \cdot [(1 - k/L)R_j + (C - Index_j) \times \beta k/L], \qquad j \in S_{k-1}\}, \tag{13}$$

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

$$PI_j = \frac{r^{\max} - n_j / N}{r^{\max}} \cdot Index_j, \quad (14)$$

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

$$PI_j = \frac{r^{\max} - n_j / M}{r^{\max}} \cdot (C - Index_j). \quad (15)$$

RESULTS

Results of Ability Estimation

The ability estimations obtained from different MCAT algorithms were compared with respect bias and MSE statistics. Figure 1 depicts mean bias of the three ability dimensions under each item selection method and item exposure control methods with differing correlation between dimensions.

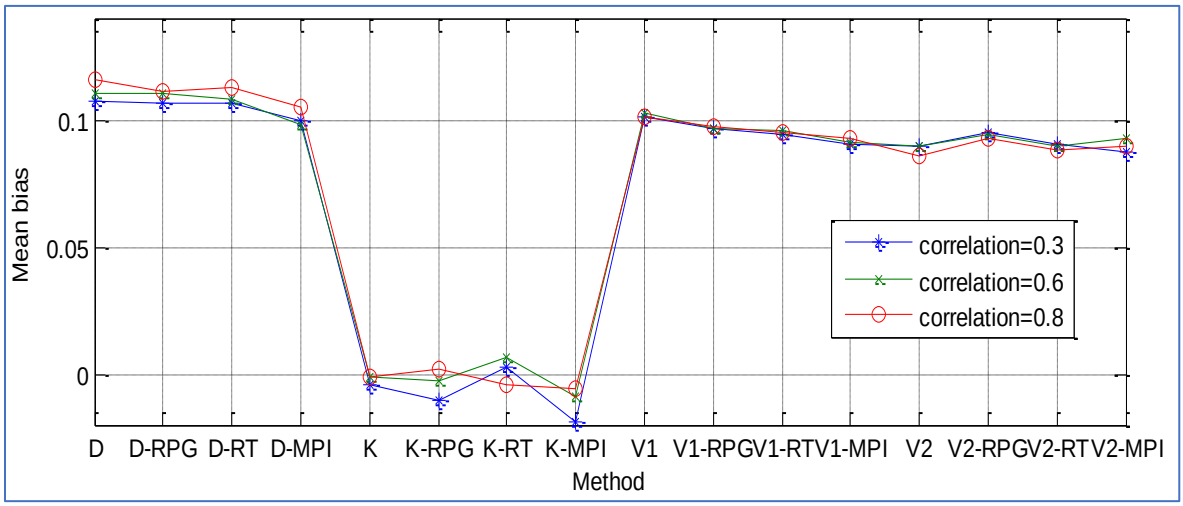


Figure 1. Mean Bias of the Three Ability Dimensions Under Each Item Selection Method

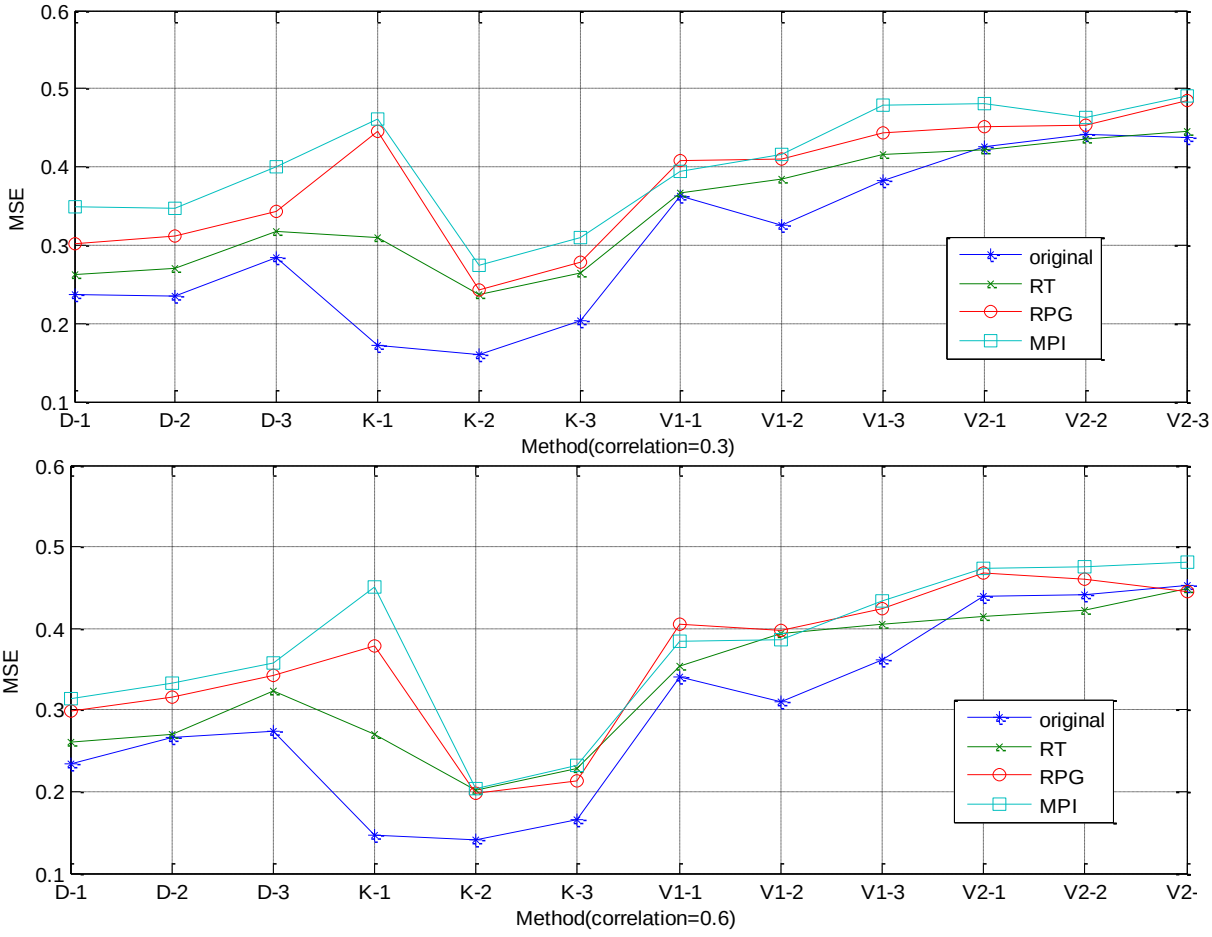
Figure 1 shows that the differences in bias between two arbitrary dimensions of each method were negligible regardless of item selection and exposure control methods. Moreover, one can observe from Figure 1 that the bias associated with D-optimality, V1, and V2 were similar, while greater than the bias produced by KLP which indicates that KLP outperformed other item selection method and effect of item exposure controlling methods on KLP and other ability estimation methods were negligible small.

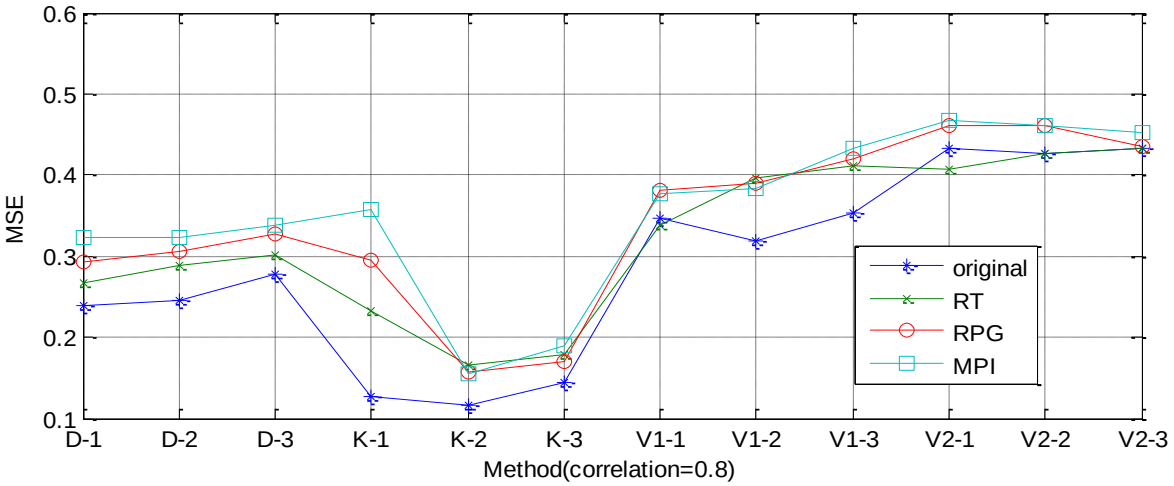
Figure 2 presents the distribution of the MSEs of each ability dimension across the different item selection and exposure controlling methods at each correlation level.

MSE statistics provided in Figure 2 shows that, for each dimension, KLP produces the smallest MSE and it was followed by D-optimality, V1, and V2. Generally, it is easy to sort the item selection methods into descending order of KLP, D-optimality, V1, and V2 according to their measurement precision. All three item exposure strategies led to an increase in MSE except for V2 item selection method. 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 bank utility. Overall, measurement precision tends to decrease when an exposure controlling method is employed

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

Finally, the RT exposure controlling method always produced the lowest MSE values, thus, giving higher measurement precision compared to RPG and MPI. Although their precision under different item selection indices varied to some degree, RPG and MPI performed similarly. 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, respectively.





(Note: Original=items selection methods without item exposure controlling strategies; D=D-optimality; K=KLP; '- 1','-2', and '-3'denote the first, second and third dimensions)

Figure 2. MSE of Each Ability Dimension Under Different Item Selection and Exposure Controlling Methods

Results of Item Exposure Rates

The item exposure rates and chi-square statistics associated with each item selection method with and without exposure controlling were presented in Table 1 and distribution of these statistics across different conditions were depicted in Figure 3 and Figure 4, respectively.

First, it is easy to infer from Table 1 that the exposure rates were distributed unevenly for D-optimality, KLP, V1, and V2. For instance, D-optimality and KLP yielded the largest test overlap and overexposed item rates and the lowest item bank usage rates which were depicted in Figure 3. Although the number of never-reached items in V1 and V2 was close to 0, and the test overlap rates and χ^2 values were smaller than those of D-optimality and KLP, yet, these exposure rate control methods still produced 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).

Table 1. Item Exposure Statistics Associated with Each Method

Item selection method	Exposure controlling method	*Overlap rate*			χ^2		
		r=.30	*r*=.60	*r*=0.80	*r*=.30	*r*=.60	*r*=0.80
D-Optimality	without exposure controlling	0.408	0.23	0.23	152.6	75.14	75.14
	RPG	0.067	0.065	0.068	3.78	2.53	3.97
	RT	0.123	0.122	0.123	25.63	24.89	24.86
	MPI	0.075	0.073	0.069	0.97	0.974	0.96
KLP	without exposure controlling	0.145	0.238	0.325	42.02	78.54	96.15
	RPG	0.078	0.074	0.074	7.23	3.40	3.45
	RT	0.121	0.119	0.118	24.45	23.47	23.10
	MPI	0.087	0.098	0.098	10.35	14.29	14.19
V1	without exposure controlling	0.253	0.241	0.237	83.5	78.78	76.29
	RPG	0.124	0.124	0.124	25.90	25.95	25.83
	RT	0.099	0.101	0.098	14.76	14.72	14.84
	MPI	0.072	0.073	0.072	2.52	2.59	2.55
V2	without exposure controlling	0.114	0.113	0.113	21.37	20.83	20.81
	RPG	0.124	0.125	0.124	15.89	25.92	15.90
	RT	0.092	0.086	0.093	11.64	8.61	11.88
	MPI	0.074	0.077	0.074	3.29	4.44	3.29

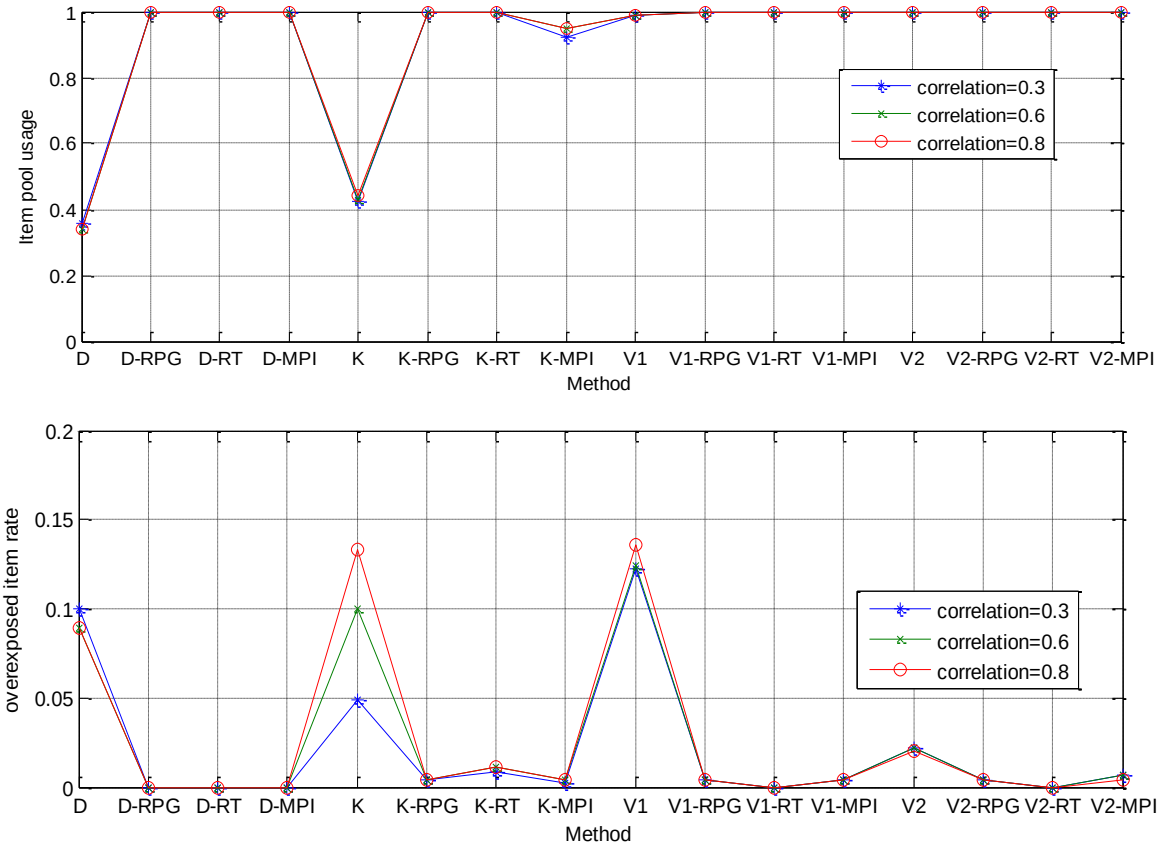


Figure 3. Item Bank Usage and Overexposed Item Rates for Each Method Under Different Correlations.

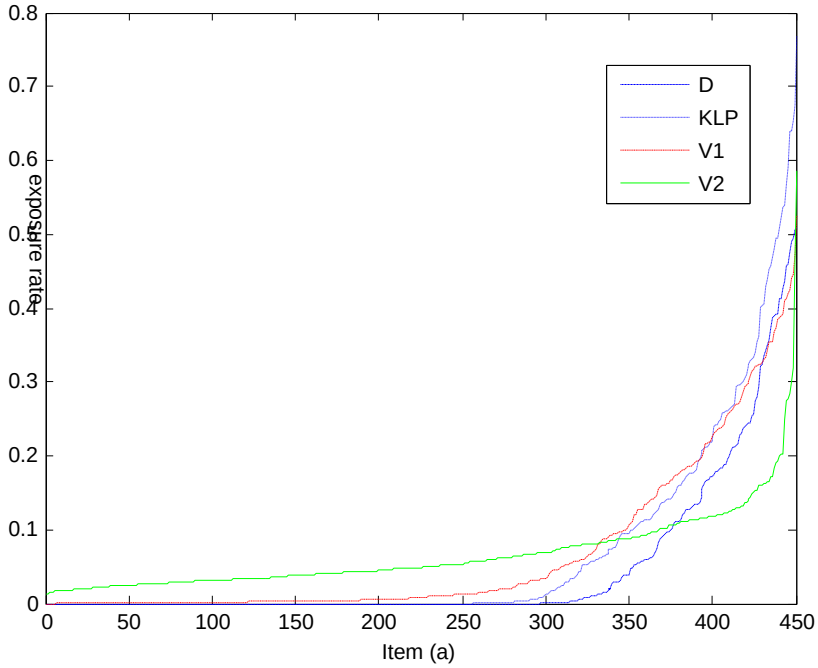
Second, all the exposure control methods improved the uniformity of exposure rates substantially in terms of increasing item bank usage and decreasing the overexposed item rates, test overlap rates, and χ^2 statistics. Although MPI performed similarly, RPG outperformed the other methods in most cases. It is apparent that all the item exposure distributions followed the same pattern when different item selection indices were 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. One can observe from Figure 3 that the item bank usage rate reaches 100% for all methods except KLP-MPI condition. In other words, all item exposure methods improve the item bank usage substantially. Checking the overexposed items, both RPG and MPI produced more overexposed items than RT under most test conditions. Generally, RT was able to control the item exposure rates to be lower than the allowable maximum value, whereas both RPG and MPI resulted 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 everyone taking the test. 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.

4(a) the four item selection indices without item exposure control



4(b) the three item exposure control methods combined with KLP.

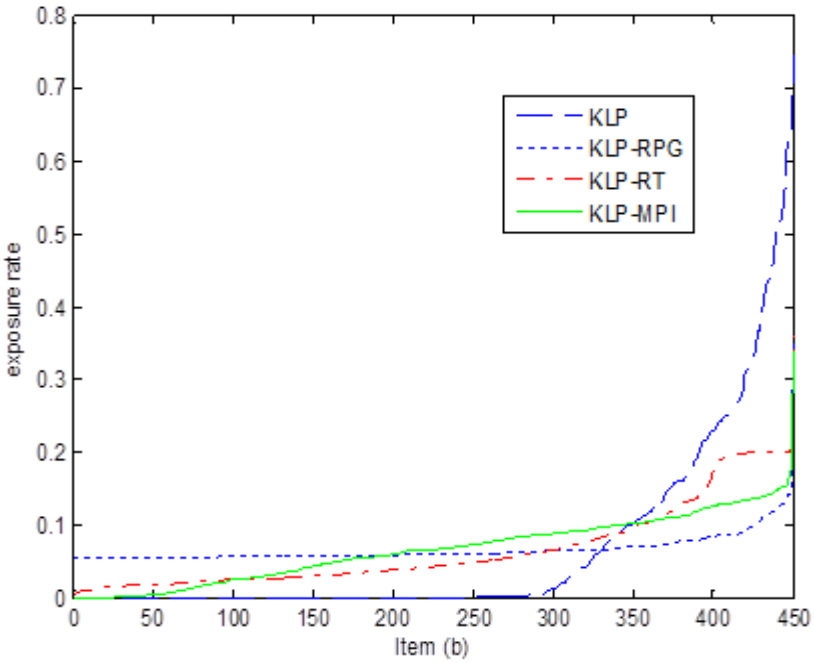


Figure 4. Item Exposure Rates of Different Methods Under the Correlation of 0.6

Overall, although the item exposure control strategies produced different patterns of item exposure rates, they all considerably improved the balance of the item exposure distribution. This can be seen from comparing Figure 4(a) and 4(b). In addition, one can infer from the results that there appear to be trade-off between the measurement precision and employing the item exposure controlling methods.

CONCLUSIONS AND DISCUSSIONS

Many studies have acknowledged the advantages of CAT over P&P tests and computer-based tests with respect to 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 methods combined with different exposure control methods in MCAT. Simulations showed that V2 outperformed 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 methods without using item exposure control were unsatisfactory with respect to item exposure statistics. The results also indicate that without using item exposure control, the item selection indices could 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 for all item selection method.

When the item exposure rate distribution obtained from different item exposure control methods were compared, the RPG and MPI outperformed the other methods in most cases, while the RT method showed the worst performance. Furthermore, each item exposure control method yielded the same exposure rate pattern under different item selection methods. When it comes to comparing the measurement precision, the performance of the different exposure control methods could be ordered as RT, RPG, and MPI. This kind of trade-off between measurement precision, utility of item bank, and evenness of item exposure rate has been observed in many studies (Chang & Twu, 1998). In 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 examine how to deal with non-statistical constraints in MCAT.

Third, polytomous items such as essay-type and constructed-response 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(4),* doi:522-532. 10.1037/0021-9010.76.4.522

Bloxom, B. M., & Vale, C. D. (1987, June). *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. doi: 10.1177/0146621603258350

Chang, S. W., & Twu, B. Y. (September 1998). *A comparative study of item exposure control methods in computerized adaptive testing*. ACT Research Report Series, ACT-RR-98-3.

Chang, H.: H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222. doi: 10.1177/01466210122032181

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

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. doi:10.1348/000711008X304376

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:0.1111/j.1745-3984.2009.01070.x

Huebner, A. R., Wang, C., Quinlan, K., & Seubert, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods, 48*(4), 1443-1453. 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. doi:10.1177/0013164407307007

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404. doi:10.1177/014662169602000406

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. http://www.dtic.mil/dtic/tr/fulltext/u2/a125099.pdf

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria. *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(77-101)*. Springer, New York.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354. doi: 10.1007/BF02294343

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 94–5). Princeton, NJ: Educational Testing Service. doi:10.1002/j.2333-8504.1994.tb01578.x. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1994.tb01578.x

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(4)*, 398-412. doi:10.3102/10769986024004398

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(4)*, 398-418. doi: 10.3102/1076998606298044

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67(4)*, 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(3)*, 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(1)*, 13-39. DOI: 10.1007/s11336-010-9186-0

Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99-122. DOI: 10.1177/0146621612463422

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48(3)*, 255-273. DOI: 10.1111/j.1745-3984.2011.00145.x

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychologica Measurement, 28(5)*, 295–316. DOI: 10.1177/0146621604265938.

Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement, 47(3)*, 339-360. doi:10.1111/j.1745-3984.2010.00117.x

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77 (3)*, 495-523. doi: 10.1007/s11336-012-9265-5.

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(1)*, 18-38. doi:10.1111/jedm.12032

Yao, L. (2014b). Multidimensional item response theory for score reporting. In Cheng, Y. & Chang, H.-H. (Eds.), *Advancing methodologies to support both summative and formative assessments (147-182)*. 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(8)*, 614-631. doi:10.1177/0146621614541514

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(6)*, 3-23. doi:10.1177/0146621605284537

Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Testlerde Madde Kullanım-Sıklığı Yöntemlerinin Madde Seçim Yöntemleri Üzerindeki Etkisinin Incelenmesi

Giriş

Binlerce öğrencinin aynı oturumda aynı sorulara cevap verdiği geleneksel test yöntemine alternatif olarak, öğrencilerin yetenek düzeyleri ile madde özelliklerinin bilgisayar ortamında eşleştirildiği bilgisayar ortamında bireyselleştirilmiş test yöntemleri her geçen gün yaygınlaşmaktadır. Bireyselleştirilmiş test uygulamalarının yaygınlaşmasında, geleneksel kâğıt kalem testlerine göre, uygulanmasının daha az zaman alması, testteki madde sayısını önemli ölçüde azaltması ve test biter bitmez bireye dönüt verebilmesi gibi faktörlerin etkili olduğu söylenebilir. Bireyselleştirilmiş

testlerin bir diğer avantajı ise tek boyutlu, çok boyutlu madde tepki kuramları (MTK) veya bilişsel tanı modelleri gibi farklı ölçme modellerinin (measurement models) kullanılmasına olanak sağlamasıdır. Farklı ölçme modellerinin kullanılmasına olanak sağlaması hem model-veri uyumunun incelenmesi hem de farklı puanlama yöntemlerinin kullanılmasına olanak sağlaması acısından önemli görülmektedir.

Çok boyutlu bilgisayar ortamında bireyselleştirilmiş testler ise hem çok boyutlu MTK modellerinin kullanılmasına olanak sağlaması hem de bireyselleştirilmiş olması acısından diğer yöntemlere göre avantajlı görülmektedir. Diğer taraftan farklı madde ve test seçme algoritmalarının kullanıldığı bireysel testlere ilişkin yapılan birçok çalışmada, çok boyutlu bireyselleştirilmiş testlerin tek boyutlu bireyselleştirilmiş testlere göre daha avantajlı olduğunu vurgulamaktadır. Örneğin, Segall (1996) gerçek verilere dayalı yapmış olduğu simülasyon çalışmasında tek boyutlu bireyselleştirilmiş test uygulamaları ile karşılaştırıldığında, çok boyutlu bireyselleştirilmiş testlerin test uzunluğunun üçte-bir oranında daha az olduğu ve benzer veya daha yüksek güvenirlik katsayılarına sahip olduğu bulgusuna ulaşmıştır. Luecht (1996) Yapmış olduğu çalışmada çok boyutlu bireyselleştirilmiş testlerin test uzunluğunu %25 ile %40 oranında azalttığını belirtmiştir. Ayrıca çok boyutlu modeller öğrencinin birden fazla yeteneğinin aynı anda ölçülmesine olanak sağladığından bireyin ölçülen yeteneği hakkında daha fazla bilgi sağlamaktadır. Bundan dolayı bazı geniş ölçekli test uygulamalarında tek boyutlu bireyselleştirilmiş test yerine çok boyutlu bireyselleştirilmiş testler kullanılmaktadır. Nitekim Terra Nova (Yao, 2010), American College Testing (ACT) (Veldkamp & van der Linden, 2002) ve ASVAB (Segall, 1996; Yao, 2012, 2014a) gibi testlerde gerçek madde havuzları kullanılarak çok boyutlu bireyselleştirilmiş test yöntemleri kullanılmıştır.

Çok boyutlu bireyselleştirilmiş test uygulamalarında güvenilir ve geçerli sonuçlar elde edilebilmesi ve başarılı bir şekilde uygulanabilmesinde madde secim yöntemleri önemli bir yere sahiptir (Wang & Chang, 2011). Fakat güvenilir ve geçerli sonuçlar vermelerine karşın bazı maddelerin sık uygulanması (overexposed items) veya az uygulanması (underexposed items) problemlerini çözmede yetersiz kalmaktadırlar. Bu probleme bir çözüm olarak farklı madde kullanım sıklığı yöntemleri geliştirilip, madde seçim yöntemleri ile birlikte uygulanmaya başlanmıştır.

Bu araştırmada çok boyutlu bireyselleştirilmiş testlerde kullanılan farklı madde kullanım sıklığı kontrol yöntemlerinin madde seçim yöntemleri üzerindeki etkisinin incelenmesi amaçlanmaktadır. Ayrıca, bu çalışmada madde kullanım sıklığı kontrol yöntemlerinden restrictive threshold (RT) ve restrictive progressive (RPG) yöntemlerinin madde kullanım sıklığı oranını ve diğer maddelere göre daha az uygulanan maddelerin kullanım sıklığını nasıl etkilediği incelenmiştir.

Yöntem

Bu çalışmada Monte Carlo simülasyon yöntemi ile dört farklı madde secim yönteminin farklı madde kullanım sıklığı yöntemlerinin kullanıldığı ve kullanılmadığı durumlardaki performansları karşılaştırılmıştır. Çok boyutlu MTK ya dayalı modellerin kullanıldığı simülasyon çalışmalarında genellikle boyut olarak iki veya üç boyut, madde ve yetenek parametresini kestirmek için ise çok boyutlu modellerden ise 2 parametreli veya 3 parametreli MTK modelleri tercih edildiği görülmektedir. (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). Bu simülasyon çalışmasında madde ve yetenek parametrelerinin simülasyonunda 2-parametreli MTK modelleri kullanılmış ve testler üç boyuttan oluşmaktadır. Özellikle madde havuzunda yer alan 450 maddeye ait ayırt edicilik parametreleri (a_{j1}, a_{j2}, a_{j3}) logaritmik normal dağılımdan üretilirken ($\log N(0, 0.5)$) madde güçlük parametreleri ise standart normal dağılımdan ($N(0,1)$) üretilmiştir. Her bir test için örneklem büyüklüğü 5000 olarak belirlenmiş ve bireylerin maddelere verdiği cevaplar çok değişkenli normal dağılımdan üretilmiştir. Nitekim daha önceki çalışmalarda benzer simülasyon koşulları kullanılmıştır (Wang & Chang, 2011; Yao, Pommerich, & Segall, 2014; Wang et al., 2013).

Bu çalışmada, madde secim yöntemlerinden, D-optimality, Kullback–Leibler bilgi yontemi (Kullback–Leibler information-KLP), V1 (the minimized error variance of linear combination score with equal weight) ve V2 (the composite score with optimized weight) yöntemleri kullanılmıştır. Ayrıca, madde kullanım sıklığını kontrol etmek amacıyla tek boyutlu bireyselleştirilmiş testler için geliştirilen MPI (the maximum priority index) ve bilişsel tanı modelleri için geliştirilen RT ve RPG yöntemleri kullanılmıştır. Test sürecinde yetenek parametrelerinin kestirilmesi ve güncellenmesi için Bayesyen yetenek kestirim yöntemlerinden MAP (maximum a posteriori) yöntemi kullanılmıştır. Belirlenen her bir koşul için 100 tekrar yapılmıştır.

Yukarıda belirtilen farklı çok boyutlu bireyselleştirilirmiş test koşullarından elde edilen yetenek parametrelerini karşılaştırmak için yanlılık ve standart hata ortalamaları hesaplanmıştır. Madde kullanım sıklığı yöntemlerinin etkisini incelemek için ise her bir koşula ait (a) hiç uygulanmayan madde sayısı (b) kullanım sıklığı oranı 0,2`den yüksek madde sayısı (c) ki-kare istatistiği ve (d) çakışma oranı (test overlap) istatistikleri kullanılmıştır.

Sonuç ve Tartışma

Bu çalışmada dört farklı madde seçim yöntemi ile birlikte farklı madde kullanım sıklığı yöntemlerinin kullanıldığı çok boyutlu bireyselleştirilmiş testlerin performansları karşılaştırılarak, madde kullanım sıklığı yöntemlerinin madde seçim yöntemleri üzerindeki etkisi incelenmiştir. Araştırma sonucunda, V2 madde secim yönteminin madde havuzu kullanım oranı, sık uygulanan madde oranı ve testlerdeki madde çakışma oranı açısından diğer madde secim yöntemlerine göre daha iyi sonuç verdiği bulgusuna ulaşılmıştır. Buna karşın, genel olarak, dört madde seçim yönteminin de madde kullanım sıklığı istatistikleri açısından yetersiz olduğu söylenebilir.

Madde kullanım sıklığı oranlarının dağılımı incelendiğinde, RT madde kullanım sıklığı kontrol yöntemine göre, RPG ve MPI yöntemlerinin daha iyi sonuç verdiği görülmektedir. Diğer taraftan, madde kullanım sıklığı yöntemlerinin diğer madde seçim yöntemleri ile birlikte uygulandığında maddelerin kullanım sıklığı oranı dağılımlarının benzer olduğu bulgusuna ulaşılmıştır. Ölçmenin kesinliği (measurement precision) istatistiklerine göre karşılaştırıldığında, RT yönteminin en yüksek güvenirliğe sahip olduğu ve bunu RPG ve MPI yöntemlerinin takip ettiği görülmektedir. Bu sonuçlara göre madde havuzu kullanımı ve madde kullanım sıklığı oranlarının eşitliğinin sağlanması için madde kullanım sıklığı kontrol yöntemleri uygulandığında, ölçmenin kesinliğinde belli oranda düşüşün olacağı gerçeğinin göz önünde bulundurulması gerekir (Chang & Twu, 1998). Diğer bir değişle madde kullanım sıklığı oranını istenilen düzeyde tutmak ölçmenin kesinliğinde belirli bir düzeyde düşüşü göze almayı gerektirir.

Bu çalışmada maddelerin ikili puanlandığı (0,1) çok boyutlu bireyselleştirilmiş testlerde farklı madde kullanım sıklığı yöntemlerinin madde seçim yöntemleri üzerindeki etkisi incelenmiştir. Benzer koşulların farklı madde türlerinden oluşan (örneğin çoklu puanlanan maddeler) bireyselleştirilmiş testlerde de incelenmesi önerilmektedir. Ayrıca bu çalışma çok boyutlu bireyselleştirilmiş testlerde kullanılan madde seçim ve madde kullanım sıklığı yöntemleri ile sınırlıdır. Farklı yetenek kestirim yöntemleri ve durdurma kurallarının uygulandığı test koşullarının çok boyutlu bireyselleştirilmiş testler üzerindeki etkisinin incelenmesi önerilmektedir.