Article

Applied Psychological Measurement
2016, Vol. 40(5) 346–360
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616639305
apm.sagepub.com
SAGE

A Comparison of Constrained Item Selection Methods in Multidimensional Computerized Adaptive Testing

Ya-Hui Su[1]

Abstract

The construction of assessments in computerized adaptive testing (CAT) usually involves fulfilling a large number of statistical and non-statistical constraints to meet test specifications. To improve measurement precision and test validity, the multidimensional priority index (MPI) and the modified MPI (MMPI) can be used to monitor many constraints simultaneously under a between-item and a within-item multidimensional framework, respectively. As both item selection methods can be implemented easily and computed efficiently, they are important and useful for operational CATs; however, no thorough simulation study has compared the performance of these two item selection methods under two different item bank structures. The purpose of this study was to investigate the efficiency of the MMPI and the MPI item selection methods under the between-item and within-item multidimensional CAT through simulations. The MMPI and the MPI item selection methods yielded similar performance in measurement precision for both multidimensional pools and yielded similar performance in exposure control and constraint management for the between-item multidimensional pool. For the within-item multidimensional pool, the MPI method yielded slightly better performance in exposure control but yielded slightly worse performance in constraint management than the MMPI method.

Keywords

CAT, priority index, multidimensional, item selection, constrained, IRT

Computerized adaptive testing (CAT) has been widely used in educational and psychological assessments because it can obtain efficient and precise ability estimation with fewer items than traditional paper-and-pencil tests. One of the important issues in CAT is the item selection algorithm. Test specifications specify a series of constraints for including items in a test (Swanson & Stocking, 1993). The constraints can be both statistical (i.e., psychometric) and non-statistical (i.e., non-psychometric) on item properties. Examples of the statistical constraints include target item and test information, whereas examples of the non-statistical constraints include content

[1] National Chung Cheng University, Chiayi, Taiwan

Corresponding Author:
Ya-Hui Su, Department of Psychology, National Chung Cheng University, 168 University Road, Minhsiung Township, Chiayi County 62102, Taiwan.
Email: psyyhs@ccu.edu.tw

specifications and key balancing. Although different algorithms perform item selection sequentially or simultaneously in the test assembly, the item selection in CAT is sequential by nature (van der Linden, 2005). Therefore, it is challenging, while constructing assessments, to meet the various constraints in CAT simultaneously.

Many item selection methods have been proposed to handle constraints in CAT; in general, these methods can be classified into mathematical programming approaches and heuristic approaches (Cheng & Chang, 2009). The mathematical programming approaches, such as the network-flow programming method (Armstrong, Jones, & Kunce, 1998) and the shadow-test approaches (van der Linden, 2000, 2005; van der Linden & Chang, 2003; van der Linden & Reese, 1998; van der Linden & Veldkamp, 2004, 2007; Veldkamp & van der Linden, 2002, 2008), are very effective in managing constraints, but computation may be intensive when a large number of constraints are considered. Another important issue is feasibility; the mathematical programming approaches can have a solution only when a test can satisfy all the constraints (Timminga, 1998). In addition, the mathematical programming approaches often rely on external commercial software, such as CPLEX and LINDO (Chang, 2007).

However, the heuristic approaches, such as the weighted deviation modeling (WDM; Stocking & Swanson, 1993) method and the priority index (PI) approaches (Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009; Su, 2015; Su & Huang, 2015; Yao, 2011, 2012, 2013, 2014), can avoid the issues of computational intensity and infeasibility. Because the heuristic approaches can be implemented easily and computed efficiently, they are widely used in operational CAT (Cheng & Chang, 2009). In addition, the heuristic approaches make the item selection process very fast and the algorithms can always find a solution. A drawback of the heuristic approaches is that the assembled tests may not be ''optimal'' nor meet all the constraints because items are selected sequentially. In practice, the WDM method can satisfy most of the non-statistical constraints. To achieve optimal results, the WDM method would need to adjust the weights for constraints through a remarkably time-consuming process (Leung, Chang, & Hau, 2005). In contrast, the PI approaches do not require adjusting weights for constraints (Cheng & Chang, 2009). The PI approaches can be used to monitor many non-statistical constraints simultaneously and efficiently in unidimensional CAT (Cheng & Chang, 2009; Cheng et al., 2009) and multidimensional CAT (Su, 2015; Su & Huang, 2015; Yao, 2011, 2012, 2013, 2014).

Several studies were conducted to compare the performance of some of the methods discussed above (Cheng & Chang, 2009; He, Diao, & Hauser, 2014; Robin, van der Linden, Eignor, Steffen, & Stocking, 2005; van der Linden, 2005), but the findings were inconclusive. Robin et al. (2005) compared the shadow-test approaches and the WDM method, and found that two methods performed in a comparable manner; van der Linden (2005) also compared these two methods, but he found that the WDM yielded some violations, larger bias, and inaccurate ability estimators. Cheng and Chang (2009) compared the PI and the WDM item selection methods, and found that the PI method yielded fewer constraint violations and better exposure control than the WDM method while maintaining the same level of measurement precision. He et al. (2014) compared the shadow-test approaches and three heuristic approaches, the WDM, the PI, and the weighted penalty model (Shin, Chien, Way, & Swanson, 2009). They found that the shadow-test approaches yielded the best performance in terms of measurement accuracy and constraint management, and the three heuristic approaches performed in a comparable manner with one another.

Many educational and psychological tests, such as the Minnesota Multiphasic Personality Inventory–2 (MMPI-2; Hathaway & McKinley, 1989), have been analyzed with multidimensional models. Multidimensional CAT can provide higher precision and reliability or reduce test length, compared with unidimensional CAT (Segall, 1996; Wang & Chen, 2004; Yao, 2012).

However, there are difficulties in the use of multidimensional CAT in practice, one of which is time-consuming to obtain ability estimates for high dimensional structure. Therefore, more studies on multidimensional CAT are needed. The multidimensional priority index (MPI; Yao, 2011, 2012, 2013, 2014) method was developed to handle constraints for the Armed Services Vocational Aptitude Battery (ASVAB), which is a between-item multidimensional CAT test. Under the between-item multidimensional framework, items in the same battery are assumed to measure only one distinct latent trait, and the overall assessment is assumed to measure different latent traits. In contrast, some other tests might have a within-item multidimensional structure, such that individual items are intended to assess multiple latent traits. For instance, an arithmetic item in a mathematics test can be used to assess both symbolic representation and calculation, that is, this mathematics test has a within-item multidimensional structure.

Su and Huang (2015) argued that the MPI method was not appropriate for item selection in within-item multidimensional CAT. To extend the MPI method to a within-item multidimensional framework, the modified MPI (MMPI; Su & Huang, 2015) method was proposed to handle various constraints for item selection. In Su and Huang's study, some items assessed multiple latent traits and some items assessed single latent traits to form a within-item multidimensional pool. The proposed MMPI method was compared with several item selection methods, and it could accommodate various non-statistical constraints simultaneously. Since Yao (2013) found that high-quality items tend to be administered to examinees who take the test earlier and suggested that *a*-stratification (Chang & Ying, 1999; Chang, Qian, & Ying, 2001) can be integrated with the MPI method, administering ''low-information'' items first. The rationale is that less discriminating items can be used at the initial stage of testing when the latent trait estimation is not reliable, and high discriminating items can be used at the later stages of testing when the latent trait estimation is of greater certainty. Therefore, the MMPI method was integrated with *a*-stratification and exposure control (Su & Huang, 2015) to obtain better pool usage and lower test overlap rates; however, it yielded some slight loss in measurement precision and constraint management.

The MPI method was developed for item selection in multidimensional CAT and showed its advantages in between-item multidimensional CAT (Yao, 2011, 2012, 2013, 2014). Because Su and Huang (2015) argued that the MPI method was not appropriate in within-item multidimensional CAT, the MMPI method was proposed and only investigated under the within-item multidimensional framework. In practice, the MMPI method can be used for item selection in both between-item and within-item multidimensional CATs (Su & Huang, 2015). In addition, as the MPI and MMPI methods can be implemented easily and computed efficiently, they are important and useful for operational CAT; however, no thorough simulation study has been done to compare the performance of these two item selection methods under two different item bank structures. Therefore, the purpose of this study was to compare the efficiencies of the MMPI and MPI methods for item selection in between-item and within-item multidimensional CATs.

The MPI Method

Yao (2011) defined the MPI for each item i as

$$\mathrm{MPI}_i = \prod_{d=1}^{D} f_{id}^{c_{id}}, \tag{1}$$

where c_{id} is the loading information for item i on domain d such that $c_{id} = 1$ if item i is from domain d and $c_{id} = 0$ otherwise. To apply a stopping rule of measurement precision, Yao (2013) included the estimated domain score precision, item exposure rate, and content constraints with upper and lower bounds for each domain to define f_{id} as

$$f_{id}=\left[\max\left\{\left[1-\left(\frac{p_d}{\widehat{p}_d}\right)^a+\varepsilon_1\right],0\right\}\right]\left[\max\left\{\left(\frac{r_i-n_i/N}{r_i}\right),0\right\}\right]$$
$$\left[1_{x_d\leq l_d}\left(\frac{l_d-x_d}{l_d}+\varepsilon_2/x_d\right)+1_{x_d>l_d}\max\left\{1-\left(\frac{x_d}{u_d}\right)^b,0\right\}\right], \tag{2}$$

where p_d and $\hat{p}_d$ are the required standard error of measurement (SEM) and the SEM estimates based on the administered items for the domain d ability estimates, respectively; the larger the SEM, the smaller the precision. The second term in Equation 2 is used to ensure that no item is selected more than a pre-specified item exposure rate r_i. N is the total number of examinees. For each item selection step, n_i is the number of examinees who have seen item i. The third term in Equation 2 is used to monitor content specifications. The lower and upper bounds of each domain d are l_d and u_d, respectively. For each item selection step, x_d is the number of selected items from domain d. To consider item selection criteria or information measures in multidimensional CATs, the MPI in Equation 1 can be modified by multiplying the minimum angle (Reckase, 2009), maximum volume or the maximum determinant of the Fisher information matrix (MDFIM; Segall, 1996), minimum error variance of the linear combination (van der Linden, 1999), minimum error variance of the composite score with the optimized weight (Yao, 2010), or Kullback–Leibler information (Chang & Ying, 1996). Then, an item with the largest value will be selected for administration. According to the algorithm, no further items will be selected for a specific constraint if the constraint is met. Here, the smaller the values of a and b, the larger the weight given to the precision. The terms ε_1 and ε_2 are small numbers, so that the precision of the estimates can be slightly above the required precision, and the minimum required number of items for each domain can be administered first, respectively.

Su and Huang (2015) argued that the MPI method was developed to handle constraints for the between-item multidimensional test, so it might not be appropriate to be used directly in a within-item multidimensional framework. When an arithmetic item is used to assess symbolic representation and calculation, the terms regarding measurement precision, exposure control, and content balancing in Equation 2 are calculated within each dimension. For the arithmetic item, the MPI needs to be calculated over symbolic representation and calculation dimensions. Obviously, an item with within-item multidimensional structure is unlikely to be selected because many multiplications make the MPI much smaller.

The MMPI Method

Based on the PI framework, Su and Huang (2015) extended the MPI method to a within-item multidimensional framework. The MMPI for each item i is defined as

$$\mathrm{MMPI}_i=Inf_i\times\prod_{k=1}^{j}\omega_k f_k^{c_{ik}}\times\sqrt{\sum_{k=j+1}^{K}\left[\omega_k c_{ik} f_k\right]^2}, \tag{3}$$

where $c_{ik}=1$ represents constraint k being relevant to item i and $c_{ik}=0$ otherwise. Each constraint k is associated with a weight w_k. Different constraints are given different weights depending on their importance. The item with the largest MMPI will be selected for administration. The first term in Equation 3 is the item information criterion, which the determinant of the Fisher information matrix was used in Su and Huang's study. The second term in Equation 3 includes the constraints between dimensions, such as item exposure control and key balancing. For a unidimensional or between-item multidimensional pool, constraints with regard to content balancing are included in the second term. When only the first two terms in Equation 3 are included, it is reduced to the PI item selection method in unidimensional CAT. For a within-item multidimensional pool, constraints considering within dimensions, such as content balancing, are included in the third term of Equation 3.

When constraints are considered with regard to flexible content balancing, each flexible content balancing constraint involves a lower bound l_k and an upper bound u_k. The number of items to be selected from content area k is denoted as μ_k. Then, $l_k \leq \mu_k \leq u_k$ and $\sum_{k=1}^{K} \mu_k = L$, where L is test length. A one-phase item selection strategy can be used by incorporating both upper and lower bounds. The term f_k in Equation 3 can be replaced with $f_{1k}\, f_{2k}$, defined as

$$f_{1k} = \frac{1}{u_k}(u_k - x_k) \tag{4}$$

and

$$f_{2k} = \frac{(L - l_k) - (t - x_k)}{L - l_k}, \tag{5}$$

where t is the number of items that have already been administered and $t = \sum_{k=1}^{K} x_k$. The terms f_{1k} and f_{2k} measure the closeness to the upper and lower bounds, respectively. When the f_{2k} in Equation 5 is equal to 0, it indicates that the sum of items from other domains has reached its maximum. Then, the $f_{1k}\, f_{2k}$ for constraint k is defined as 1 when $f_{2k} = 0$. When item exposure control is considered, the term f_k can be calculated as

$$f_k = \frac{1}{r_{\max}}\left(r_{\max} - \frac{n_i}{N}\right), \tag{6}$$

where N is the number of examinees who have taken the CAT, and n is the number of examinees who have seen item i. After N examinees have taken the CAT, the n_i/N is the provisional exposure rate of item i.

Method

Item Pool

The multidimensional three-parameter logistic (M3PL; Reckase, 1985) model was used in this study. The probability of getting a correct response for examinee n with d-dimensional latent traits $\boldsymbol{\theta}'_n = (\theta_1, \theta_2, \ldots, \theta_d)$ is defined as

$$p_{ni1} = c_i + (1 - c_i)\frac{\exp[\mathbf{a}_i'(\boldsymbol{\theta}_n - b_i\mathbf{1})]}{1 + \exp[\mathbf{a}_i'(\boldsymbol{\theta}_n - b_i\mathbf{1})]}, \tag{7}$$

where $\mathbf{a}_i$ is a $d \times 1$ vector of the discrimination parameter of item i; b_i and c_i are the difficulty and the guessing parameters of item i, respectively; and $\mathbf{1}$ is a $d \times 1$ vector of 1s.

The item parameters in this study were adapted from Su and Huang's study (2015). Two simulated pools were used in the study. The discrimination parameters were drawn from a uniform distribution at the interval of real numbers (0.5, 1.5) for each dimension, difficulty parameters were drawn from a standard normal distribution, and guessing parameters were drawn from a uniform distribution at the interval of real numbers (0, 0.4). For the between-item two-dimensional pool, one thousand M3PL items were generated, in which 60% and 40% of the items measured the first and second dimensions, respectively. For the within-item two-dimensional pool, one thousand M3PL items were generated, in which 40% of the items measured the first dimension, 30% of the items measured the second dimension, and the remaining 30% of the items measured both dimensions. The numbers of content areas simulated for these two dimensions were 3 and 2, and items were randomly assigned to these areas with equal probability. All 5,000 simulated examinees were drawn from a multivariate standard normal distribution with correlations .8 and .4, indicating high and low correlation. Item responses were generated according to Equation 7.

Simulation Conditions

Four item selection methods were considered in this study: two constrained item selection methods and two control methods. In this study, two constrained methods were the MPI and the MMPI; two control methods were the MDFIM and the randomized (R) item selection. When the R method was applied, items were selected randomly for administration. When the MDFIM method was applied, an item with the maximum determinant of the Fisher information matrix was selected for administration. The R method selected items randomly, so the performance of measurement precision was the worst scenario among all the item selection methods. The MDFIM method selected items with the maximum information criterion, so the performance of measurement precision was the best scenario among all the item selection methods.

The constraints and weights for the constrained item selection methods in between-item and within-item multidimensional pools are listed in Table 1. Eleven constraints, including content balancing, key balancing, item exposure control, and item information, were considered in the study. For the item exposure control constraint, 0.2 was the target maximum item exposure rate in the study. For the item information constraint, the determinant of the Fisher information matrix of unadministered items in the pool was calculated in the study during each item selection step. As the item information criterion is already included by the MMPI in Equation 3, the determinant of the Fisher information matrix is multiplied by the MPI in Equation 1 before item selection. When the MPI or the MMPI method was applied, an item with the maximum value among unadministered items in the pool was selected for administration. Similar to Yao's study (2013), 3 and 1 were used for b and ε_2 in Equation 2, respectively. The total test length was 30. The Maximum-a-Posteriori (MAP) estimation with a prior, matching examinees' ability distribution, was used to estimate $\hat{\boldsymbol{\theta}}$.

Evaluation Criteria

The results of the simulations were analyzed and discussed based on the following criteria: (a) measurement precision, (b) exposure control, and (c) constraint management. With respect to measurement precision, the bias, the root-mean-square error (RMSE), and a measure of relative efficiency were recorded for each item selection method. The formulas for bias and RMSE are given as follows:

$$\text{Bias} = \frac{1}{N}\sum_{n=1}^{N}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right) \tag{8}$$

and

Table 1. Constraints and Weights for Between-Item and Within-Item Multidimensional Pools.

	Between-item pool			Within-item pool		
Constraints	Weight	Lower bound	Upper bound	Weight	Lower bound	Upper bound
Dimension 1—Content 1	1	3	5	1	5	9
Dimension 1—Content 2	1	5	7	1	7	13
Dimension 1—Content 3	1	4	6	1	6	11
Dimension 2—Content 1	1	5	8	1	6	14
Dimension 2—Content 2	1	6	9	1	7	16
Answer Key-A	1	5	10	1	5	10
Answer Key-B	1	5	10	1	5	10
Answer Key-C	1	5	10	1	5	10
Answer Key-D	1	5	10	1	5	10
Item exposure control	1		0.2	1		0.2
Fisher information	1			1		

Note. The Fisher information constraint refers to the determinant of the Fisher information matrix.

$$\text{RMSE} = \left[\frac{1}{N}\sum_{n=1}^{N}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n\right)^2\right]^{1/2}, \tag{9}$$

where $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n$ are the estimated and true abilities, respectively. To evaluate the latent trait recovery, the relative efficiency is defined as the ratio of RMSE of each item selection method to that of the MDFIM method.

With respect to exposure control, the maximum item exposure rate, the number of overexposed items, the number of unused items, and the skewness of the item exposure rate distribution were recorded. The item exposure rates were recorded for all items, and only the maximum item exposure rate was reported. The overexposed items were items with item exposure rates higher than 0.2. The unused items were items that were never exposed. The chi-square statistic was used to measure the skewness of the item exposure rate distribution (Chang & Ying, 1999).

$$\chi^2 = \frac{1}{L/I}\sum_{i=1}^{I}(r_i - L/I)^2, \tag{10}$$

where r_i is the exposure rate of item i and L is the test length; the smaller the chi-square statistic, the better the item exposure control.

Constraint management is checking whether the tests met the specified constraints for examinees. The number of constraints that were violated in each test (for each examinee) was recorded, and then the proportion of tests violating a certain number of constraints was calculated. Finally, the averaged number of violated constraints ($\bar{V}$) for each item selection method was calculated by

$$\overline{V} = \frac{\sum_{n=1}^{N} V_n}{N}, \tag{11}$$

where V_n represents the number of constraint violations in the nth examinees' test.

Results

The results of the simulations are summarized according to measurement precision, exposure control, and constraint management in Tables 2, 3, and 4, respectively.

Measurement Precision

With respect to measurement precision, the bias, RMSE, and relative efficiency of the four item selection methods for the between-item and within-item multidimensional pools are listed in Table 2. The MDFIM and R methods were the baselines in this study. Because the MDFIM method selected items with the maximum information criterion, the MDFIM method's performance of measurement precision is the best scenario among all the item selection methods. Indeed, among the four item selection methods, the MDFIM method yielded the best measurement precision in terms of the smallest bias and RMSEs for two multidimensional item pools. When the correlation between latent traits was .8, the RMSEs of two dimensions for the MDFIM method were 0.24 and 0.27 for the between-item multidimensional pool, and were 0.23 and 0.24 for the within-item multidimensional pool.

Because the R method selected items randomly, the R method's performance of measurement precision was the worst scenario among all the item selection methods. Indeed, among the four item selection methods, the R method yielded the worst measurement precision in terms of the largest bias, the largest RMSE, and the smallest relative efficiency. When the correlation between latent traits was .8, the RMSEs of two dimensions for the R method were 0.43 and 0.48 for the between-item multidimensional pool, and were 0.43 and 0.44 for the within-item multidimensional pool.

The MPI and the MMPI methods performed very similarly in terms of the bias, RMSE, and relative efficiency. When the correlation between latent traits was .8, the RMSEs of two dimensions for the MPI method were 0.32 and 0.36 for the between-item multidimensional pool, and were 0.28 and 0.35 for the within-item multidimensional pool; the RMSEs of two dimensions for the MMPI method were 0.34 and 0.34 for the between-item multidimensional pool, and were 0.31 and 0.29 for the within-item multidimensional pool.

Compared with the MDFIM method, the relative efficiency of the MPI method ranged between 0.67 and 0.76 for the between-item multidimensional pools, and ranged between 0.66 and 0.83 for the within-item multidimensional pools. The relative efficiency of the MMPI method ranged between 0.70 and 0.78 for the between-item multidimensional pools, and ranged between 0.67 and 0.85 for the within-item multidimensional pools. In general, the measurement precision of the item selection methods was slightly better in the within-item multidimensional pool than in the between-item multidimensional pool; was slightly better for correlation .8 conditions than for correlation .4 conditions. With respect to measurement precision, the R method yielded the worst performance, whereas the MDFIM method yielded the best performance; the MPI and the MMPI methods yielded similar performance.

Exposure Control

With respect to exposure control, the actual item exposure rates of each item were recorded for the four item selection methods under the between-item and within-item multidimensional pools. The maximum item exposure rate, the number of overexposed items, the number of unused items, and the chi-square statistic measuring the skewness of the item exposure rate distribution were calculated for each item selection method. The results of exposure control for the four item selection methods are listed in Table 3. For both multidimensional item pools, the

Table 2. Measurement Precision Results for the Item Selection Methods.

Item selection methods	bias		RMSE		Relative efficiency	
	Dimension 1	Dimension 2	Dimension 1	Dimension 2	Dimension 1	Dimension 2
Between-item pool						
Correlation = .8						
R	0.026	0.033	0.430	0.479	0.559	0.557
MDFIM	0.007	0.015	0.240	0.267	1.000	1.000
MPI	0.015	0.019	0.317	0.363	0.759	0.734
MMPI	0.016	0.026	0.335	0.344	0.717	0.776
Correlation = .4						
R	0.029	0.036	0.444	0.546	0.580	0.503
MDFIM	0.008	0.021	0.258	0.275	1.000	1.000
MPI	0.022	0.035	0.337	0.410	0.764	0.670
MMPI	0.025	0.027	0.367	0.372	0.702	0.738
Within-item pool						
Correlation = .8						
R	0.019	0.020	0.432	0.442	0.540	0.553
MDFIM	0.002	0.009	0.233	0.244	1.000	1.000
MPI	0.010	0.024	0.282	0.353	0.826	0.691
MMPI	0.011	0.011	0.308	0.288	0.757	0.847
Correlation = .4						
R	0.009	0.033	0.451	0.481	0.570	0.551
MDFIM	0.002	0.014	0.257	0.265	1.000	1.000
MPI	0.017	0.038	0.308	0.401	0.834	0.661
MMPI	−0.001	0.019	0.377	0.344	0.681	0.770

Note. Four item selection methods in this study were (a) the randomized (R) item selection, (b) the MDFIM, (c) the MPI, and (d) the MMPI. RMSE = root mean square error; MDFIM = maximum determinant of the Fisher information matrix; MPI = multidimensional priority index; MMPI = modified multidimensional priority index.

MDFIM method yielded the worst exposure control, with the largest values for the maximum exposure rate, the number of overexposed items, the number of unused items, and the chi-square statistic. When the correlation between latent traits was .8, the maximum exposure rates for the MDFIM method were 0.70 and 0.53 (both exceed the pre-specified value of 0.2), the numbers of the overexposed items were 59 and 53, the numbers of unused items were 779 and 743, and the chi-square statistics were 279.90 and 206.46 for between-item and within-item multidimensional pools, respectively.

By contrast, for both multidimensional item pools, the R method yielded the best exposure control, with the smallest values for the maximum exposure rate, the number of overexposed items, the number of unused items, and the chi-square statistic. When the correlation between latent traits was .8, the maximum exposure rates for the R method were 0.04 and 0.04 (both less than the pre-specified value of 0.2), the numbers of overexposed items were 0 and 0, the numbers of unused items were 0 and 0, and the chi-square statistics were .20 and .20 for between-item and within-item multidimensional pools, respectively.

For the between-item multidimensional pool, the MPI and the MMPI methods performed very similarly. When the correlation between latent traits was .8, the maximum exposure rates for the MPI and the MMPI methods were 0.08 and 0.08 (both less than the pre-specified value of 0.2), the numbers of overexposed items were 0 and 0, the numbers of unused items were 0 and 0, and the chi-square statistics were 4.15 and 3.77, respectively. However, for the within-

Table 3. Exposure Control Results for the Item Selection Methods.

Item selection methods	Maximum exposure rate	Overexposed items	Unused items	χ^2
Between-item pool				
Correlation = .8				
R	0.040	0	0	0.198
MDFIM	0.702	59	779	279.896
MPI	0.078	0	0	4.146
MMPI	0.076	0	0	3.774
Correlation = .4				
R	0.039	0	0	0.195
MDFIM	0.774	58	790	297.330
MPI	0.084	0	0	5.313
MMPI	0.079	0	0	4.402
Within-item pool				
Correlation = .8				
R	0.038	0	0	0.195
MDFIM	0.527	53	743	206.456
MPI	0.081	0	0	5.172
MMPI	0.104	0	138	23.616
Correlation = .4				
R	0.038	0	0	0.196
MDFIM	0.573	58	751	239.314
MPI	0.078	0	0	6.224
MMPI	0.102	0	130	21.181

Note. Four item selection methods in this study were (a) the randomized (R) item selection, (b) the MDFIM, (c) the MPI, and (d) the MMPI. MDFIM = maximum determinant of the Fisher information matrix; MPI = multidimensional priority index; MMPI = modified multidimensional priority index.

item multidimensional pool, the MPI method yielded better performance than the MMPI method because the MPI method obtained a smaller maximum exposure rate, fewer unused items, and a smaller chi-square statistic. When the correlation between latent traits was .8, the maximum exposure rates for the MPI and the MMPI methods were 0.08 and 0.10, the numbers of overexposed items were 0 and 0, the numbers of unused items were 0 and 138, and the chi-square statistics were 5.17 and 23.62, respectively.

In general, the R method yielded the best performance in exposure control, followed by the MMPI and the MPI methods, and the MDFIM method yielded the worst performance for both multidimensional item pools. The MPI and the MMPI methods yielded better performance in exposure control for the between-item multidimensional pool than for the within-item multidimensional pool. With respect to exposure control, although the MMPI and the MPI methods yielded similar performance for the between-item multidimensional pools, the MPI method yielded better performance than the MMPI method for within-item multidimensional pools. For both multidimensional pools, four item selection methods yielded slightly better performance on exposure control in the correlation .8 conditions than in the correlation .4 conditions.

Constraint Management

As the violation was considered at each examinee level, only the first nine constraints in Table 1 were included to evaluate the efficiency of the constraint management. The proportions of assembled tests violating a certain number of constraints and the average number of violated constraints for the four item selection methods are listed in Table 4. When the tests were assembling, the R method yielded the severest violation, followed by the MDFIM method, and the MMPI and the MPI methods yielded the best constraint management for the between-item multidimensional pool. For the between-item multidimensional pool, the averaged violations of the R, the MDFIM, the MPI, and the MMPI methods were 3.31, 2.68, 0.00, and 0.00 when the correlation between latent traits was .8, respectively. Similar pattern was found for the within-item multidimensional pool. For the within-item multidimensional pool, the averaged violations of the R, the MDFIM, the MPI, and the MMPI methods were 2.01, 1.96, 0.06, and 0.00 when the correlation between latent traits was .8, respectively.

Table 4. Constraint Management Results for the Item Selection Methods.

	Number of violations										
Item selection methods	0	1	2	3	4	5	6	7	8	9	Averaged
Between-item pool											
Correlation = .8											
R	0.07	0.37	0.40	0.14	0.01	0.00	0.00	0.00	0.00	0.00	3.31
MDFIM	0.17	0.40	0.35	0.08	0.00	0.00	0.00	0.00	0.00	0.00	2.68
MPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MMPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Correlation = .4											
R	0.08	0.38	0.39	0.14	0.01	0.00	0.00	0.00	0.00	0.00	3.27
MDFIM	0.16	0.43	0.34	0.07	0.00	0.00	0.00	0.00	0.00	0.00	2.64
MPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MMPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Within-item pool											
Correlation = .8											
R	0.32	0.42	0.20	0.05	0.01	0.00	0.00	0.00	0.00	0.00	2.01
MDFIM	0.29	0.47	0.23	0.02	0.00	0.00	0.00	0.00	0.00	0.00	1.96
MPI	0.97	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
MMPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Correlation = .4											
R	0.31	0.42	0.20	0.06	0.01	0.00	0.00	0.00	0.00	0.00	2.06
MDFIM	0.31	0.45	0.21	0.03	0.00	0.00	0.00	0.00	0.00	0.00	1.93
MPI	0.94	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
MMPI	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note. Four item selection methods in this study were (a) the randomized (R) item selection, (b) the MDFIM, (c) the MPI, and (d) the MMPI. MDFIM = maximum determinant of the Fisher information matrix; MPI = multidimensional priority index; MMPI = modified multidimensional priority index.

In general, the R and the MDFIM methods yielded better performance in constraint management for the within-item multidimensional pool than for the between-item multidimensional pool; but the MPI and the MMPI methods yielded better performance in constraint management for the between-item multidimensional pool than for the within-item multidimensional pool. With respect to constraint management, the MPI and the MMPI methods yielded similar performance for the between-item multidimensional pool, but the MMPI methods yielded better performance than the MPI method for the within-item multidimensional pool. For both multidimensional pools, four item selection methods yielded slightly better performance on constraint management in the correlation .4 conditions than in the correlation .8 conditions.

In summary, the MDFIM method had the best results in measurement precision, but it lost some control in exposure control and constraint management. The R method had the best results in exposure control but lost some control in measurement precision and constraint management. The MMPI and the MPI methods obtained similar measurement precision for both multidimensional pools; and obtained similar exposure control and constraint management for the between-item multidimensional pool. For the within-item multidimensional pool, the MPI method yielded better exposure control but yielded worse constraint management than the MMPI method. When the correlation between latent traits was .8, the four item selection methods yielded slightly better performance in measurement precision and exposure control, and did slightly worse performance in constraint management.

Discussion

Today, CAT is making a crucial influence on how people are selected, classified, and diagnosed; CAT studies will lead to better assessments, and hence benefit society (Chang, 2015). One of the main challenges in educational and psychological measurement is to develop item selection methods for CAT. Assembling tests in CAT usually requires meeting many statistical and non-statistical constraints simultaneously. The MPI item selection method (Yao, 2011, 2012, 2013, 2014) can be used to handle many constraints for between-item multidimensional tests, whereas the MMPI item selection method (Su, 2015; Su & Huang, 2015) can be used to handle many constraints for within-item multidimensional tests. This study compared the performances of the MMPI and the MPI methods for item selection under two different bank structures through simulations. The results from the study show that both item selection methods have great potential in operational CAT. The MMPI and the MPI item selection methods obtained similar measurement precision for both multidimensional pools. These two item selection methods also obtained similar exposure control and constraint management for between-item multidimensional pool. However, the MPI method yielded better exposure control but yielded worse constraint management than the MMPI method for the within-item multidimensional pool.

The research findings from this study will advance our knowledge of item selection in multidimensional CAT. However, there are also some limitations to the current study. First, the test length was fixed at 30 in the study. When a stopping rule of fixed length is considered, the precisions vary at different examinee levels, resulting in a high misclassification rate, which might be costly. To achieve the same measurement precision, it is important to apply the MMPI method to variable-length CAT. Second, even though both MPI and MMPI methods successfully kept all the items from being overexposed, the MMPI method still left some items unused. As developing items is very expensive, it is important to have the item pool fully utilized. The *a*-stratified design (Chang & Ying, 1999; Chang et al., 2001) can be integrated with the MMPI item selection method; however, because there is more than one discrimination parameter for the within-item multidimensional items, the method to stratify the pool needs to be investigated. Third, item exposure and test overlap rates are two popular indices to track item exposure in CAT. The item exposure rate is the administered proportion of an item. The test overlap rate is the proportion of items shared by pairs of exams, averaged across all possible pairwise comparisons. By considering both indices, item exposure can be monitored at both the item and test levels (Chen & Lei, 2005). Hence, it is worthwhile to integrate the test overlap rate with the MMPI item selection method and investigate its performance in a further study. Fourth, the determinant of the Fisher information matrix was considered one of the constraints in this study. The item information criterion might play an important role during item selection. It deserves further study when other selection criteria or information measures are considered, such as minimum angle (Reckase, 2009), minimum error variance of the linear combination (van der Linden, 1999), minimum error variance of the composite score with the optimized weight (Yao, 2010), and Kullback–Leibler information (Chang & Ying, 1996).

Thanks to an anonymous reviewer, who pointed out that the correlation between latent traits and the dimensional structure are important factors in CATs. With respect to the correlation factor, one more level of correlation, .4, was considered as low correlation between latent traits in the study. The correlation factor showed small effect on the performance of item selection methods. With respect to the dimensional structure factor, due to space limitation of the current paper, 80% and 20% of the items measured the first and the second dimensions were considered for the between-item multidimensional pool; and 70%, 20%, 10% of the items measured the first, the second, and both dimensions were considered for the within-item multidimensional pool. The correlation between latent traits was .8. Results from these two multidimensional pools were similar to those in the simulations. With respect to measurement precision, the RMSEs of two dimensions for the MPI item selection method were 0.34 and 0.34 for the between-item multidimensional pool, and were 0.30 and 0.31 for the within-item multidimensional pool. The RMSEs of two dimensions for the MMPI item selection method were 0.36 and 0.32 for the between-item multidimensional pool, and were 0.29 and 0.30 for the within-item multidimensional pool. With respect to exposure control, for the between-item multidimensional pools, the maximum exposure rates for the MPI and the MMPI item selection methods were 0.08 and 0.09, the numbers of overexposed items were 0 and 0, the numbers of unused items were 0 and 0, and the chi-square statistics were 4.17 and 3.85, respectively. For the within-item multidimensional pools, the maximum exposure rates for the MPI and the MMPI item selection methods were 0.11 and 0.12, the numbers of overexposed items were 0 and 0, the numbers of unused items were 0 and 2, and the chi-square statistics were 6.24 and 18.73, respectively. With respect to constraint management, for the between-item multidimensional pools, the averaged violations of the MPI and the MMPI item selection methods were 0.00 and 0.00, respectively. For the within-item multidimensional pools, the averaged violations of the MPI and the MMPI item selection methods were 0.06 and 0.05, respectively. That is, the MMPI and the MPI item selection methods had similar measurement precision for both multidimensional pools, and had similar exposure control and constraint management for the between-item multidimensional pool. However, the MPI item selection method had slightly better exposure control but had slightly worse constraint management than the MMPI method for the within-item multidimensional pool. The performance of item selection methods might be affected by bank structure, constraints, administered population, and so on. Future research can be carried out along this line in a large-scale CAT program. Other interesting research questions include how to apply the MMPI item selection method to polytomously scored models, and how to incorporate constraints for testlets (Wainer, Bradlow, & Wang, 2007).

Acknowledgment

The author is indebted to the editor and two anonymous reviewers for their suggestions and comments on the earlier versions of the manuscripts.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is partially supported by the National Science Council in Taiwan

(Research Grants NSC 99-2410-H-194-004, NSC 101-2410-H-194-060, and NSC 102-2918-I-194-005) and a research grant from the Chinese Testing International Co. Ltd.

References

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*, 237-246.

Chang, H.-H. (2007). Book review: Linear models for optimal test design. *Psychometrika*, *72*, 279-281.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1-20. doi: 10.1007/s11336-014-9401-5

Chang, H.-H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage CAT with b-blocking. *Applied Psychological Measurement*, *25*, 333-341.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

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

Chen, S.-Y., & Lei, P.-W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, *29*, 204-217.

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.

Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, *69*, 35-49.

Hathaway, S. R., & McKinley, J. C. (1989). *Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. *Educational and Psychological Measurement*, *74*, 677-696.

Leung, C., Chang, H., & Hau, K. (2005). Computerized adaptive testing: A mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology*, *58*, 239-257.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Robin, F., van der Linden, W. J., Eignor, D. R., Steffen, M., & Stocking, M. L. (2005). *A comparison of two procedures for constrained adaptive test construction* (ETS Research Report No. RR-04–39). Princeton, NJ: Educational Testing Service.

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

Shin, C., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs*. Pearson. Retrieved from http://images.pearsonassessments.com/images/tmrs/tmrs_rg/WeightedPenaltyModel.pdf?WT.mc_id=TMRS_Weighted_Penalty_Model_for_Content_Balancing

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

Su, Y.-H. (2015). The performance of the modified multidimensional priority index for item selection in variable-length MCAT. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research* (pp. 89-97). Switzerland: Springer.

Su, Y.-H., & Huang, Y.-L. (2015). Using a modified multidimensional priority index for item selection under within-item multidimensional computerized adaptive testing. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (pp. 227-242). Switzerland: Springer.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151-166.

Timminga, E. (1998). Solving infeasibility problems in computerized test assembly. *Applied Psychological Measurement*, *22*, 280-291.

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. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston, MA: Kluwer-Nijhoff.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*, 283-302.

van der Linden, W. J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, *27*, 107-120.

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

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*, 273-291.

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.

Veldkamp, B. P., & van der Linden, W. J. (2008). Implementing Sympson–Hetter item-exposure control in a shadow-test approach to constrained adaptive testing. *International Journal of Testing*, *8*, 272-289.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological 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. (2011, October). *Multidimensional CAT item selection procedures with item exposure control and content constraints*. Paper presented at the 2011 International Association of Computer Adaptive Testing (IACAT) Conference, Pacific Grove, CA.

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

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement*, *37*, 3-23.

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