Article

Item Selection With Collaborative Filtering in On-The-Fly Multistage Adaptive Testing

Applied Psychological Measurement
2022, Vol. 46(8) 690–704

DOI: 10.1177/01466216221124089
journals.sagepub.com/home/apm


Jiaying Xiao[1] **and Okan Bulut**[2]

Abstract

An important design feature in the implementation of both computerized adaptive testing and multistage adaptive testing is the use of an appropriate method for item selection. The item selection method is expected to select the most optimal items depending on the examinees' ability level while considering other design features (e.g., item exposure and item bank utilization). This study introduced collaborative filtering (CF) as a new method for item selection in the *on-the-fly assembled multistage adaptive testing* framework. The user-based CF (UBCF) and item-based CF (IBCF) methods were compared to the maximum Fisher information method based on the accuracy of ability estimation, item exposure rates, and item bank utilization under different test conditions (e.g., item bank size, test length, and the sparseness of training data). The simulation results indicated that the UBCF method outperformed the traditional item selection methods regarding measurement accuracy. Also, the IBCF method showed the most superior performance in terms of item bank utilization. Limitations of the current study and the directions for future research are discussed.

Keywords

collaborative filtering, multistage adaptive testing, item selection, measurement accuracy

With the rapid advancement of information technologies and robust computer systems, more and more large-scale testing programs (e.g., Graduate Management Admission Test) have transitioned from traditional paper-and-pencil testing to computerized adaptive testing (CAT) over the past 20 years. However, in both research and practice, CAT indicated several limitations, such as not allowing examinees to review completed items or skip items (Wainer, 1993), the lack of control over the context effects (Hendrickson, 2007), and overestimation or underestimation of

[1]College of Education, University of Washington, Seattle, WA, USA
[2]Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Jiaying Xiao, Measurement and Statistics, College of Education, University of Washington, 2012 Skagit Ln, Seattle, WA 98105, USA.
Email: jxiao6@uw.edu

examinees' abilities in short tests (Chang & Ying, 2008). Therefore, new adaptive testing frameworks have been proposed to address these issues by allowing response review and revision in CAT (Wang et al., 2017) or combining the design features of both CAT and multistage adaptive testing (MST), such as the hybrid designs (Bao et al., 2021; Wang et al., 2016;) and *on-the-fly assembled multistage adaptive testing* (OMST; Zheng & Chang, 2011, 2015). The current study was based on the OMST framework.

On-the-fly assembled multistage adaptive testing is a group-sequential design in which items are grouped into several modules. Modules in the first stage are preassembled at a moderate difficulty level, while modules for the subsequent stages are assembled on the fly (i.e., in real-time). Therefore, each examinee receives a different set of items in the second and third stages based on their provisional ability estimates (Zheng & Chang, 2015). Unlike typical MST in which preassembled modules are administered at each stage, OMST builds new modules after the first module and creates a uniquely tailored test for each examinee. Although Zheng and Chang (2015) indicated that OMST provides a flexible framework of sequential testing and controls several psychometric properties adequately, further studies in this direction have been scant so far (Wang et al., 2016). Limited literature has shown that OMST has better test security and flexibility (Tay, 2015); therefore, the current study was based on the OMST design.

A successful adaptive testing application requires implementing an appropriate item selection method, and OMST is no exception to this condition. In OMST without non-statistical constraints (e.g., content coverage), the maximum Fisher information (MFI) method (Thissen & Mislevy, 1990) can be used for assembling a new module by selecting items that maximize the Fisher information at the latest provisional ability estimate. To maximize the Fisher information within a module, the MFI method tends to choose the items that are highly discriminating and have difficulty levels closer to the provisional ability estimate. However, this behavior of the MFI method could lead to some undesirable effects in practice. For example, some items from the item bank may be selected very frequently while the remaining items are never or hardly ever used, resulting in overexposure and underexposure of the items (Eggen, 2001). Highly uneven item selection also affects the utilization of the item bank negatively. Another potential problem with the MFI method and its variants is that solely relying on maximizing the Fisher information leads to selecting items where the examinee's probability of answering the items correctly is roughly 50%.[1] Previous studies showed that depending on their motivation levels, some examinees may perceive such adaptive tests as much harder than conventional tests and thus perform with lower effort, compared to those who are more motivated to take the test (e.g., Kim & McLean, 1995; Tonidandel et al., 2002). As Wise (2014) pointed out, this situation could pose a significant threat to the validity of inferences and interpretations to be made from such adaptive tests.

Researchers are dedicated to developing new item selection methods for adaptive tests to address the challenges mentioned above. Recently, there has been an upward trend in the use of data mining and machine learning algorithms in education (Nehm et al., 2012). One of these promising algorithms is *collaborative filtering* (CF), which is a method widely used by commercial applications such as Netflix for producing user-specific recommendations of items (e.g., movies) based on a user's ratings or usage (e.g., liked or disliked movies) or similar users' ratings (Sarwar et al., 2001). This algorithm can also be divided into two main categories: user-based (UBCF) and item-based (IBCF) approaches (Breese et al., 1998). The former recommends items liked by similar users, and the latter recommends items similar to those that a user liked or preferred in the past (Lu, Wu, Mao, Wang, & Zhang, 2015). The primary advantages of the CF algorithm include its computational efficiency in searching for the most suitable item for each user among many available options and its accuracy in recommending a suitable item in the presence of data sparsity (Hu et al., 2017).

To date, a large number of studies have shown superior performance of the CF algorithm for predicting ratings or recommending products in intelligent recommender systems, but their applications to educational assessments are rarely discussed. Toscher and Jahrer (2010) used the CF algorithm for predicting students' abilities to respond to items correctly, which achieved the same goal as a traditional item response theory (IRT) model that estimates the probability of an examinee answering the item correctly. Thai-Nghe et al. (2012) conducted a similar study in which they used the CF algorithm to encode the prevailing latent factors implicitly (i.e., "slip" and "guess") for predicting student performance. Furthermore, Bergner et al. (2012) formalized the relationship between IRT and CF by using the CF algorithm to estimate "difficulty-like" and "discrimination-like" parameters. Other studies applied CF methods to summative and formative assessments to provide students with personalized feedback (de Schipper et al., 2021) and generate personalized test administration schedules (Bulut et al., 2020; Shin & Bulut, 2021). These studies demonstrated the utility of the CF algorithm as a psychometric method and highlighted its main strength of finding the most suitable items efficiently. The same computational strength also makes the CF algorithm a plausible approach for selecting items in adaptive testing.

This study aims to utilize the CF algorithms as item selection methods under the OMST framework and compare their performance with the MFI method under different test conditions. The rest of this study is organized as follows. First, item selection based on the MFI method is briefly explained. Next, the item selection procedures based on the CF algorithms are introduced. Then, simulation studies are presented to compare the performances of the item selection methods in terms of accuracy of ability estimates and item bank utilization in OMST. Finally, conclusions and future directions are discussed.

Maximum Fisher Information Method

The MFI method can be used for item selection when an adaptive test does not involve any non-statistical constraints, such as content-balancing requirements. This method was proposed by Birnbaum (1968) to explain the information function for dichotomous items. It describes the extent to which an item contributes to the quality of ability estimation. For example, in the three-parameter logistic (3PL) model, item information at a given ability level θ can be calculated as follows

$$I(\theta) = \frac{(1-c)a^2\exp[a(\theta-b)]}{\{1+\exp[a(\theta-b)]\}^2\{1-c+c\{1+\exp[a(\theta-b)]\}\}}, \quad (1)$$

where a is the discrimination parameter, b is the difficulty parameter, c is the lower asymptote, and $I(\theta)$ is the item information level at the ability level of θ. Using the MFI method in the OMST framework, a set of items that maximize the information at the latest provisional ability level can be selected after each stage to build a custom module for each examinee (Zheng & Chang, 2015). Equation (1) shows that an item can provide the highest amount of information when θ is matched (or closely matched) to the b value, the a value is relatively high, and the c value is closer to zero. Thus, item selection based on MFI is more likely to choose items with large a and small c values, which results in high usages (i.e., exposure) of some items from the item bank and leads to lower test security since commonly used items can be memorized by some examinees in real test administrations (Chang & Ying, 1999). To address this issue, several item exposure control strategies have been proposed under two categories (Stocking, 1993): methods including an exposure-rate parameter to each item to control the maximum exposure (e.g., Sympson & Hetter, 1985) and methods adding a random component to MFI (e.g., McBride & Martin, 1983;

Kingsbury & Zara, 1989). The current study focused on Kingsbury and Zara's (1989) randomesque method with MFI (denoted as MFI-R). For example, if the desired number of items is 20, we select the 40 most informative items (instead of the 20 most informative items) and randomly select 20 items from them to compose adaptive modules.

Collaborative Filtering

The CF methods typically utilize raw data (e.g., users' ratings of movies) as a rating matrix to find similarities between users or items in the prediction stage. In this study, we used the item information values as a rating matrix instead of raw data (i.e., dichotomous item responses). To apply the CF methods to the item selection procedure in OMST, the critical part of our setting is to build an $N \times J$ person-item rating matrix $\mathbf{R}$ as a training dataset where each row $(n = 1, 2, 3, \ldots, N)$ represents an examinee in the examinee pool (i.e., examinees with known ability levels), each column $(j = 1, 2, 3, \ldots, J)$ represents an item from the item bank (i.e., items with known parameters), and cell values are the item information values computed based on the examinees' ability levels and the item parameters (see Equation (1)). The rating matrix $\mathbf{R}$ has no missing values because the item information can be computed for each combination or examinee j and item n based on their known parameters.

Using item information values instead of raw data (i.e., dichotomous item responses) in the training dataset has two major advantages. First, compared with dichotomous item responses, continuous values of item information are more suitable for constructing a rating matrix required for the CF methods. The CF methods can find similar items based on their information levels using the item information matrix and thereby recommend the most informative items. Second, the item response data from a typical OMST administration would be highly sparse since each examinee answers a unique set of items after the first module. When the rating matrix is highly sparse, the CF methods fail to produce accurate recommendations due to insufficient information (Huang et al., 2004). Using the item information values as a rating matrix solves the sparsity problem because item information can be calculated for both answered and unanswered items based on previous examinees' ability estimates in the training dataset and new users' provisional ability estimates during the OMST administration.

To describe the item selection procedure with the CF methods within an OMST administration, assume that r_{n_j} refers to item j's information for examinee u_n. Then, the item set for examinee u_n can be denoted as I_n in which the items examinee u_n has not answered yet are recorded as missing values. Finding similarities between similar examinees or items from the training dataset and calculating missing responses in I_n is called *prediction*. Lastly, a *recommendation* process is implemented to create a top-N list that includes N items with the highest predicted information for examinee u_n. Then, these items can be used to assemble a new module to be administered to the examinee in the next stage. Note that since the item information is calculated based on the provisional ability estimate for each examinee, I_n must be updated after each stage of OMST.

User-Based Collaborative Filtering. User-based collaborative filtering (UBCF) utilizes the training dataset (i.e., an $N \times J$ person-item rating matrix $\mathbf{R}$) to search for similar users called *neighbors* who rated the items similarly to the target examinee u_n (Breese et al., 1998). In our study, *neighbors* were the examinees who had similar item information values as the target examinee u_n. The similarity between the target examinee u_n and other examinees in the training dataset can be measured using either cosine similarity or Pearson correlation. The cosine similarity index can be computed as

$$sim(x,y) = \cos(x,y) = \frac{\sum_{i \in I_{xy}} x_i y_i}{\sqrt{\sum_{i \in I_{xy}} x_i^2} \sqrt{\sum_{i \in I_{xy}} y_i^2}}, \tag{2}$$

and the Pearson correlation coefficient can be computed as

$$sim(x,y) = cor(x,y) = \frac{\sum_{i \in I_{xy}} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i \in I_{xy}} \left(x_i - \overline{x}\right)^2 \sum_{i \in I_{xy}} \left(y_i - \overline{y}\right)^2}}, \tag{3}$$

where x, y denote two examinees' item information values, I_{xy} denotes the set of items answered by both examinees, and $\overline{x}$ and $\overline{y}$ denote the average item information values for the examinees. Then, the neighbors of the examinee u_n can be selected by either taking the k-nearest neighbors or setting a particular similarity threshold based on the cosine similarity or the Pearson correlation coefficient. Once the neighbors are identified, the UBCF algorithm combines item information values to form a prediction or top-N list. The easiest way to aggregate the results is to average neighbors' item information values.

Item-Based Collaborative Filtering. Unlike UBCF that utilizes a user-item rating matrix in the prediction process, IBCF focuses on the similarity between items and calculates a $J \times J$ item-to-item similarity matrix $\mathbf{S}$ (Sarwar et al., 2001). The underlying assumption of IBCF is that users would prefer items similar to those they rated highly in the past. In our setting, examinees would be recommended items that provide similar or higher information than the administered items. The similarity calculation between two items i and j is based on the information from examinees who have answered both items. Either cosine similarity or Pearson correlation can also be used to calculate item similarities, but the cosine similarity should be adjusted to offset examinee differences by subtracting each examinee's average item information values separately (Sarwar et al., 2001). The revised formula can be expressed as follows

$$sim(i,j) = \frac{\sum_{u \in \mathrm{U}} \left(r_{u,i} - \overline{r}_u\right)\left(r_{u,j} - \overline{r}_u\right)}{\sqrt{\sum_{u \in \mathrm{U}} \left(r_{u,i} - \overline{r}_u\right)^2} \sqrt{\sum_{u \in \mathrm{U}} \left(r_{u,j} - \overline{r}_u\right)^2}}, \tag{4}$$

where $r_{u,i}$, $r_{u,j}$ denote the item information of examinee u on items i and j, and $\overline{r}_u$ denotes the examinee u's average item information value. To improve the space complexity and reduce the computation time, only the k most similar items to each item are stored, where k is smaller than the number of items. In this computational process, k refers to the model size. The next step is to check how many of those k items have been administered to the target examinee u_n. Finally, examinee u_n 's information values for new items can be predicted by computing a weighted sum of the information values on similar items. Equation (5) shows how to calculate the prediction information of the item j for examinee u_n

$$P_{nj} = \frac{\sum_i (s(j,i) v_{ni})}{\sum_i |s(j,i)|}, \tag{5}$$

where i denotes the item that is supposed to be similar to item j, $s(j,i)$ denotes the similarity between items j and i, and v_{ni} denotes item j's information for examinee u_n. Based on the prediction results, t items with the highest information will be recommended to examinee u_n.

Cold-Start Problem. The CF methods are known to produce less accurate predictions or recommendations when there is no prior information available about new users, which is known as *the cold-start problem* (e.g., Biswas et al., 2017; Zhao, 2016). When applying the UBCF and IBCF methods to item selection in OMST, the cold-start problem does not occur because the first stage of OMST is based on a preassembled module, and thus there is no adaptive item selection. Information obtained from the first stage can be incorporated into the CF methods for selecting the items adaptively for the second and subsequent stages. Using the information obtained from earlier stages, the UBCF method recommends items based on the examinees with similar item information values as the target examinee, while the IBCF method recommends items that yield similar or higher information than the target examinee's administered items.

Methods

The current study follows a Monte Carlo simulation approach since it aims to compare the performances of four item selection methods based on the accuracy of ability estimates under the OMST design. The simulation conditions included the size of the item bank (300 or 600 items), test length (30 or 60 items), and item selection methods (MFI, MFI-R, UBCF, or IBCF). The simulation study was implemented using the xxIRT (Luo, 2016), mirt (Chalmers, 2012), and recommenderlab (Hahsler, 2015) packages in R (R Core Team, 2021).

Data Generation

The item bank was constructed based on dichotomous items from an operational CAT program used for measuring K-12 students' reading abilities in the United States. Two sets of item parameters calibrated with the 3PL model (i.e., 300 and 600 items) were randomly selected from the original item bank. The selected item parameters were used to generate item responses, calculate item information, and estimate ability parameters in the simulations. Table 1 shows descriptive statistics for each item parameter. The current study followed Leung, Chang, and Hau's (2002) approach for generating ability parameters. First, true ability parameters (θ) were generated between -3 and 3 with equal intervals of 0.4, resulting in a vector of 16 unique θ values (i.e., $\theta = [-3, -2.6, \ldots, 2.6, 3]$). Next, 500 examinees were simulated for each ability point, and the total sample size was 8000 (i.e., 500 examinees x 16 ability points). Finally, a response matrix $\mathbf{A}$ with 8000 examinees for the item banks with 300 items (8000×300) and 600 items (8000×600) was simulated based on the item bank and true ability parameters.

Table 1. Descriptive Statistics for the Item Parameters in the Item Banks.

Bank size	Item parameter	*M*	*SD*	*Min*	*Max*
300	*A*	1.696	0.592	0.650	3.704
	B	−0.293	1.036	−2.873	2.770
	C	0.241	0.032	0.060	0.299
600	*A*	1.699	0.609	0.645	3.704
	b	−0.326	1.015	−3.000	2.770
	c	0.242	0.030	0.060	0.2999

The OMST Design

For the first stage in OMST, an automatic test assembly process based on mixed integer programming (van der Linden, 1998) was used for creating three equivalent modules. Then, each examinee was randomly assigned to one of the three modules, and their responses to the items in the selected modules were selected from the response matrix **A**. For all methods, modules in the first stage maximized the test information function between $\theta = -0.8$ and $\theta = 0.8$ (e.g., Verschoor & Eggen, 2014). Modules in the subsequent stages were adaptive (i.e., built on the fly based on provisional ability estimates). For MFI, either 10 items (for 30-item design) or 20 items (for 60-item design) maximizing the information at the latest provisional ability level were selected to compose adaptive modules in the subsequent stages. As for MFI-R, 10 or 20 items were selected randomly from 20 or 40 most informative items to create the adaptive modules.

As explained earlier, the CF methods require a training dataset to find similar examinees or items before recommending any items for the target examinee. In this study, two training datasets were created based on a sample of 2000 examinees for the item banks with 300 items (2000 × 300) and 600 items (2000 × 600). After generating ability parameters between −3 and 3 with equal intervals of 0.4, Equation (1) was used to calculate the expected Fisher information $I(\theta)$ for the entire item bank with the mirt package in R, yielding training datasets without any missing attributes. However, one could also take a more conservative approach by creating a training dataset where the Fisher information is only calculated for the items with valid responses, but not for not-administered items, yielding a sparse dataset. Depending on the sparsity level in the training dataset, the accuracy of the predictions produced by the CF methods can decrease significantly (Huang et al., 2004; Ma et al., 2007). To evaluate the performance of the CF methods in the presence of missing values, we created two sparse datasets by randomly deleting 20% and 50% of the Fisher information values. Then, we used the UBCF and IBCF algorithms for item selection in the second and third stages of the OMST using the complete training dataset and the sparse training datasets with 20% missingness (UBCF-20 and IBCF-20) and 50% missingness (UBCF-50 and IBCF-50). The expected a posteriori method was used for estimating ability parameters after each stage of OMST. 100 replications were conducted across all simulation factors.

Evaluation Criteria

The performances of the four item selection methods (MFI, MFI-R, UBCF, and IBCF) were evaluated based on the accuracy of final ability estimates under each simulation factor. Evaluation criteria included bias, root mean square error (RMSE), and reliability statistics (Lin, 2021). Bias, RMSE, and reliability values for each replication were calculated as follows

$$Bias = \frac{\sum_{i=1}^{N}\left(\overline{\theta}_i - \theta_i\right)}{N}, \tag{6}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\overline{\theta}_i - \theta_i\right)^2}{N}}, \text{and} \tag{7}$$

$$Reliability = \left(\frac{cov\left(\overline{\theta}_i, \theta_i\right)}{\sigma_{\overline{\theta}_i}\sigma_{\theta_i}}\right)^2, \tag{8}$$

where $\overline{\theta}_i$ represents examinee i's final ability estimate, θ_i represents examinee i's true ability level, and N represents the sample size, $cov(\overline{\theta}_i, \theta_i)$ represents the covariance between estimated and true ability levels, and $\sigma_{\overline{\theta}_i}$ and σ_{θ_i} represent the standard deviations of $\overline{\theta}_i$ and θ_i, respectively. Smaller values of RMSE and the absolute value of bias and larger values of reliability indicated more accurate ability estimates. Positive bias values indicated overestimated ability levels, whereas negative bias values represented underestimated ability levels.

In addition to the accuracy of final ability estimates, the item bank utilization was also examined based on the maximum item usage rate and the proportion of unused items. Item usage was rate calculated based on the proportion of the number of examinees who answered the item to the total sample size. The maximum value across all items was the maximum item usage rate (Zheng & Chang, 2015). The proportion of unused items was calculated based on the proportion of the number of unselected items to the total number of items in the item bank. The higher the proportion of unused items, the worse the item bank utilization.

Results

Accuracy of Ability Estimates

Table 2 shows the average bias, RMSE, and reliability statistics across all simulation conditions. The average bias values for UBCF, UBCF-20, MFI, and MFI-R methods under all conditions were negative, indicating that they yielded slightly underestimated ability levels. The smallest absolute bias value occurred when adopting UBCF-50 for the 30-item design or 60-item design, and the item bank consisted of 300 items, or adopting IBCF-20 for the 60-item design when the item bank size was 300. Across all conditions, the UBCF methods (i.e., UBCF, UBCF-20, and UBCF-50)

Table 2. Average Values of Bias, RMSE, and Reliability for the Item Selection Methods.

Item Bank Size	Method	30-item design			60-item design		
		Bias	RMSE	Reliability	Bias	RMSE	Reliability
300	UBCF	−0.019	0.362	0.970	−0.013	0.277	0.981
	IBCF	−0.007	0.428	0.962	0.002	0.341	0.974
	UBCF-20	−0.018	0.363	0.970	−0.013	0.277	0.981
	IBCF-20	0.005	0.436	0.961	0.001	0.342	0.974
	UBCF-50	−0.001	0.384	0.968	−0.003	0.291	0.980
	IBCF-50	−0.001	0.431	0.962	0.006	0.349	0.973
	MFI	−0.016	0.369	0.969	−0.013	0.279	0.981
	MFI-R	−0.021	0.393	0.966	−0.012	0.297	0.979
		−0.010	0.341	0.973	−0.010	0.256	0.984
600	UBCF	0.044	0.464	0.957	0.015	0.335	0.975
	IBCF	−0.010	0.341	0.973	−0.010	0.256	0.984
	UBCF-20	0.044	0.466	0.957	0.018	0.341	0.974
	IBCF-20	0.020	0.386	0.968	0.010	0.286	0.980
	UBCF-50	0.011	0.417	0.964	0.012	0.331	0.976
	IBCF-50	−0.011	0.350	0.972	−0.011	0.257	0.983
	MFI	−0.012	0.365	0.970	−0.011	0.270	0.982
	MFI-R	−0.019	0.362	0.970	−0.013	0.277	0.981

Note. UBCF: UBCF learning from a training response dataset with no missing values; IBCF: IBCF learning from a training response dataset with no missing values; UBCF-20: UBCF learning from a training response dataset with 20% missingness; IBCF-20: IBCF learning from a training response dataset with 20% missingness; UBCF-50: UBCF learning from a training response dataset with 50% missingness; IBCF-50: IBCF learning from a training response dataset with 50% missingness; MFI: Maximum Fisher information; MFI-R: MFI with random item selection.

provided more accurate results than the IBCF methods (i.e., IBCF, IBCF-20, and IBCF-50) in terms of RMSE and reliability. UBCF also outperformed both MFI and MFI-R in terms of RMSE, although the differences among these methods were mostly negligible in terms of reliability.

Increasing the item bank size (from 300 items to 600 items) and test length (from 30 items to 60 items) improved the performance of all the item selection methods based on the average RMSE and reliability values. The proportion of missingness in the training data had a negligible impact on UBCF and IBCF in the small item bank condition (i.e., 300 items). However, for the large item bank condition (i.e., 600 items), using a training dataset with 20% and 50% missingness improved the performance of IBCF. However, it deteriorated the performance of UBCF, regardless of the test length. A possible reason for this finding is that IBCF could still capture item similarity accurately with a larger item bank and offset the effects of missing values in the training data. In contrast, user similarity matching through UBCF became less accurate due to incomplete user profiles in the training dataset.

Figures 1 and 2 present bias and RMSE values from the item selection methods at each ability point (i.e., $\theta = [-3, -2.6, \ldots, 2.6, 3]$). The results show that the item selection methods had similar patterns for ability estimation across different test lengths and item bank sizes. Bias and RMSE values were small within the range of $\theta = -2$ and $\theta = 2$, indicating higher accuracy in ability estimation. However, lower levels of true ability parameters were overestimated, whereas higher levels of true ability parameters were underestimated. Also, the variation around each ability point was consistent across the item selection methods.

Item Bank Utilization

Item bank utilization was evaluated based on two criteria: the maximum item usage rates and the proportion of unused items. Table 3 presents the item bank utilization results for each item selection method. When the test length was 60 items, MFI-R (i.e., MFI with randomesque) outperformed the other item selection methods based on the maximum item usage rates and the proportion of unused items in the item bank. However, when the test length was 30 items, the results were mixed. MFI-R produced the smallest values for the maximum usage rates, while the IBCF methods yielded the smallest values for the proportion of unused items in the item bank. Although the UBCF and IBCF methods produced similar results regarding the maximum item usage rates, IBCF outperformed UBCF in terms of the proportion of unused items. UBCF-50 (i.e., UBCF with 50% missingness in the training dataset) was the worst-performing method regarding both the maximum item usage rate and the proportion of unused items.

The sparsity of the training dataset had different effects on the CF methods in terms of the item usage control. Specifically, UBCF-20 controlled the maximum item usage rates more effectively than UBCF-50. In contrast, IBCF-20 performed worse than IBCF-50 regarding the maximum item usage rates. Increasing the test length from 30 items to 60 items increased the maximum item usage rates because the chance of selecting the same item increased in the longer test. Also, regarding the proportion of unused items, increasing the sparsity of the training dataset from 20% to 50% deteriorated the performance of UBCF (i.e., UBCF-20 and UBCF-50) but had a negligible impact on IBCF (i.e., IBCF-20 and IBCF-50). When the test length was increased from 30 to 60 items, the proportion of unused items decreased for all item selection methods. Similarly, increasing the item bank size resulted in worse item bank utilization for all item selection methods because more unique items were selected from a larger item bank.

Discussion

The OMST framework proposed by Zheng and Chang (2015) has motivated several researchers and practitioners who aim to design and implement better adaptive tests (Du, Li, & Chang, 2019).

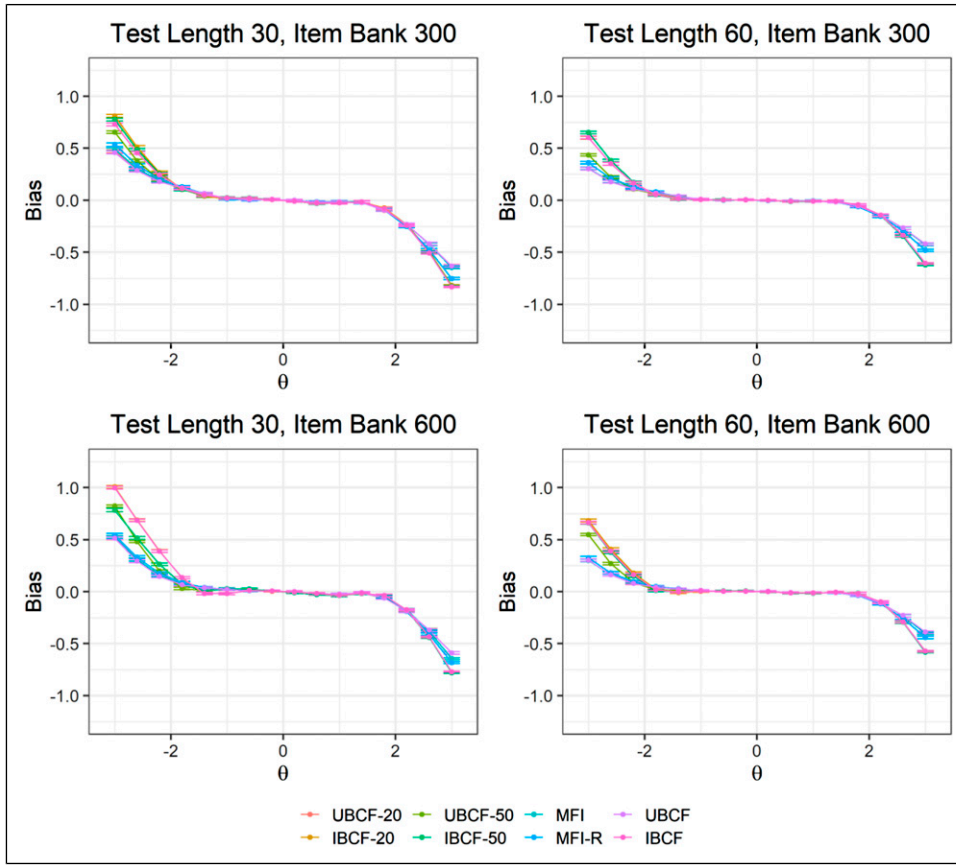

Figure 1. Bias values for the item selection methods across the ability points.

Among all the elements of adaptive testing, the item selection procedure plays a highly critical role and thus deserves to be investigated in more depth. Previous research has already investigated the effects of using MFI as an item selection method on the accuracy of ability estimation and item exposure rates in traditional adaptive tests (Chang & Ying, 1999). However, item selection methods in the OMST design still need to be explored. Therefore, this study proposed new item selection methods for the OMST design based on the CF algorithms. Many researchers demonstrated the superior performance of the CF algorithms (UBCF and IBCF) in selecting and recommending items in the context of intelligent recommender systems (e.g., Li et al., 2016). This study utilized the user-based and item-based forms of the CF algorithms (i.e., UBCF and IBCF) as potential item selection methods in the OMST design. In addition, this study proposed a combination of MFI with a randomesque method (called MFI-R) to improve the item bank utilization and compared the CF methods with MFI and MFI-R based on the accuracy of ability estimates and item bank utilization.

The results indicated that with the complete training dataset and a training dataset with 20% missingness, UBCF performed the best in terms of the accuracy of ability parameters. However, when the sparsity level increased to 50%, the performance of UBCF was less accurate than MFI and MFI-R. These findings, while preliminary, suggest that implementing the UBCF method as an item selection method can yield accurate results in OMST. This study also compared item bank

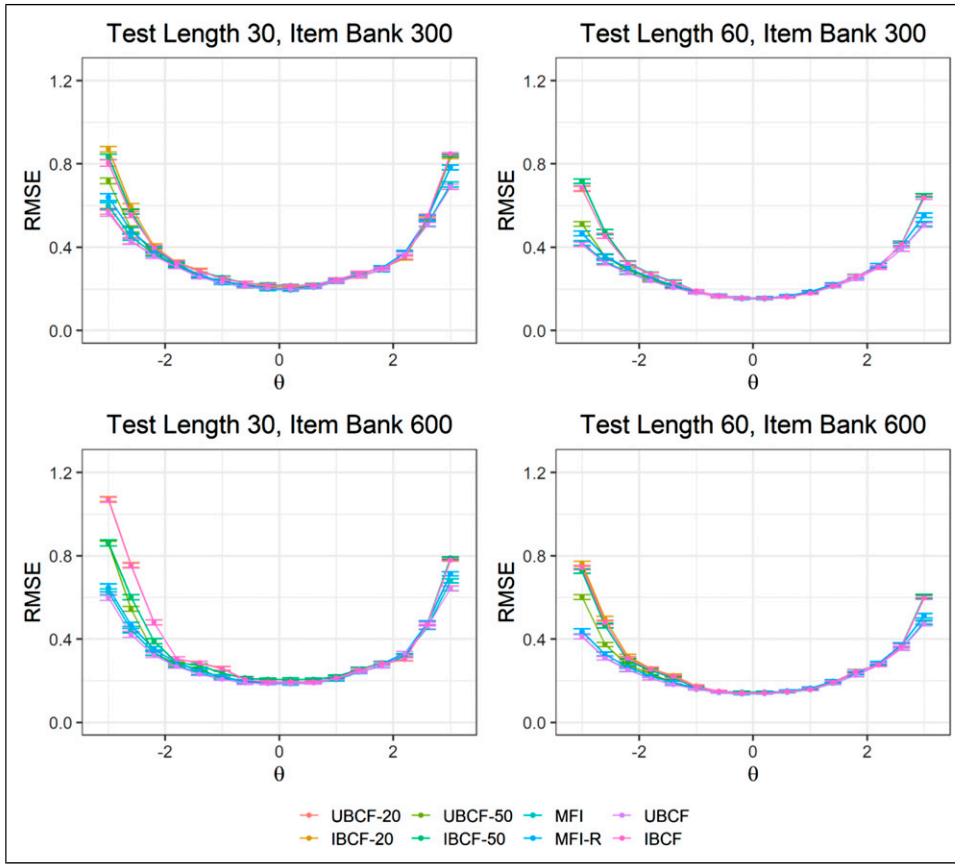

Figure 2. RMSE values for the item selection methods across the ability points.

utilization across different item selection methods based on the maximum item usage rate and the proportion of unused items in the item bank. When the test length was short (i.e., 30 items), IBCF outperformed the other methods, including MFI-R, regarding item bank utilization. However, when the test length was increased (i.e., 60 items), MFI-R produced the best item bank utilization results. Overall, the results showed a significant trade-off between the accuracy of ability parameters and the maximum item usage rates because exposing the same items to many examinees yielded more accurate ability estimates at the expense of increased item exposure rates (Zheng & Chang, 2015). Low rates of unused items also indicated that both IBCF and MFI-R are highly effective in increasing the usage of different items in the item bank.

Overall, the current study demonstrated the feasibility of using the CF algorithms as item selection methods under the OMST framework. The UBCF method can produce accurate ability parameter estimates comparable to those from the traditional item selection methods (MFI and MFI-R). Our findings also suggest that the IBCF method can utilize the item bank more effectively at the cost of sacrificing measurement accuracy. In the OMST design, we recommend the UBCF method for testing conditions where the highest priority is to estimate accurate ability parameters and the IBCF method for adaptive testing programs that prioritize reducing the number of unused items in the item bank while controlling for item exposure rates. Our findings also indicate that the performance of the CF methods relies on the conditions of the training dataset. For example, using

Table 3. Average Values of Item Bank Utilization Indices for the Item Selection Methods.

		30-item design		60-item design	
Item Bank size	Method	Maximum item usage rates	Proportion of unused items	Maximum item usage rates	Proportion of unused items
300	UBCF	0.826	0.566	0.823	0.307
	IBCF	0.819	0.377	0.825	0.269
	UBCF-20	0.820	0.575	0.831	0.314
	IBCF-20	0.815	0.386	0.825	0.269
	UBCF-50	0.830	0.623	0.840	0.397
	IBCF-50	0.817	0.371	0.823	0.254
	MFI	0.824	0.587	0.827	0.317
	MFI-R	0.767	0.468	0.782	0.153
600	UBCF	0.826	0.747	0.827	0.568
	IBCF	0.814	0.658	0.815	0.510
	UBCF-20	0.827	0.758	0.826	0.584
	IBCF-20	0.814	0.654	0.816	0.503
	UBCF-50	0.821	0.788	0.832	0.629
	IBCF-50	0.806	0.600	0.812	0.462
	MFI	0.815	0.756	0.819	0.585
	MFI-R	0.766	0.668	0.770	0.445

Note. UBCF: UBCF learning from a training response dataset with no missing values; IBCF: IBCF learning from a training response dataset with no missing values; UBCF-20: UBCF learning from a training response dataset with 20% missingness; IBCF-20: IBCF learning from a training response dataset with 20% missingness; UBCF-50: UBCF learning from a training response dataset with 50% missingness; IBCF-50: IBCF learning from a training response dataset with 50% missingness; MFI: Maximum Fisher information; MFI-R: MFI with random item selection.

a highly sparse training dataset may negatively affect the accuracy of estimated ability parameters. Also, the similarity between the examinees in the training dataset and the target examinees taking the test can affect the quality of the training process for the CF methods and thereby influence their performance in the item selection process.

Limitations and Future Research

This study has several limitations. First, the present study compared the CF methods with MFI and MFI-R based on measurement accuracy and item bank utilization. However, other non-statistical constraints (e.g., answer key balancing and content balancing) were not considered. Standardized tests need to have a similar content distribution and accurate ability estimates for all examinees (van der Linden, 2005). Therefore, future studies are needed to investigate the performance of the CF methods when both statistical and non-statistical constraints are considered in the OMST design. Second, previous studies also developed several item selection methods to better control content balancing, such as the maximum priority index method (Cheng & Chang, 2009) and the weighted-deviations method (Stocking & Swanson, 1993). Future studies can involve item selection methods with content balancing capabilities when examining the performance of the CF methods. Third, the CF methods require a training dataset to learn, which means the items must be pretested or calibrated using on-the-fly calibration (Kingsbury, 2009; Verschoor et al., 2019).

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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jiaying Xiao https://orcid.org/0000-0001-9513-6477
Okan Bulut https://orcid.org/0000-0001-5853-1267

Notes

1. It should be noted that this information is only applicable to the Rasch, 1PL, and 2PL IRT models. For more complex IRT models (e.g., 3PL and 4PL), when the Fisher information is maximized, the probability of answering the item correctly may not be equal to 50%.

References

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement*, *45*(1), 3–21. https://doi.org/10.1177/0146621620965730

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, 19–21 June 2021, (pp. 95–102). http://educationaldatamining.org/EDM2012/uploads/procs/Full_Papers/edm2012_full_4.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Biswas, S., Lakshmanan, L. V., & Roy, S. B. (2017). Combating the cold start user problem in model based collaborative filtering. ArXiv, abs/1703.00397. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Madison, WI, 24-26 July 1998, (pp. 43–52).

Bulut, O., Cormier, D. C., & Shin, J. (2020). An intelligent recommender system for personalized test administration scheduling with computerized formative assessments. *Frontiers in Education*, *5*, 1–11. https://doi.org/10.3389/feduc.2020.572612

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chang, H.-H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, *73*(3), 441–450. https://doi.org/10.1007/s11336-007-9047-7

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*(2), 369–383. https://doi.org/10.1348/000711008X304376

de Schipper, E., Feskens, R., & Keuning, J. (2021). Personalized and automated feedback in summative assessment using recommender systems. *Frontiers in Education*, *6*, 77. https://doi.org/10.3389/feduc.2021.652070

Du, Y., Li, A., & Chang, H. H. (2019). Utilizing response time in on-the-fly multistage adaptive testing. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.) *Quantitative psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics* (Vol. 265, pp. 107–117). Springer.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports 2001–1. CITO Groep.

Hahsler, M. (2015). *Recommenderlab: A framework for developing and testing recommendation algorithms*. http://CRAN.R-project.org/package=recommenderlab

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44–52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hu, Y., Shi, W., Li, S., & Hu, X. (2017). Mitigating data sparsity using similarity reinforcement-enhanced collaborative filtering. *ACM Transactions on Internet Technology*, *17*(3), 1–20. https://doi.org/10.1145/3062179

Huang, Z., Chen, H., & Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 116–142. https://doi.org/10.1145/963770.963775

Kim, J., & McLean, J. E. (1995). The influence of examinee test-taking motivation in computerized adaptive testing. [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Francisco , CA.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN, USA, 2 June 2009. http://iacat.org/sites/default/files/biblio/cat09kingsbury.pdf

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375. https://doi.org/10.1207/s15324818ame0204_6

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*(4), 376–392. https://doi.org/10.1177/014662102237795

Li, S., Karatzoglou, A., & Gentile, C. (2016). Collaborative filtering bandits. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016 (pp. 539–548). ACM.

Lin, Y. (2021). Reliability estimates for IRT-based forced-choice assessment scores. *Organizational Research Methods*, *25*(3), 575–590. https://doi.org/10.1177/1094428121999086

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. *Decision Support Systems*, *74*, 12–32. https://doi.org/10.1016/j.dss.2015.03.008

Luo, X. (2016). xxIRT: Practical item response theory and computer-based testing in R [Computer software].

Ma, H., King, I., & Lyu, M. R. (2007). Effective missing data prediction for collaborative filtering Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain (pp. 39-46).

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223–236). Academic Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In 10th international Conference on World Wide Web, Hong Kong, 1–5 May 2015 (pp. 285–295).

Shin, J., & Bulut, O. (2021). Building an intelligent recommendation system for personalized test scheduling: A reinforcement learning approach. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01602-9

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm*. Educational Testing Service. Technical Report RR 93-2).

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, *17*(3), 277–292. https://doi.org/10.1177/014662169301700308

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the military testing association, San Diego, CA, 21–25 October 1985 (pp. 973–977). Navy Personnel Research and Development Center.

Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* [Doctoral dissertation, University of Illinois at Urbana-Champaign, USA]. https://www.proquest.com/openview/1bd7b72d39c9e19966faaa362bff55c0/1?pq-origsite=gscholar&cbl=18750

Thai-Nghe, N., Drumond, L., Horváth, T., Krohn-Grimberghe, A., Nanopoulos, A., & Schmidt-Thieme, L. (2012). Factorization techniques for predicting student performance. In O. C. Santos, & J. G. Boticario (Eds.), *Educational recommender systems and technologies: Practices and challenges* (pp. 129–153). IGI Global.

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 101–134). Erlbaum.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320–332. https://doi.org/10.1037/0021-9010.87.2.320

Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. In *KDD cup 2010: Improving cognitive models with educational data mining*.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*(3), 195–211. https://doi.org/10.1177/01466216980223001

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, *42*(3), 283–302. https://doi.org/10.1111/j.1745-3984.2005.00015.x

Verschoor, A., Berger, S., Moser, U., & Kleintjes, F. (2019). On-the-fly calibration in computerized adaptive testing. In *Theoretical and practical advances in computer-based educational measurement* (pp. 307–323). Springer.

Verschoor, A. J., & Eggen, T. J. H. M. (2014). Optimizing the test assembly and routing for multistage testing. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing. Theory and applications* (pp. 135–150). CRC Press.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20. https://doi.org/10.1111/j.1745-3992.1993.tb00519.x

Wang, S., Fellouris, G., & Chang, H. H. (2017). Computerized adaptive testing that allows for response revision: Design and asymptotic theory. *Statistica Sinica*, *27*(4), 1987–2010. https://doi.org/10.5705/ss.202015.0304

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45–62. https://doi.org/10.1111/jedm.12100

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, *2*(1), 1–17. https://doi.org/10.7333/1401-0201001

Zhao, X. (2016). *Cold-start collaborative filtering* [Doctoral dissertation, University College London]. https://discovery.ucl.ac.uk/id/eprint/1474118/

Zheng, Y., & Chang, H.-H. (2011). Automatic on-the-fly assembly for computer adaptive multistage testing. [Paper presentation]. Annual Meeting of the National Council of Measurement in Education, San Francisco, CA.

Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519