

Collaborative Filtering Item Selection Methods for On-the-Fly Assembled Multistage Adaptive

Testing

by

Jiaying Xiao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in

Measurement, Evaluation, and Data Science

Department of Educational Psychology
University of Alberta

© Jiaying Xiao, 2019

Abstract

An important issue in the design and implementation of adaptive testing is the use of an appropriate item selection method. This study employed Collaborative Filtering (CF) methods for selecting items under *on-the-fly assembled multistage adaptive testing* (OMST) framework. Traditional item selection methods were compared to CF methods concerning the accuracy of ability estimation and item bank utilization. The simulation results indicated that CF item selection methods were comparable to traditional item selection methods (Maximum Fisher Information and α -stratification) for ability estimates. Furthermore, CF methods indicated more superior performance in terms of item bank utilization. Limitations for the current study, as well as directions for future research, are discussed.

Preface

This thesis is an original work by Jiaying Xiao. No part of this thesis has been previously published.

Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Okan Bulut. Without your support and help, it is impossible for me to finish this work. Your patient guidance has promoted my development in programming skills and your deep insights have sparked my interest in the educational measurement field. Also, I would like to thank my committee members, Drs. Maria Cutumisu and Damien Cormier for their time and advice.

Moreover, I would like to thank all my friends who always encourage me and keep my company. Finally, I would extend my gratitude to my parents for their unconditional love and support in this journey. I wish you all the best in the future.

Table of Contents

Abstract	ii
Preface	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
Chapter 1: Introduction	1
Background	2
Purpose of Current Study	5
Chapter 2: Literature Review	7
Overview of MST	7
Overview of OMST	9
Overview of Maximum Fisher Information Method	10
Overview of A-stratification Method	11
Overview of Collaborative Filtering	12
User-based collaborative filtering (UBCF)	13
Item-based collaborative filtering (IBCF)	14

Chapter 3: Method	18
Overview of Design.....	18
Data Generation.....	19
Automated Test Assembly (ATA) Procedure.....	20
Item Selection Methods	23
Data Analysis	25
Chapter Summary.....	26
Chapter 4: Results.....	28
Measurement Accuracy Results	28
Item Bank Utilization Results	31
The Proportion of Remaining Examinees Results	32
Chapter Summary.....	33
Chapter 5: Discussion	34
Results for Research Question 1	35
Results for Research Question 2	36
Results for Research Question 3	36
Conclusion	37
Limitations of the Study and the Directions for Future Research.....	37
References	39

List of Tables

Table 1 Distributions of Item Parameters.....	19
Table 2 Mean Values of Measurement Accuracy Indices for Different Item Selection Methods	29
Table 3 Mean Values of Item Bank Utilization Indices for Different Item Selection Methods....	32
Table 4 Mean Values of the Proportion of Remaining Examinees for Different Item Selection Methods.....	33

List of Figures

Figure 1. Panel #1 with 1-3-3 MST.	8
Figure 2. A general framework of OMST.	10
Figure 3. A general CF process.	12
Figure 4. A user-item rating matrix \mathbf{R} and estimated ratings for the active user.	14
Figure 5. An example of item-to-item similarity matrix \mathbf{S} and the calculation process of IBCF. .	16
Figure 6. The frequency distribution histogram of discrimination parameters.	20
Figure 7. The test information plots for parallel modules under the maximum test information design.	21
Figure 8. The test information plots for parallel modules under the b-parameter close to 0 design.	22
Figure 9. A brief description of the entire OMST process.	22
Figure 10. Mean RMSE values by different item selection method and test structure.	30
Figure 11. Mean correlation values by different item selection method and test structure.	30

Chapter 1: Introduction

With the rapid advancement in the computer and information technologies, more and more large-scale testing programs, such as the Graduate Management Admission Test (GMAT), have made the transition from traditional paper-and-pencil testing to computerized adaptive testing (CAT) over the past 20 years. However, in both application and research, CAT indicated several problems, such as providing unreliable scores in the GRE CAT system (Carlson, 2000), not allowing examinees to review completed items or skip items, and lack of control over the context effects (Hendrickson, 2007). Due to these shortcomings, there has been an increasing interest in replacing CAT with multistage adaptive testing (MST). MST refers to a group sequential design, in which items are grouped into modules that are matched to an examinee's provisional ability estimates (Chang, 2015). Previous research indicated several benefits of MST over a traditional CAT application. First, MST can alleviate the problem of overestimation or underestimation of ability levels at the beginning of a test (Zheng & Chang, 2014). Second, through its panel design, MST allows examinees to navigate back and forth inside a module and revisit their questions, which may decrease test anxiety among examinees. Third, MST enables better control over non-statistical design properties of tests – such as content specifications (Hendrickson, 2007).

While MST has many advantages over traditional adaptive testing, it also has its limitations, including the necessity to have several pre-assembled modules, which may not target some examinees properly (Tay, 2015). Therefore, researchers developed a new adaptive testing design called *on-the-fly assembled multistage adaptive testing* (OMST) by combining the design features of both CAT and MST (Zheng & Chang, 2015). Like MST, OMST also adapts between

multiple stages; however, modules in every stage are assembled on the fly, as examinees move from one module to another.

Background

According to Hendrickson (2007), MST has taken several terms and forms, such as *computerized mastering testing* (CMT; Lewis & Sheehan, 1990), *two-stage testing* (Kim & Plake, 1993), *computer-adaptive sequential testing* (CAST; Luecht & Nungester, 1998), *bundled multistage adaptive testing* (BMAT; Luecht, 2003), and *multiple form structures* (MFS; Armstrong, Jones, Koppel, & Pashley, 2004).

Betz and Weiss (1973) pointed out that the early idea of MST was advanced in 1957 (Cronbach & Gleser, 1957) and after one year, the first empirical study was published (Angoff & Huddleston, 1958). But at that time, the framework of MST was based on classical test theory (CTT) and MST was only administered in paper-and-pencil settings. Then, Lord (1971) made the first attempt to establish MST under the item response theory (IRT) framework. With the innovative development of modern computing, MST was improved as an adaptive version as well as implemented online for several large-scale assessments (Luecht, 2000; Luecht, Brumfield & Briethaupt, 2006; Breithaupt, Ariel & Hare, 2010; Melican, Breithaupt & Zhang, 2010).

Typically, test developers design modules of MST at different difficulty levels using the automated test assembly (ATA) methods based on several test requirements and constraints (e.g., content specifications, number of items, desired levels of reliability). During the test, examinees receive pre-assembled modules based on their estimated ability levels (Armstrong & Roussos, 2005). However, this traditional design was often criticized due to its difficulty to create parallel pre-constructed modules to satisfy all psychometric and content constraints and its inaccurate estimation of ability levels (Sari, Yahsi-Sari, & Huggins-Manley, 2016). Therefore, Zheng and

Chang (2015) proposed OMST to assemble modules on the fly by utilizing item selection methods in CAT and results showed that this hybrid design could control several psychometric properties adequately. While OMST provides a flexible framework of sequential testing, further studies in this direction have been scant so far (Wang, Lin, Chang, & Douglas, 2016).

Previous studies have already shown that MST designs can be influenced by different scenarios, such as the number of stages (two or three stages). Patsula (1999) found that 2-stage MST designs had higher error in the ability estimates than 3-stage MST designs. Since OMST is developed based on MST designs, it is also necessary to investigate the influence of the number of stages under OMST framework. As Zheng and Chang (2015) indicated, the more stages an OMST has, the more its properties may be similar to those of traditional CAT, and the fewer stages it has, the more its properties may be similar to those of MST.

An important issue in the design and implementation of on-the-fly testing applications such as CAT and OMST is the use of an appropriate item selection algorithm because the accuracy of estimated ability levels highly depends on the selection of suitable items for each examinee. The most popular method to select a bundle of items is the *Maximum Fisher Information* method. Based on this method, a bundle of items is selected after each module according to the maximum item information based on the latest provisional ability estimates (Zheng & Chang, 2015). However, previous studies found that this method is inclined to selecting items with higher a -parameter (i.e., discrimination) values, which increases the selection of such items substantially and results in high item exposure and test security problems (Han, 2018). Therefore, researchers show enormous interests in developing other item selection algorithms and item exposure control methods. For instance, Chang and Ying (1999) proposed an approach called *a-stratification* method where the item bank is stratified based on a -parameter

values. This approach selects items with b -parameter values closest to estimated ability values as well as reserves items with higher a -parameter values in the later stages. Research showed the efficiency of this approach in minimizing item exposure rates without reducing the accuracy of ability estimation in CAT (Chang & Ying, 1999; Chang, Qian, & Ying, 2001). Although Zheng and Chang (2011) developed the item bank stratification design based on this approach and proved its effectiveness, they still indicated the need for future studies to investigate this method more thoroughly and expand its capabilities under OMST framework.

While these two methods have shown excellent performance in the literature, researchers are still dedicated to developing new item selection methods in order to better control item exposure and enhance test security (Georgiadou, Triantafillou, & Economides, 2007). Recently, there has been an upward trend in the use of data mining and machine learning methods for improving the quality of educational assessments (Nehm, Ha, & Mayfield, 2012; Petersen & Ostendorf, 2009). Recommender systems are one of the most successful and widespread applications of machine learning methods. Recommendation of news or videos for media, product recommendation, or personalization in travel and retail can be handled by recommender systems that implement machine learning algorithms such as collaborative filtering and content-based filtering. *Collaborative filtering* (CF) can recommend items based on a user's tastes (likes or dislikes) or other similar users' ratings (Sarwar, Karypis, Konstan, & Riedl, 2001). This algorithm can also be divided into two main categories: user-based (UBCF) and item-based (IBCF) approaches (Breese, Heckerman, & Kadie, 1998). The former recommends items liked by similar users and the latter recommends items similar to those users have liked or preferred in the past (Lu, Wu, Mao, Wang, & Zhang, 2015).

In educational assessments, we can consider examinees' responses as their tastes of items: an incorrect answer (0) means they disliked this item, and a correct answer (1) means that they liked this item, which enables the use of CF algorithm as an item selection method in OMST applications. Furthermore, previous research formalized the relationship between item response theory (IRT) and CF methods (Bergner et al., 2012), which suggested the feasibility to adopt CF methods in IRT contexts. Therefore, this study utilized CF methods to select items in OMST designs and investigated their performance by comparing with traditional item selection methods across different conditions.

Purpose of Current Study

Research on OMST is still emerging, and there are several ways to improve its current design to enhance the accuracy of estimating individuals' abilities and to reduce threats to item exposure and test security. Although CF methods have shown excellent estimation results in recommendation systems (e.g., Resnick et al., 1994; Linden, Smith, & York, 2003; Li, Karatzoglou, & Gentile, 2016), no study has applied CF methods to the item selection procedure in adaptive testing designs. Therefore, the purposes of the current study are twofold. First, we aim to implement UBCF and IBCF methods as item selection techniques in the OMST framework. Second, we aim to compare the effectiveness of CF methods with that of traditional item selection methods (Maximum Fisher Information and a -stratification) under 2-stage and 3-stage OMST designs. The following research questions will be addressed in the current study:

- 1) Do CF methods produce more accurate ability estimates compared to traditional item selection methods under 2-stage or 3-stage OMST designs?
- 2) Do CF methods control the item exposure better compared to traditional item selection methods under 2-stage or 3-stage OMST designs?

- 3) Do CF methods utilize more items in the item bank compared to traditional item selection methods under 2-stage or 3-stage OMST designs?

Chapter 2: Literature Review

In this chapter, I introduce some fundamental concepts and terminology about both multistage testing (MST) and on-the-fly multistage testing (OMST) frameworks. Then, item selection methods, such as Maximum Fisher Information, α -stratification and collaborative filtering, are described. Some articles about applying collaborative filtering to educational assessments are also introduced.

Overview of MST

In order to understand OMST framework, it is necessary to know the general process of MST and MST-related terms. Under MST framework, there are several *panels* (Luecht & Nungester, 1998) which are pre-assembled based on several requirements (e.g., psychometric and content constraints). The panel is the basic structure of MST. It can be divided into several stages and each stage consists of sets of items called *modules* (Luecht & Nungester, 1998) or *testlets* (Wainer & Kiely, 1987) targeted at predetermined difficulty levels. Figure 1 shows a panel under the “1-3-3” model (Luecht, Brumfield, & Breithaupt, 2006). This panel has three stages and “1-3-3” means that the first stage (also called routing stage) only has one module (i.e., 1 moderate-difficulty module), the second stage has three modules (2 easy, 2 moderate, 2 difficult modules) and the third stage has three modules (3 easy, 3 moderate, 3 difficult modules). Apart from the “1-3-3” model, there are other MST models, such as “1-3” model (Adema, 1990), “1-3-4-4” model (Luecht, 2003) and “1-2-4” model (Chen, 2011). A *pathway* is a combination of modules that are likely to be administered from a panel (Luecht & Nungester, 1998). For example, Figure 1 uses arrows to represent nine possible pathways across the three stages. Solid lines refer to primary pathways that are the most likely to be administered to examinees, and dashed lines refer to secondary pathways that may not be very common (Luecht

et al., 2006). According to some routing rules, pathways allow examinees at different ability levels to get modules targeted different difficulty levels. For example, based on the “1-3-3” model, higher proficiency examinees are more likely to receive “2 Difficult” module in the second stage and “3 Difficult” module in the third stage. In contrast, lower proficiency examinees are more likely to receive “2 Easy” module in the second stage and “3 Easy” module in the third stage (see Figure 1).

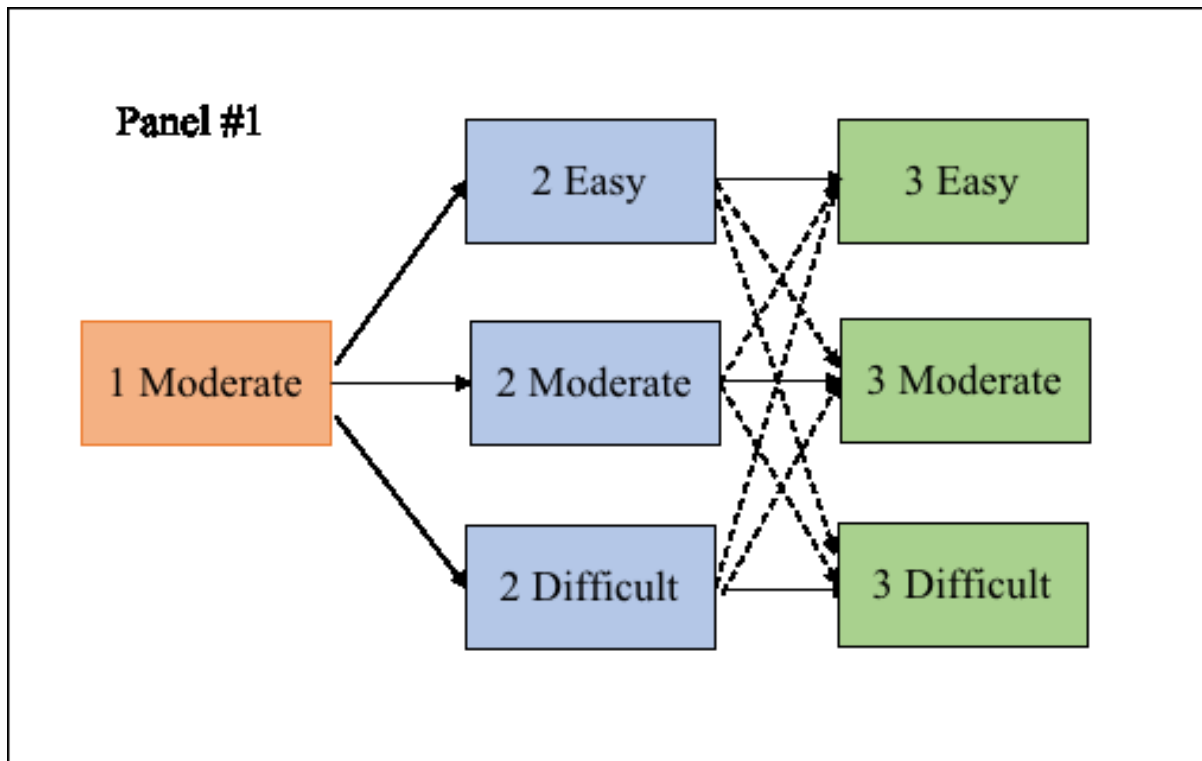


Figure 1. Panel #1 with 1-3-3 MST.

Under MST, test developers often design an automated test assembly (ATA) algorithm to assemble multiple parallel panels efficiently as well as satisfy all specifications such as the test information function (TIF), reliability, and content coverage. In a real test administration, each examinee is randomly assigned to one of the pre-assembled parallel panels and begins the test from the module at the moderate difficulty level in the first stage. After completing the moderate

module, the examinee is routed adaptively to the module at the best-matched difficulty level in the second stage based on provisional ability estimates from the first stage. A similar process takes place in each of the following stages. After completing the entire panel, the examinee can get scores based on his or her responses to all administered items.

Different from CAT where individual tests are designed for each examinee on the fly, modules of MST are pre-assembled. It is a stage-level adaptive test instead of an item-level adaptive test. Due to its preconstruction attribute, the content coverage and test security of MST can be well controlled. Its feasibility and efficiency have been proven in large-scale assessments such as the National Assessment of Educational Progress (NAEP) and the Law School Admission Test (LSAT; Bock & Zimowski, 1998; Schnipke & Reese, 1997).

Overview of OMST

Figure 2 shows a general framework of three-stage OMST proposed by Zheng and Chang (2015). Like MST, modules in the first stage are pre-assembled at a moderate difficulty level and each examinee's provisional ability estimate is calculated. But different from MST, the subsequent stages in OMST are assembled on the fly, which means that each examinee receives a different set of items in the second and third stages, based on the provisional ability parameter after each stage. In other words, OMST tests are uniquely tailored for each examinee, after the first module. For example, after an examinee finishes the second stage, the ability estimate will be updated based on responses to all items in the first and second stages, and the third stage is assembled on the fly to adapt with the new ability estimate. Finally, the final ability estimate is obtained based on responses to all administered items.

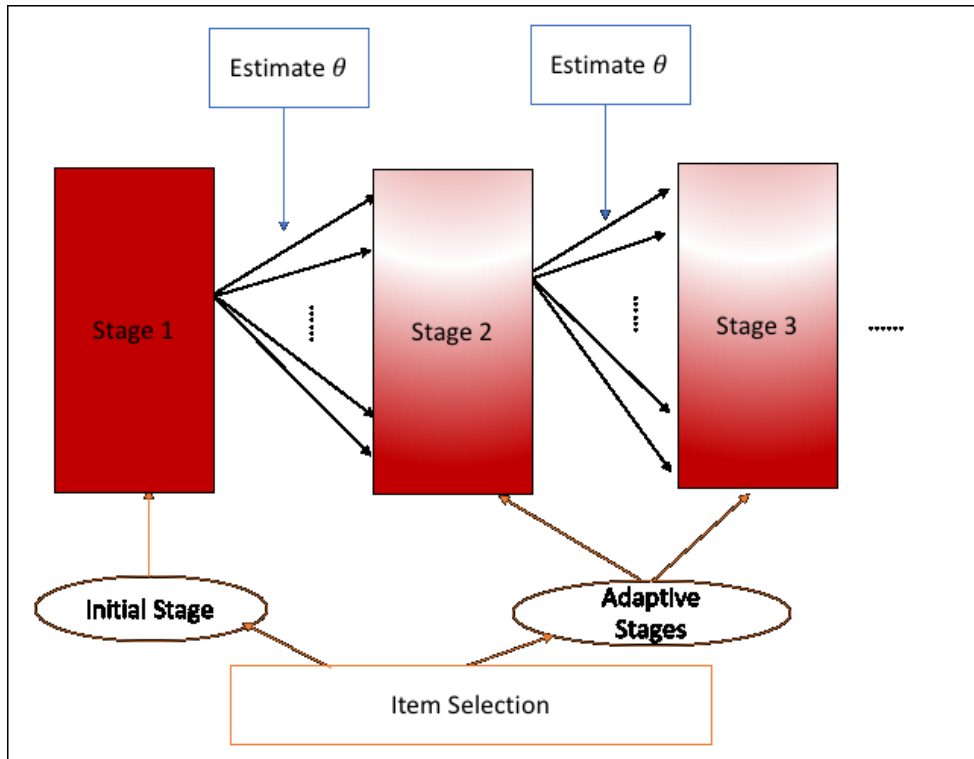


Figure 2. A general framework of OMST.

In summary, OMST framework is flexible in many aspects including but not limited to the number of items in each module and the number of stages. In OMST framework, the selection of items in on-the-fly stages can be done based on test specifications such as the maximum test information, exposure control, and content balancing.

Overview of Maximum Fisher Information Method

Maximum Fisher Information can be used to select items when the test does not have non-statistical constraints, such as content balancing. It was proposed by Birnbaum (1968) to explain the information function for dichotomous items. It describes how much an item contributes to the quality of ability estimation. In a three-parameter logistic (3PL) model, an item's information at a given ability level θ is expressed 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 refers to the discrimination parameter, b refers to the difficulty parameter, c refers to the pseudo-guessing parameter. In each stage under OMST framework, a set of items are selected one by one as a module to maximize the information at the latest provisional ability level (Zheng & Chang, 2015). The formula 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 relatively low. Due to this attribute, item selection based on Maximum Fisher Information 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 lower test security, as commonly used items can be memorized by examinees (Chang & Ying, 1999).

Overview of A-stratification Method

The a -stratification method was proposed by Chang and Ying (1999) to reduce or equalize exposure rates for items in adaptive testing. If the test has K stages ($k = 1, 2, \dots, K$), the item bank will be partitioned into K levels based on items' a values, which means items with similar a values will be grouped together. Items with relatively lower a values are used in the early stages of the test, and those with higher a values are saved for the later stages. In the k^{th} stage, only items from the k^{th} level will be selected and the priority is based on the similarity between b and the latest provisional ability level.

Evidently, this method can increase exposure rates for lower discriminating items and decrease exposure rates for higher discriminating items. In addition, it is unnecessary to use items with a high discrimination power at the early stages of the test because the examinee's true ability level is still unclear at the beginning of the test (Chang, 2015). Also, utilizing higher

discriminating items at the early stages may cause “big jumps” for ability estimation (Chang & Ying, 2008). Due to these concerns, a -stratification is a frequently used method for controlling exposure rates in adaptive tests.

Overview of Collaborative Filtering

The aim of collaborative filtering (CF) is to predict ratings and/or recommend a list of items (top- N recommendation list) to a particular user, called the active user (u_a) (Hahsler, 2015). In a typical CF scenario, there is a $m \times n$ user-item rating matrix \mathbf{R} where each row represents a user and each column represent an item. r_{jk} refers to the user u_j 's rating for item i_k . This user-item rating matrix often includes a large number of missing values and a set of unknown items to user u_a is denoted as I_a . Calculating missing ratings in I_a from other data in \mathbf{R} is called *prediction*. Then, *recommendation* process is to create a top- N list which includes the N items with the highest predicted ratings (Sarwar et al., 2001). The CF process is shown in Figure 3. In general, researchers divide CF algorithms into two categories, user-based (memory-based) CF and item-based (model-based) CF (Breese, Heckerman, & Kadie 1998).

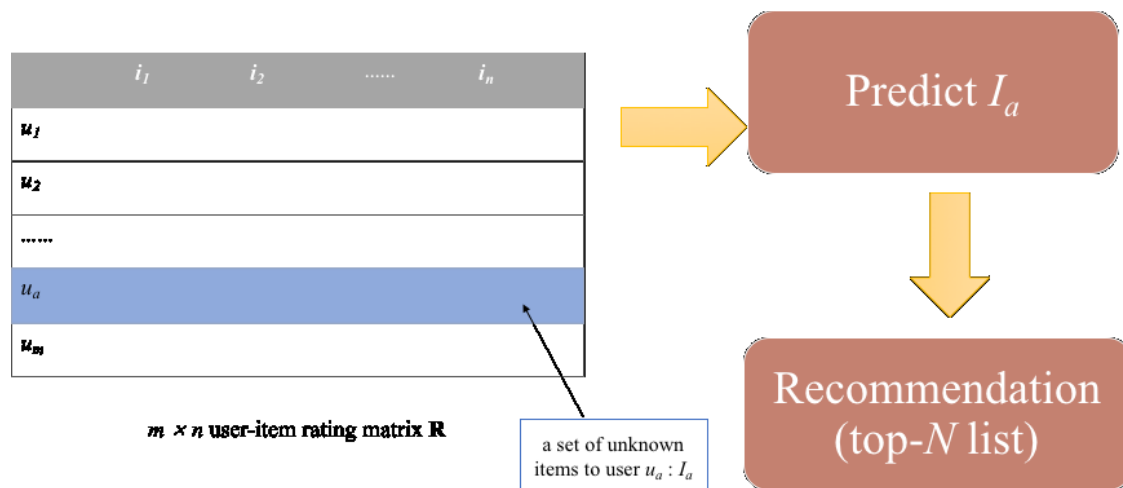


Figure 3. A general CF process.

User-based collaborative filtering (UBCF). UBCF utilizes the entire dataset \mathbf{R} to search for similar users called *neighbors* who rate items similarly as the active user u_j . The similarity can be measured by the *Cosine similarity* and the *Pearson correlation coefficient*. The *Cosine similarity* can be calculated as

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

and the *Pearson correlation coefficient* can be calculated as

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

where x, y denote two users' ratings, I_{xy} denotes the set of items rated by both users and \bar{x} and \bar{y} denote the average ratings for the users. Then *neighbors* for the user u_j can be selected by either taking the k nearest neighbors or setting a similarity threshold. Once these *neighbors* are identified, UBCF algorithm combines their ratings to form a prediction or top- N list. The easiest way to aggregate them is to average *neighbors*' ratings. Figure 4 shows a rating matrix \mathbf{R} with 5 users and 8 items and estimated ratings. Firstly, the similarity between the active user u_a and other users was calculated to get the k nearest neighbors. In this example, three neighbors (u_1, u_3 and u_4) were selected. Then, their average ratings for each item (i_3, i_4, i_5 and i_6) not rated by the active user u_a were calculated to get the estimated ratings. A top- N list was created based on the estimated ratings. Items i_6 and i_3 were recommended to the active user u_a (when $N = 2$). Although this approach is prominent, it has several shortcomings such as data sparsity, scalability, and expensive computation (Hahsler, 2015; Sarwar et al., 2001; Wang, De Vries, & Reinders, 2006).

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	4.0	2.0	NA	1.0	NA	4.0	NA	2.0
u_2	NA	NA	1.0	NA	NA	1.0	1.0	NA
u_3	NA	1.0	4.0	1.0	2.0	NA	4.0	2.0
u_4	NA	2.0	3.0	2.0	NA	NA	3.0	3.0
u_5	NA	1.0	3.0	5.0	1.0	NA	NA	NA
u_a	4.0	1.0	NA	NA	NA	NA	5.0	3.0
\overline{R}_a			3.5	1.3	2.0	4.0		

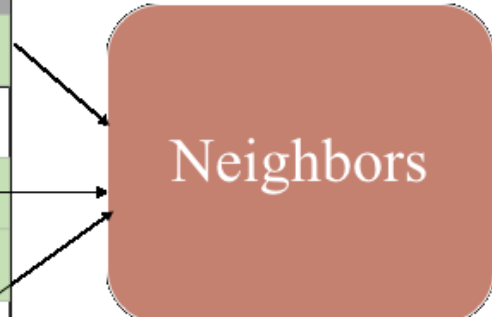


Figure 4. A user-item rating matrix \mathbf{R} and estimated ratings for the active user.

Item-based collaborative filtering (IBCF). Instead of constructing a user-item rating matrix in UBCF, IBCF calculates a $n \times n$ item-to-item similarity matrix \mathbf{S} . The underlying assumption of IBCF is that users have a preference to items that are similar to those that they like. The similarity computation $s_{i,j}$ between two items i and j is only based on ratings from users who have rated both of these items. In the example from Figure 4, the similarity $s_{3,4}$ between two items i_3 and i_4 was calculated only based on ratings from users u_3 , u_4 and u_5 . *Cosine similarity* and the *Pearson correlation coefficient* can also be used to calculate item similarities, but *Cosine similarity* should be adjusted to offset user differences by subtracting each user's average ratings separately (Sarwar et al., 2001). The new formula is calculated as follows:

$$sim(i, j) = \frac{\sum_{u \in U} (R_{u,i} - \overline{R}_u)(R_{u,j} - \overline{R}_u)}{\sqrt{\sum_{u \in U} (R_{u,i} - \overline{R}_u)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \overline{R}_u)^2}} \quad (4)$$

where $R_{u,i}$, $R_{u,j}$ denote the ratings of user u on item i and j , \overline{R}_u denotes the u th user's average ratings. In order to improve the space complexity and save time, only the k most similar items for

each item are stored, where k is smaller than the number of items. k refers to the *model size*. The next step is to check how many of those k items are rated by the active user u_a . Finally, the active user's ratings for new items can be predicted by computing the weighted sum of the ratings by this user on similar items. Equation 5 presents how to calculate the prediction rating for the item j from the active user u_a :

$$P_{aj} = \frac{\sum_i (s(j,i)v_{ai})}{\sum_i |s(j,i)|}, \quad (5)$$

where i denotes items that are similar to item j , $s(j, i)$ denotes the similarity between items j and i , v_{ai} denotes the rating for item j from the active user u_a . Based on the prediction results, t items with the highest ratings will be recommended to the active user. Figure 5 shows an 8×8 item-to-item similarity matrix \mathbf{S} with $k = 3$ and the calculation process of IBCF. Based on item similarity coefficients, for each item, we selected the 3 most similar items. Then, only those items had already been rated by the active user u_a were used to calculate the weighted sum of the ratings of i_2 , i_5 and i_8 . Based on the results, i_8 with the highest estimated rating would be recommended to the active user u_a . While reducing the model size could potentially impact recommendation quality, it significantly reduced the calculation complexity and saved time (Sarwar et al., 2001). Also, previous studies have already shown that IBCF is more efficient as well as provide more accurate results than UBCF (Deshpande & Karypis 2004; Papagelis & Plexousakis, 2005).

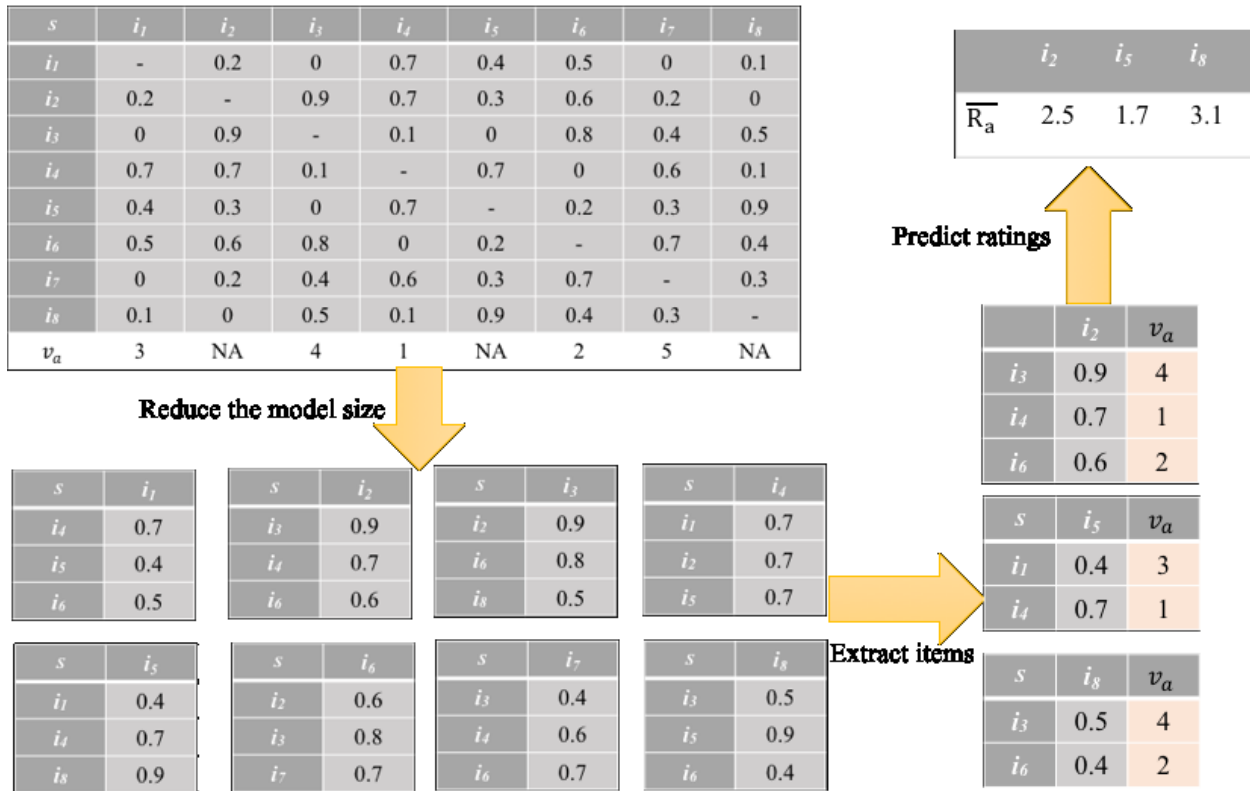


Figure 5. An example of item-to-item similarity matrix S and the calculation process of IBCF.

A large number of studies have shown the excellent performance of CF methods in recommendation systems to predict ratings or recommend products, but their applications to educational assessments are rarely discussed. Previous research utilized CF methods to predict students' abilities to answer questions correctly (Toscher & Jahrer, 2010), which achieved the same goal as IRT models to predict the probability of an examinee answering an item correctly. Furthermore, another study formalized the relationship between IRT and CF methods by using CF methods to estimate “difficulty-like” parameters and “discrimination-like” parameters (Bergner et al., 2012). Results indicated the feasibility of applying CF methods to IRT models. In addition, the idea behind CF methods was to select suitable items for users, which was similar to other item selection methods in adaptive tests. Furthermore, CF methods were flexible to

combine with other algorithms. In this case, it was possible to blend CF methods with Maximum Fisher Information or a -stratification methods.

While no study investigated the applications of CF methods to the item selection process in adaptive tests, especially in OMST, we expected that adopting CF methods to select items would be feasible. Also, it was worth comparing the performance of traditional item selection methods with CF methods and their combined methods in estimating ability parameters. The specific manipulation would be described in the next chapter.

Chapter 3: Method

Overview of Design

This study follows a Monte Carlo simulation approach with 20 OMST designs across several conditions and aims to examine the effects of different item selection methods on ability estimates. The designs can be characterized by the number of stages (two stages or three stages), the routing stage development (maximum information or difficulty parameters close to estimated ability values), and item selection methods (Maximum Fisher Information, α -stratification, UBCF or IBCF).

The simulation study was implemented using the *xxIRT* (Luo, 2016), *mstR* (Magis, Yan, & Von Davier, 2017), *mirt* (Chalmers, 2012), *recommenderlab* (Hahsler, 2015) packages in R (R Core Team, 2018). Across all designs, the test length was fixed to 60 items extracted from a simulated item bank with 300 items and 1000 examinees. Expected a-posteriori (EAP) was applied for provisional and final estimations of ability parameters.

The simulation process involved several steps. First, a complete response matrix \mathbf{A} (1000 x 300) was simulated based on the item bank. Second, the modules in the routing stage were assembled based on ATA algorithms and each examinee's responses to these items were selected from the response matrix \mathbf{A} . Third, provisional ability parameters were estimated, and different item selection methods were adopted to select items for the next stage. In the second stage, each examinee was assigned to a unique module and their responses were again retrieved from the response matrix \mathbf{A} . In 2-stage OMST design, when the test was finished, final ability parameters were estimated. But in 3-stage OMST design, the procedure as described above was continued until the third stage was implemented and then final ability parameters were estimated. 100 replications were conducted across all designs. The performance of different OMST designs was

evaluated in terms of the accuracy of ability estimates, the maximum item exposure rate (i.e., how many times each item has been repeatedly administered to examinees), and item bank utilization (i.e., what percentage of the items in the item bank has been actually used in the test administration).

Data Generation

The item bank was simulated based on the item parameters of a large-scale adaptive reading assessment in the United States. In total, 300 dichotomous items were selected from the original item bank. All of the selected items were calibrated under the three-parameter logistic (3PL) model. Table 1 shows descriptive statistics for each item parameter and Figure 6 shows the frequency distribution of discrimination parameters.

Table 1

Distributions of Item Parameters

Item Parameter	Mean	Standard Deviation	Minimum	Maximum
a	2.13	0.49	1.50	3.73
b	-0.34	0.86	-2.71	1.61
c	0.23	0.03	0.06	0.29

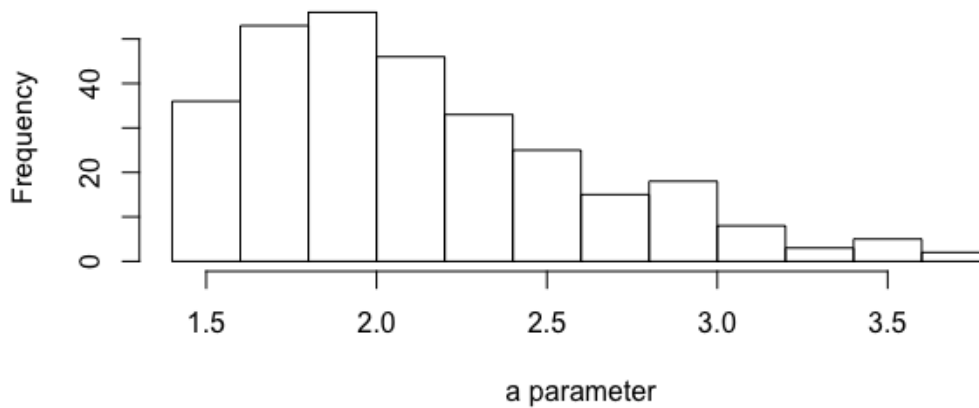


Figure 6. The frequency distribution histogram of discrimination parameters.

In order to generate the response matrix, true ability parameters for 1000 examinees were drawn from a normal distribution, $\theta \sim N(0, 1)$. Because the total test length was fixed to 60 items for all designs, each stage had 30 items for 2-stage OMST designs and 20 items for 3-stage OMST designs.

Automated Test Assembly (ATA) Procedure

In the routing stage, three parallel modules were assembled by using ATA algorithms to control the item exposure rates. In general, modules in the routing stage often keep moderate difficulty levels (Luecht, Brumfield, & Breithaupt, 2006). Each examinee received one module randomly. This study implemented two ATA algorithms. One maximized the test information function over the ability parameters from -0.8 to 0.8. Another firstly partitioned the item bank into two (2-stage design) or three (3-stage design) levels from low to high based on discrimination parameters. Then, three parallel modules selected items from the first level which had relatively lower discrimination parameters. At the same time, these items' difficulty parameters satisfied their mean close to 0 and the standard deviation close to 0.5. Considering

the item bank is finite, ATA algorithms were set to allow items to be used *at most* twice between the modules. The test information plots for the parallel modules based on these two algorithms were shown in Figures 7 and 8, respectively. Because CF methods require that examinees already have responses for some items, both UBCF and IBCF item selection methods can only be conducted in the subsequent stages.

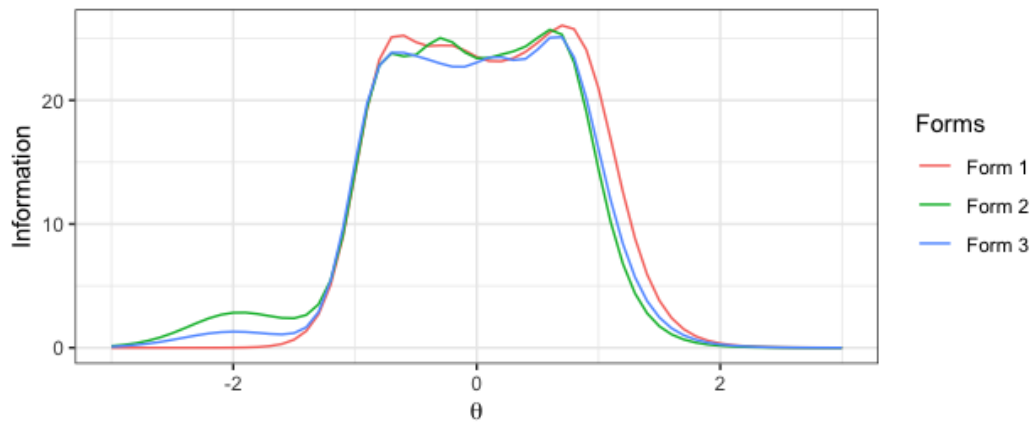


Figure 7. The test information plots for parallel modules under the maximum test information design.

In summary, when the routing stage used the maximum information function to create modules, Maximum Fisher Information, UBCF or IBCF method would be adopted for next stages and evaluation measures would be compared among these designs. However, if another ATA algorithm (b parameter close to 0) was applied in the routing stage, a -stratification, UBCF or IBCF method would be conducted for next stages and evaluation measures would be compared among these designs. Figure 9 illustrates the process of the current study.

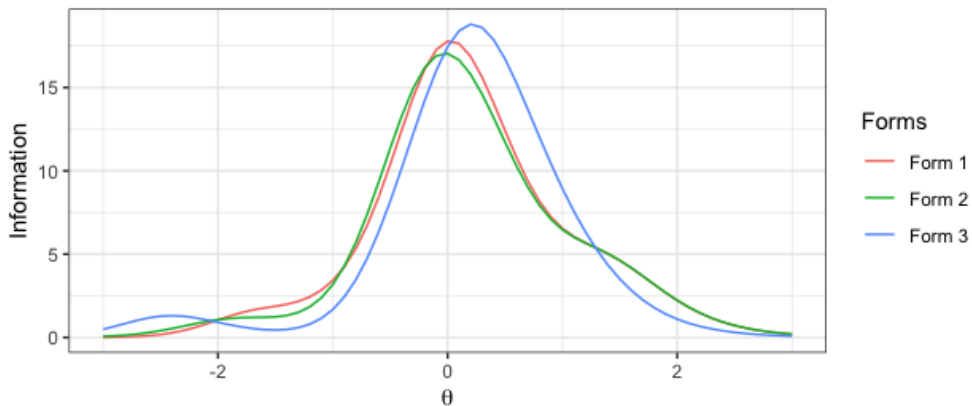


Figure 8. The test information plots for parallel modules under the b-parameter close to 0 design.

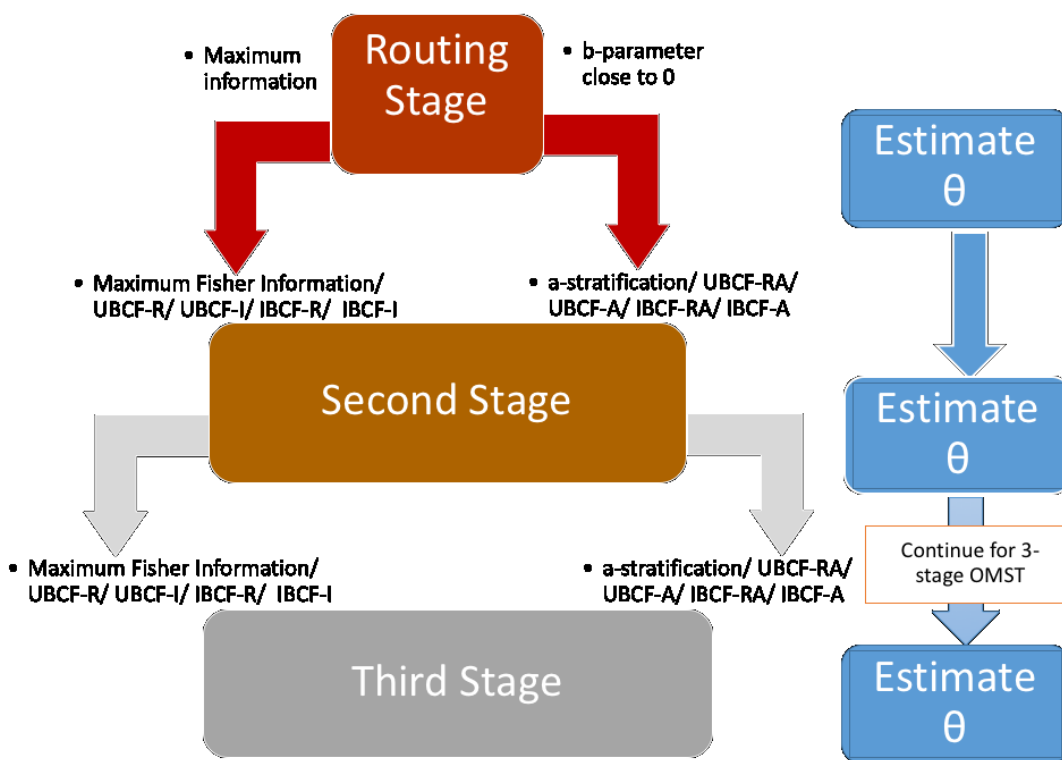


Figure 9. A brief description of the entire OMST process.

Item Selection Methods

As described earlier, Maximum Fisher Information is one of the most common item selection methods and several packages in R (i.e., *mirt*) already have functions to implement it for item selection. Under OMST design, once provisional ability parameter of each examinee was estimated, 20 (3-stage design) or 30 (2-stage design) items with the maximum information for estimated ability parameter would be selected to construct modules in the subsequent stages.

As for α -stratification method, adaptive modules in the second (and third) stage(s) consisted of items whose b parameters were closest to the latest provisional ability parameters (Zheng & Chang, 2015). Under 3-stage OMST design, items in the second stage can only be selected from the second level of the item bank and the third stage's items can only be chosen from the third level. 2-stage OMST design only partitioned the item bank into two levels, and therefore, the second stage's items were from the second level of the item bank as well as met the requirement.

CF methods recommend items based on other users' preferences, which suggests that we need to collect as much data as possible to improve the accuracy of recommendations being made (Sarwar et al., 2001). However, the response matrix of the routing stage could only provide 1,000 examinees' tastes for no more than 30 items (some items might be used twice), and thus there was no information about their preferences (i.e., responses) to other items in the item bank. Consequently, it is impossible to provide item recommendations. To address this problem, this study generated another complete response matrix \mathbf{B} (2000 x 300) based on the item bank and 2,000 examinees' ability parameters also followed a normal distribution, $\theta \sim N(0, 1)$. In the second stage, both UBCF and IBCF methods recommended items to examinees from the

response matrix **A** based on their similarities with others or items' similarities knowing from the response matrix **B**.

As mentioned previously, items with the highest predicted ratings in the top- N list are recommended in general. It should be noted that previous studies rated items according to the 5-point or 7-point scales, but items' ratings in the current study were dichotomous (i.e., 1 = correct; 0 = incorrect). The narrow range of ratings might influence the performance of the top- N list considering that large numbers of items probably had the same predicted ratings. In order to address this issue, the current study improved UBCF and IBCF in three different ways. In the second stage, this study still adopted the top- N list, but the list had 40 or 60 items. The first way was to select 20 or 30 items *randomly* from the list and compose adaptive modules. This method was denoted as UBCF-R or IBCF-R. If the item bank was partitioned previously, UBCF and IBCF would only create a top- N list from the second level of the item bank and select 20 or 30 items randomly. In this case, this method was denoted as UBCF-RA or IBCF-RA since it was adapted from a -stratification method.

The second way was to calculate each item's information based on provisional ability parameters and to choose the top 20 or 30 items with the highest item information. This method was denoted as UBCF-I or IBCF-I. The third way learned from the a -stratification method and selected either 20 or 30 items from the second level of the item bank. These items' b parameters were closest to the provisional ability parameters (Chang & Ying, 1999). This method was denoted as UBCF-A or IBCF-A. Similar ways to select items also applied in the third stage and the top- N list would have 60 items to make sure enough items could be recommended for the third stage.

Data Analysis

This study used expected a posteriori (EAP) to estimate ability parameters considering several studies adopted this technique (e.g., Sakumura, & Hirose, 2017; Sinharay, 2016). To examine the performance of each item selection method, the measurement precision of final ability parameters under each design was calculated. Indices included bias, root mean square error (RMSE), and the Pearson correlation between true ability parameters and estimated ability parameters. Bias, RMSE and correlation values were calculated as follows:

$$Bias = \frac{\sum_{i=1}^N (\bar{\theta}_i - \theta_i)}{N}, \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\bar{\theta}_i - \theta_i)^2}{N}}, \quad (7)$$

$$\rho_{\bar{\theta}_i, \theta_i} = \frac{cov(\bar{\theta}_i, \theta_i)}{\sigma_{\bar{\theta}_i} \sigma_{\theta_i}}, \quad (8)$$

where $\bar{\theta}_i$ represents examinee i 's final ability parameter based on the OMST responses, θ_i represents examinee i 's true ability parameter, N represents the sample size, $cov(\bar{\theta}_i, \theta_i)$ represents the covariance between estimated and true ability parameters, $\sigma_{\bar{\theta}_i}$ represents the standard deviation of $\bar{\theta}_i$, σ_{θ_i} represents the standard deviation of θ_i . The smaller values of bias and RMSE and higher correlation values indicated the parameter estimates were more accurate. The positive values of bias indicated the ability parameters were over-estimated and negative values represented under-estimated.

Except for the measurement precision, the item bank utilization was also investigated by using the maximum item exposure rate and the proportion of unused items. Each item's exposure rate was calculated based on the proportion of the number of examinees who answered the item to the total sample size, and the maximum value (i.e., how many times the item was used) across all items was the maximum item exposure 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.

Another index used in this study was the proportion of examinees who were left out at the end of the test. When we conducted the CF methods, we found that some examinees were not recommended any items sometimes. This situation occurs when examinees rate items without any preference. In other words, examinees do not answer any items correctly. Then, it is impossible for CF methods to create a top- N list. Also, it may occur when there are not enough items after reducing the model size in UBCF. Considering this situation, this study deleted those examinees who did not get any recommendations. In the end, the sample size might be smaller than 1,000. The proportion of examinees who completed the test would be recorded. Finally, the average values for each index over 100 replications were reported.

Chapter Summary

This study conducted 2-stage and 3-stage OMST designs separately. The test length was fixed as 60 items. 2-stage OMST designs had 30 items for each stage and 3-stage OMST designs had 20 items. The routing stages adopted test information method or b -parameter method by using ATA to assemble parallel modules. If the first method was used, an item selection method (i.e., Maximum Fisher Information, UBCF-R, UBCF-I, IBCF-R, or IBCF-I) was chosen to identify the items for the subsequent stages. On the contrary, if the second approach was

conducted, α -stratification, UBCF-RA, UBCF-A, IBCF-RA, or IBCF-A was used to identify the items. In total, this study conducted 20 OMST designs. Results would be presented in the next chapter.

Chapter 4: Results

In this chapter, the results of different OMST designs were presented and compared.

There were two test structures, 2-stage OMST and 3-stage OMST. Under each condition, 10 item selection methods were adopted.

Measurement Accuracy Results

Mean values for bias, RMSE and correlation are shown in Table 2. There were no large differences between the results of 2-stage and 3-stage OMST designs. The mean bias values for all conditions were similar and nearly zero. Using Maximum Fisher Information, UBCF-I, UBCF-R, and UBCF-A item selection methods slightly underestimated true ability parameters in 2-stage OMST designs (Bias = -0.01). In 3-stage OMST designs, Maximum Fisher Information, UBCF-I, UBCF-R, and α -stratification methods also yielded slightly underestimated values of ability parameters (Bias = -0.01).

The mean RMSE values between estimated and true ability parameters across all conditions ranged from 0.18 to 0.26. RMSE values indicated that Maximum Fisher Information method provided the most accurate estimates of ability parameters, followed by α -stratification method in both 2-stage and 3-stage conditions (see Figure 10). Compared with other methods, UBCF-RA resulted in the most inaccurate estimates of ability parameters across all conditions (RMSE = 0.26). As the number of stages increased, RMSE values remained the same or decreased slightly for all item selection methods apart from UBCF-R. Its RMSE value increased from 0.24 to 0.25.

Overall, mean correlation results showed the estimated ability parameters were very close to true ability values under each condition, ranging from 0.96 to 0.99. In 2-stage OMST test designs, Maximum Fisher Information, UBCF-I or α -stratification method produced the highest

correlation values between estimated ability parameters and true ability parameters ($\rho_{\bar{\theta}_i, \theta_i} = 0.98$), but IBCF-RA had the lowest values ($\rho_{\bar{\theta}_i, \theta_i} = 0.96$). As for 3-stage OMST test designs, Maximum Fisher Information method still had the highest values ($\rho_{\bar{\theta}_i, \theta_i} = 0.99$) and IBCF-R had the lowest values instead ($\rho_{\bar{\theta}_i, \theta_i} = 0.96$). The number of stages seemed to have no influence on correlation values for a -stratification, UBCF-RA, UBCF-A, UBCF-R, UBCF-I, IBCF-I, however, correlation values for IBCF-RA, IBCF-A and Maximum Fisher Information increased when the number of stages increased from 2 to 3. IBCF-R was an exception and its correlation decreased from 0.97 to 0.96 as the number of stages increased. Figure 11 graphically presents the mean correlation values for each condition.

Table 2

Mean Values of Measurement Accuracy Indices for Different Item Selection Methods

Method	2-stage			3-stage		
	Bias	RMSE	Correlation	Bias	RMSE	Correlation
Maximum Fisher Information	-0.01	0.19	0.98	-0.01	0.18	0.99
IBCF-I	0.00	0.22	0.97	0.00	0.22	0.97
IBCF-R	0.00	0.24	0.97	0.00	0.24	0.96
UBCF-I	-0.01	0.23	0.98	-0.01	0.23	0.98
UBCF-R	-0.01	0.24	0.97	-0.01	0.25	0.97
a -stratification	0.00	0.21	0.98	-0.01	0.20	0.98
IBCF-A	0.00	0.23	0.97	0.00	0.21	0.98
IBCF-RA	0.00	0.26	0.96	0.00	0.25	0.97
UBCF-A	-0.01	0.24	0.97	0.00	0.23	0.97
UBCF-RA	0.00	0.26	0.97	0.00	0.26	0.97

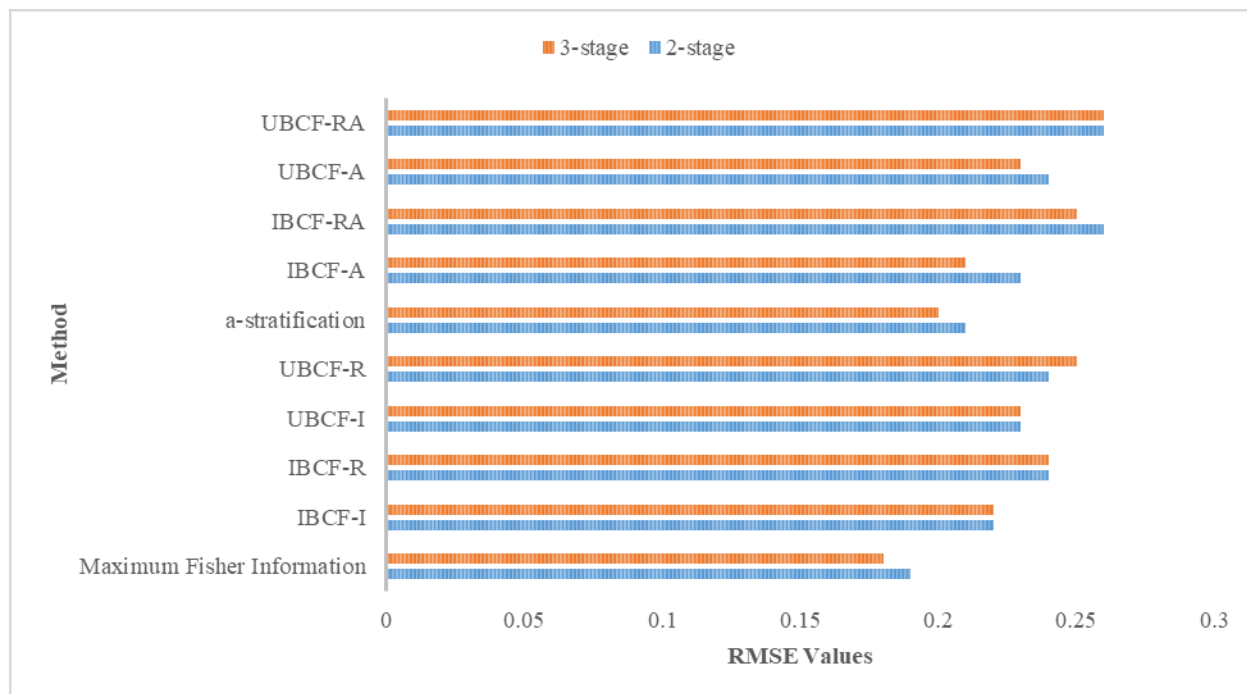


Figure 10. Mean RMSE values by different item selection method and test structure.

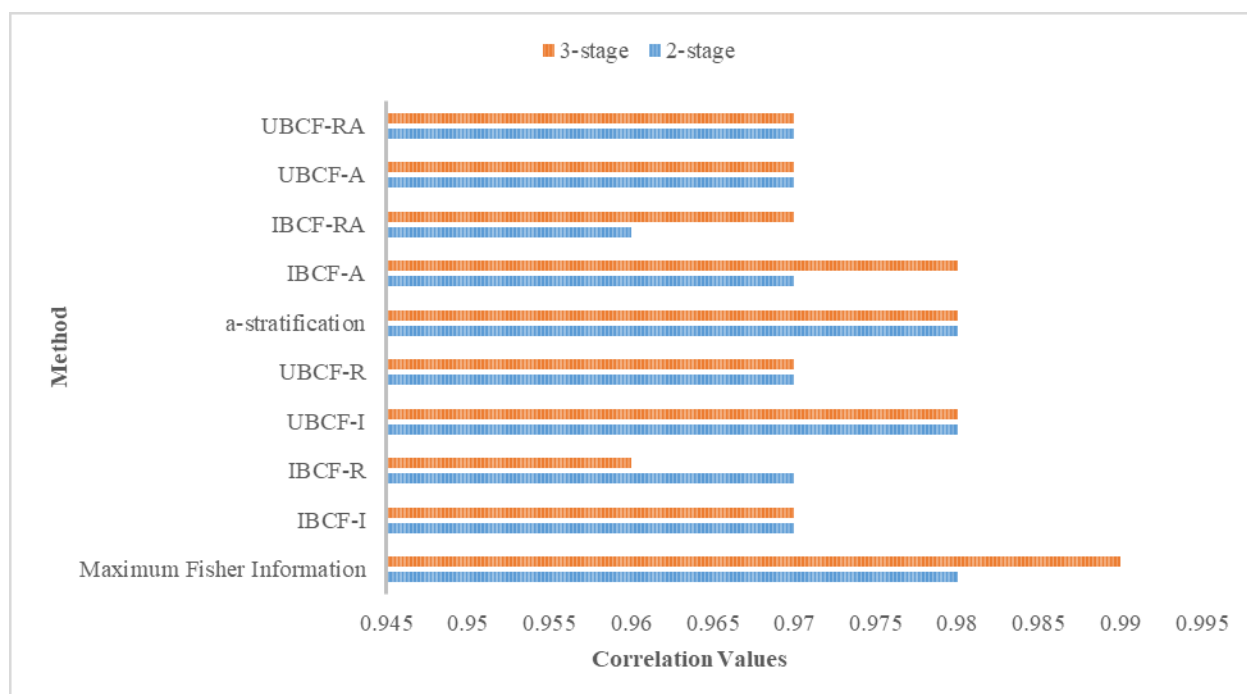


Figure 11. Mean correlation values by different item selection method and test structure.

Item Bank Utilization Results

The item bank utilization was evaluated based on the maximum exposure rates and the proportion of unused items. Table 3 presents the outcomes for each method. In 2-stage OMST designs, *a*-stratification, IBCF-RA and UBCF-RA performed the best to control item exposure and the maximum exposure rate was controlled as 0.68. The latter two methods kept the best performance to item exposure control (0.67) in 3-stage OMST designs. The maximum item exposure rates could reach up to 0.82 when using Maximum Fisher Information, UBCF-I and UBCF-R methods in the 2-stage structure. However, the maximum item exposure rate of 0.85 was achieved by IBCF-A method in 3-stage OMST designs. In general, increasing the number of stages produced higher item exposure rates, but UBCF-R, *a*-stratification, IBCF-RA and UBCF-RA could control the item exposure regardless of the increase of the stage number.

The proportion of unused items results showed that IBCF-R method performed the best in both 2-stage and 3-stage structures, followed by IBCF-I. Moreover, IBCF-R used the whole items from the item bank in 3-stage OMST designs and 97% of items in 2-stage OMST designs. When adopting Maximum Fisher Information method to selection items, nearly 39% of items were unused for 2-stage tests and 30% for 3-stage tests. When the number of stages increased, more items would be used for the test and the proportion of unused items would decrease. It was consistent across all conditions.

Table 3

Mean Values of Item Bank Utilization Indices for Different Item Selection Methods

Method	2-stage		3-stage	
	Maximum Exposure Rates	Proportion of Unused Items	Maximum Exposure Rates	Proportion of Unused Items
Maximum Information	0.82	0.39	0.83	0.30
IBCF-I	0.79	0.09	0.84	0.03
IBCF-R	0.74	0.03	0.80	0.00
UBCF-I	0.82	0.12	0.83	0.10
UBCF-R	0.82	0.13	0.77	0.11
<i>a</i> -stratification	0.68	0.34	0.68	0.22
IBCF-A	0.79	0.38	0.85	0.22
IBCF-RA	0.68	0.34	0.67	0.22
UBCF-A	0.71	0.40	0.77	0.25
UBCF-RA	0.68	0.38	0.67	0.24

The Proportion of Remaining Examinees Results

This study also compared the proportion of remaining examinees and results are presented in Table 4. As mentioned earlier, CF methods might not recommend any items to some examinees, and thus those examinees were deleted in the current study. This was not a concern for traditional item selection methods, UBCF-I and UBCF-R methods. Other CF methods would result in case deletion. In the 2-stage structure, 94% examinees remained for IBCF-IF and IBCF-R. For 3-stage OMST tests, using IBCF-RA resulted in only 90% examinees left, which meant approximate 100 out of 1,000 examinees would not get any recommendations for items.

Table 4

Mean Values of the Proportion of Remaining Examinees for Different Item Selection Methods

Method	2-stage	3-stage
Maximum Information	1.00	1.00
IBCF-I	0.94	0.93
IBCF-R	0.94	0.93
UBCF-I	1.00	1.00
UBCF-R	1.00	1.00
<i>a</i> -stratification	1.00	1.00
IBCF-A	0.97	0.97
IBCF-RA	0.97	0.90
UBCF-A	1.00	1.00
UBCF-RA	1.00	0.91

Chapter Summary

The results of the current study demonstrated that traditional item selection methods outperformed CF methods for ability estimates. However, CF methods showed their superiority for the item bank utilization. But their drawback of case deletion could not be ignored. Overall, the difference between the 2-stage and 3-stage OMST designs with regard to measurement accuracy indices was negligible, but the number of stages could influence item bank utilization results.

Chapter 5: Discussion

Since on-the-fly assembled multistage adaptive testing (OMST) was proposed, it has drawn great attention from researchers and practitioners who design and implement adaptive tests (Du, Li, & Chang, 2017). The item selection procedure, which plays a critical role in the adaptive testing design, deserves to be investigated in more depth. Previous research on popular item selection methods – such as Maximum Fisher Information and α -stratification methods – already investigated the effects of using such methods on the accuracy of ability estimation and item exposure control in traditional adaptive tests (Chang & Ying, 1996), but their applications in the OMST design still need to be explored. Except for the aforementioned item selection methods, this study also proposed new item selection methods. Due to Collaborative Filtering (CF) algorithms' excellent estimation performance in recommendation systems (Li, Karatzoglou, & Gentile, 2016), the current study employed them in the item selection process under OMST framework.

The first research purpose of this study was to develop UBCF and IBCF item selection methods. To achieve this goal, the current study created a top- N list for each examinee based on item similarity (i.e., item information, difficulty, and discrimination) and examinees' preferences (i.e., correct/incorrect answers). Then, items from the top- N list would be selected using CF algorithms. More specifically, UBCF-RA and IBCF-RA only selected items from the second or third levels of the item bank. UBCF-I and IBCF-I selected items with the maximum item information based on provisional ability parameters. As for UBCF-A and IBCF-A, items with b parameters closest to the provisional ability parameters as well as from the specific levels of the item bank were chosen.

The second research purpose was to compare the performance of CF-based item selection methods with traditional item selection methods (Maximum Fisher Information and α -stratification methods) under 2-stage and 3-stage OMST designs. Both 2-stage and 3-stage OMST designs had 60 items. 2-stage OMST design had 30 items for each stage and 3-stage OMST design had 20 items for each stage. To evaluate the performance of each item selection method, bias, RMSE and correlation values were compared. Also, the item bank utilization was evaluated by calculating the maximum item exposure rates and the proportion of unused items. 100 replications were conducted to get the average value for each index. This study was designed to address the following specific research questions: (1) Do CF algorithms produce more accurate ability estimates compared to traditional item selection methods under 2-stage or 3-stage OMST designs?; (2) Do CF algorithms control the item exposure better compared to traditional item selection methods under 2-stage or 3-stage OMST designs?; and (3) Do CF algorithms utilize more items in the item bank compared to traditional item selection methods under 2-stage or 3-stage OMST designs?

Results for Research Question 1

The present study calculated mean bias, RMSE, and correlation values to measure the accuracy of ability estimates. The results indicated that different item selection methods produced very similar ability estimates (see Table 2), with the mean bias close to 0 to two decimal places. RMSE values suggested that using Maximum Fisher Information produced the most accurate ability estimates (0.18) whereas UBCF-RA yielded the largest RMSE value (0.26). Different item selection methods also led to similar correlation values between true and estimated ability parameters, a maximum difference of 0.03 between the lowest (0.96) and highest (0.99) correlation values. Besides, results indicated increasing the number of stages from

two to three had a very negligible effect on the accuracy of ability estimation, which was inconsistent with previous research under the MST framework (Patsula, 1999). A possible explanation for the difference is that the current study was under OMST framework. Modules are adaptive in OMST designs, which means examinees receive modules that have already been tailored to closely match their ability levels. On the contrary, modules in traditional MST are typically pre-assembled and thus some examinees receive modules that may not match their true ability levels. This situation is common especially at the beginning of the test. In this case, increasing the number of stages can collect more information as well as decrease the degree of error in ability estimates in traditional MST.

Results for Research Question 2

The present study compared maximum item exposure rates across the item selection method. Results indicated IBCF-RA and UBCF-RA performed the best to control the overexposure of popular items (0.67), followed by the α -stratification method (0.68) under 3-stage OMST designs. These three methods were identical with regard to controlling item exposure (0.68) under 2-stage OMST designs (see Table 3). It is not surprising that IBCF-RA and UBCF-RA had the highest RMSE values (0.26) for ability estimates (see Table 2). There is a trade-off between RMSE values and the maximum item exposure rate because test security will be sacrificed by exposing the same items to most examinees while achieving more accurate ability estimates (Zheng & Chang, 2015). Results also suggested that increasing the number of stages from two to three had no systematic effect on the maximum item exposure rates considering that some methods controlled the item exposure (e.g., UBCF-RA, UBCF-R) but others increased item exposure rates (e.g., IBCF-A, UBCF-A).

Results for Research Question 3

The proportion of unused items was calculated as another evaluation criterion in this study. Results indicated that IBCF-I and IBCF-R were the most effective methods to utilize items as many as possible (see Table 3). While these two methods resulted in cased deletions (see Table 4), they recommended highly similar items as well as utilized the item bank completely. As the number of stages increased, the proportion of unused items would decrease. One possible explanation is that increasing the number of stages actually enhances opportunities for items to be selected.

Conclusion

The current study demonstrated the feasibility of using CF algorithms as item selection methods under OMST framework. CF methods could produce accurate ability estimates that were comparable to traditional item selection methods (Maximum Fisher Information and α -stratification) even though CF methods yielded relatively higher RMSE values for final ability estimates. Furthermore, CF methods indicated more superior performance in terms of item bank utilization. For example, even α -stratification method was outperformed by IBCF-RA and UBCF-RA for item exposure control under 3-stage designs. Therefore, it can be concluded that CF methods performed adequately for item selection in OMST designs.

Limitations of the Study and the Directions for Future Research

This study has several limitations. First, the present study compared CF methods with two popular item selection methods considering measurement accuracy and item bank utilization; but nonstatistical constraints (e.g., answer key balancing, content balancing) were not taken into consideration. It is very important for standardized tests to have similar content and reliable ability estimates for examinees (van der Linden, 2005). Previous studies also developed several item selection methods to control content balancing, such as the maximum priority index

method (MPI) (Cheng & Chang, 2009), weighted-deviations method (WDM) (Stocking & Swanson, 1993), shadow test approach (STA) (van der Linden & Reese, 1998). Future studies can set content constraints to compare the performance of CF methods with other item selection methods.

Second, the current study only explored the effects of the number of stages. Previous studies also found that other design features such as test length and the number of items within each module could impact results under the MST framework (Patsula, 1999; Chen, 2011), but their effects in the OMST design were unknown. It will be necessary for future research to investigate whether these design features will influence the accuracy ability estimates or item bank utilization.

Finally, as mentioned earlier, previous studies that utilized CF algorithms often chose rating scales with a wide range (e.g., Koren, 2009; Linden, Smith, & York, 2003), but the present study adopted dichotomous ratings (i.e., responses) under the 3PL IRT model. In the future, researchers could construct OMST designs with polytomous items to examine whether CF methods could produce more accurate ability estimates and enhance test security. Alternatively, response options can be considered nominal categories to select items based on similar or relevant distractors from multiple-choice items.

References

- Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27*(3), 241–253
- Angoff, W. H., & Huddleston, E. M. (1958). *The multi-level experiment: a study of a two-level test system for the College Board Scholastic Aptitude Test* (Report SR-58-21). Princeton, NJ: Educational Testing Service.
- Armstrong, R. D., Jones, D. H., Koppel, N. B., & Pashley, P. J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement, 28*(3), 147–164.
- Armstrong, R., & Roussos, L. A. (2005). *A method to determine targets for multi-stage adaptive tests* (Vol. 2, No. 7). Law School Admission Council.
- 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. *International Educational Data Mining Society*.
- Betz, N. E., & Weiss, D. J. (1973). *An empirical study of computer-administered two-stage ability testing*. (Research Report 73–4). Psychometric Methods Program, Department of Psychology, University of Minnesota.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- Bock, R. D., & Zimowski, M. F. (1998). *Feasibility studies of two-stage testing in large-scale educational assessment: Implications for NAEP*. Commissioned by the NAEP Validity

- Studies Panel, American Institute for Research in the Behavioral Sciences. Washington, D.C.: NCES.
- Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). *Empirical analysis of predictive algorithms for collaborative filtering*. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence (pp. 43-52). Morgan Kaufmann Publishers Inc..
- Breithaupt., K.; Ariel, A. A.; & Hare, D. R. (2010). Assembling an inventory of multistage adaptive testing systems. In W. J. van der Linden & C. E. W. Glas (Eds). *Elements of Adaptive Testing*, (pp. 247 - 266). New York: Springer.
- Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education*, 47(8), A47.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.
- Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1), 1-20.
- Chang, H-H., Qian, J., & Ying, Z. (2001). A-stratified multistage computer adaptive testing with b blocking. *Applied Psychological Measurement*, 25(4), 333–341.
- Chang, H. H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222.
- Chang, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73(3), 441.
- Chen, L-Y. (2011). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model*. The University of Texas at Austin.

- 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.
- Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. (2nd Ed.) Urbana, IL: University of Illinois Press.
- Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1), 143-177.
- Du, Y., Li, A., & Chang, H. H. (2017, July). Utilizing response time in on-the-fly multistage adaptive testing. In *The Annual Meeting of the Psychometric Society* (pp. 107-117). Springer, Cham.
- Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, 5(8).
- Han, K. C. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of educational evaluation for health professions*, 15.
- Hahsler, M. (2015). recommenderlab: A framework for developing and testing recommendation algorithms, from <http://CRAN.R-project.org/package=recommenderlab>.
- Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, 26, 44–52.
- Kim, H., & Plake, B. S. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Atlanta, GA.

- Koren, Y. (2009, June). Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 447-456). ACM.
- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.
- Li, S., Karatzoglou, A., & Gentile, C. (2016, July). Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 539-548). ACM.
- Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing, (1)*, 76-80.
- Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36*(3), 227–242.
- Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems, 74*, 12-32.
- Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Luecht, R. M. (2003, April). *Exposure control using adaptive multistage item bundles*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for the uniform CPA examination. *Applied Measurement in Education, 19*, 189 - 202.

- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35(3), 229–249.
- Luo, X. (2016). xxIRT: Practical item response theory and computer-based testing in R [Computer software]. Retrieved February 27, 2017, from <https://cran.r-project.org/package=xxIRT>
- Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.
- Melican, G. J.; Breithaupt, K.; & Zhang, Y. (2010). Designing and implementing a multistage adaptive test: the Uniform CPA Exam. In W. J. van der Linden & C. E. W. Glas (Eds). *Elements of Adaptive Testing* (pp. 167 - 189). New York: Springer.
- 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.
- Papagelis, M., & Plexousakis, D. (2005). Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents. *Engineering Applications of Artificial Intelligence*, 18(7), 781-789.
- Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.
- Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1), 89-106.
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994, October). GroupLens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work* (pp. 175-186). ACM.
- Sakumura, T., & Hirose, H. (2017). A bias reduction method for ability estimates in adaptive online IRT testing systems. *International Journal of Smart Computing and Artificial Intelligence*, 1(1), 59-72.
- Sari, H. I., Yahsi-Sari, H., & Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 7(2), 388-406.
- 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*, pp. 285-295.
- Schnipke, D. L., & Reese, L. M. (1997). *Comparison of testlet-based test designs for computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
- Sinharay, S. (2016). The choice of the ability estimate with asymptotically correct standardized person - fit statistics. *British Journal of Mathematical and Statistical Psychology*, 69(2), 175-193.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277-292.
- Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* (Doctoral dissertation, University of Illinois at Urbana-Champaign).

- Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. winner - 3rd place. *KDD Cup 2010: Improving Cognitive Models with Educational Data Mining*.
- 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. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*(3), 283-302.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.
- Wang, J., De Vries, A. P., & Reinders, M. J. (2006, August). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 501-508). ACM.
- 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.
- Zheng, Y., & Chang, H. H. (2014). Multistage testing, on-the-fly multistage testing, and beyond. *Advancing methodologies to support both summative and formative assessments*, 21-39.
- Zheng, Y., & Chang, H. H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement, 39*(2), 104-118.