Making Adaptive Testing More User-Aware and Selective: A Recommendation System-Based Item Selection Strategy

Authors: Wang Pujue, Liu Hongyun

Date: 2019-02-19T00:00:00+00:00

Abstract

Drawing on the concept of collaborative filtering from recommendation systems, we propose two CAT item selection strategies that can leverage existing examinee data: Direct Examinee-Based Recommendation (DEBR) and Indirect Examinee-Based Recommendation (IEBR). Through two simulation studies, across different item banks and test lengths, we compared the performance of these two recommendation-based strategies with two traditional item selection strategies (FMI and BAS) in terms of measurement accuracy and item exposure rate control, as well as the factors influencing the performance of the recommendation-based strategies. The results revealed that: the two recommendation-based strategies exhibited superior control over item exposure rates compared to the traditional strategies, while maintaining measurement accuracy comparable to the BAS method; specifically, DEBR emphasized item selection precision, whereas IEBR achieved the best control of item exposure rates. The characteristics and quality of the existing examinee data constitute the primary factors affecting the performance of the recommendation-based item selection strategies.

Full Text

Making Adaptive Testing Better at Knowing Examinees and Selecting Items: Item Selection Strategies Based on Recommender Systems

WANG Pujue[1]; LIU Hongyun[1,2]

([1] School of Psychology, Beijing Normal University, Beijing 100875, China)
([2] Beijing Key Laboratory of Applied Experimental Psychology, School of Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

Drawing on the principles of collaborative filtering recommendation from recommender systems, this study proposes two novel Computerized Adaptive Testing (CAT) item selection strategies that leverage existing examinee data: Direct Examinee-Based Recommender (DEBR) and Indirect Examinee-Based Recommender (IEBR). Through two simulation studies, we compared the measurement precision and item exposure control of these recommender-based strategies against two traditional methods (FMI and BAS) across different item banks and test lengths, while also examining factors influencing the performance of the recommender strategies. Results indicate that both recommender strategies outperform traditional methods in controlling item exposure rates, with measurement precision comparable to or exceeding that of BAS. DEBR prioritizes selection accuracy, while IEBR demonstrates the strongest control over item exposure rates. The characteristics and quality of existing examinee data emerge as the primary factors affecting recommender strategy performance.

Keywords: item selection strategy; past examinees' data; recommender system; collaborative filtering recommender; simulation study
Classification: B841

Computerized Adaptive Testing (CAT) employs specific item selection strategies to provide examinees with items tailored to their ability levels, thereby measuring true abilities more accurately and efficiently through a "customized" test (Weiss, 1982). As intelligent learning and testing systems have gained popularity, CAT applications have expanded considerably (Zhang & Chang, 2016), generating substantial process data from completed tests. From a data mining perspective, this data contains rich information, including response patterns, evolving ability estimates, and mappings to subsequent items. Appropriate techniques can extract useful patterns to predict unknown outcomes (Tan, Steinbach, & Kumar, 2007), effectively abstracting and establishing a new set of item selection rules that can reproduce selection accuracy similar to the original strategy while dynamically adjusting these rules to address identified problems —such as uneven item bank usage. Mao and Xin (2011) noted that a crucial direction for improving CAT item selection strategies involves making full use of prior information about examinees. For each test-taker, data from previous examinees represents a valuable yet long-neglected source of prior information. Since CAT development has primarily been grounded in Item Response Theory (IRT; Chang, 2015), most existing IRT-based selection strategies utilize only the current examinee' s response information, making it difficult to incorporate past examinee data and limiting the ability to learn selection experience from others to improve strategies.

Building more intelligent learning and testing systems to further achieve adaptive goals is an interdisciplinary challenge requiring integration of expertise from psychology, education, statistics, and machine learning (Chen, Li, Liu, & Ying, 2018; Zhang & Chang, 2016). Regarding the use of past examinee data to im-

prove CAT item selection, and given the limitations of traditional strategies, we can introduce novel technical approaches on the IRT foundation—recommender systems present a suitable choice.

Recommender Systems comprise algorithms and techniques that use existing data to recommend items to users, providing precise matches based on user needs. This represents a popular research topic in data mining (Ricci, Rokach, & Shapira, 2015), with mature algorithms achieving tremendous success in commercial, entertainment, and social applications (Covington, Adams, & Sargin, 2016; Quijano-Sánchez, Recio-García, Díaz-Agudo, & Jiménez-Díaz, 2011; Smith & Linden, 2017). In education, recommender systems can leverage large-scale learning data to predict student performance on new items with greater accuracy than traditional methods (Thai-Nghe, Drumond, Krohn-Grimberghe, & Schmidt-Thieme, 2010). Recent e-Learning developments have used recommender systems to design personalized learning plans for tens of thousands of learners (Liu et al., 2018; Klašnja-Milićević, Ivanović, & Nanopoulos, 2015). Thus, recommender systems offer viable solutions for utilizing past examinee data in item selection.

Recommender systems can also integrate with IRT to build intelligent learning systems with similar adaptive properties. Zhu et al. (2017) integrated the DINA model with matrix factorization techniques into a collaborative filtering item recommendation method that simultaneously estimates knowledge mastery and recommends items, outperforming single cognitive diagnosis models or data mining algorithms. Chen et al. (2018) combined recommender systems, multidimensional IRT models, and reinforcement learning to propose two adaptive learning system prototypes, demonstrating higher efficiency in selecting learning materials compared to random selection across two statistical metrics. They argued that the core component of adaptive learning should be a recommender system that infers latent knowledge states from performance and selects appropriate materials—remarkably similar to CAT' s core process of selecting items matched to examinee ability based on responses. In essence, CAT item selection strategies can be viewed as recommender systems. However, no precedent exists for combining recommender systems with CAT item selection. With appropriate recommendation techniques, we can fill this gap.

Collaborative Filtering Recommender from recommender systems uses extensive user data to predict and recommend items for current users, perfectly aligning with the goal of using past examinee data for item selection. Collaborative filtering assumes that if two users previously showed interest in the same items, they will likely share similar preferences in the future, thereby filtering out the most relevant items for recommendation (Pirasteh, Jung, & Hwang, 2014). Simple to implement without model training, its underlying assumptions have proven stable and effective across numerous scenarios, making it the most mature and popular recommendation method (Koren & Bell, 2015). Applying collaborative filtering to CAT item selection avoids complex calculations and constraint procedures in traditional strategies, quickly screening suitable items for cur-

rent examinees from past data. Moreover, additional rules can be incorporated based on research needs to design flexibly extensible selection strategies that can emphasize either selection accuracy or item exposure control, or balance both while ensuring test security. For example, by calculating similarity among past examinees on administered items, using a recommendation algorithm to filter candidate items for the current examinee, and then applying an exposure control method to select the final item—this satisfies accuracy requirements while ensuring uniform item bank usage.

Based on this analysis, this study aims to apply collaborative filtering recommendation from recommender systems to CAT item selection, proposing novel strategies that leverage past examinee data (hereinafter referred to as recommender-based selection strategies). Through Monte Carlo simulation studies, we examine the performance of these strategies in selection accuracy and item exposure control under various conditions.

2.1 Generating Initial Data Using Traditional Item Selection Strategies

Reliable historical user data is essential for accurate recommendations, corresponding to past examinee data in CAT. If previous examinees received items mismatched to their abilities, accumulating low-precision selection data in the database, we cannot expect recommender strategies to identify correct selection patterns for new examinees. Beyond selection accuracy, CAT strategies must also control item exposure rates. If past strategies inadequately utilized the entire item bank, creating imbalanced exposure in the data, recommender strategies may be affected, selecting items according to these historically uneven proportions.

We must first employ well-established traditional item selection strategies with distinct characteristics to generate initial past examinee data for examining the features of recommender strategies. The first strategy is Lord' s (1980) Maximum Fisher Information (MFI) method, which maximizes test information to improve selection accuracy. While the most popular CAT strategy, MFI has exposure control deficiencies (Chang, 2015). The second strategy is Chang, Qian, and Ying' s (2001) a-Stratified Strategy with b-Blocking (BAS), which increases exposure rates for low-discrimination items early in testing while reducing overexposed items. Additionally, stratified methods preserve stratification characteristics in generated data, allowing recommender strategies to narrow their search within specific strata, improving selection speed.

2.2 New Item Selection Strategies Based on Collaborative Filtering Recommendation

Collaborative filtering has two main implementations: User-Based Collaborative Filtering (e.g., Jia, Yang, Gao, & Chen, 2015) finds users most similar to the current user and recommends items from similar users' histories; Item-Based

Collaborative Filtering (e.g., Pirasteh, Jung, & Hwang, 2014) seeks items most similar to those the current user prefers. Since the number of past examinees typically exceeds the number of items in the bank, finding similar examinees is easier, and more reference information becomes available as the examinee pool grows, facilitating better item selection. Therefore, we designed selection strategies based on user-based collaborative filtering, with finding similar examinees as the first step. After each item response, we identify past examinees who answered the same item with the same result as "similar examinees" for that item, using them as a reference group for the next item recommendation. Unlike the cosine similarity commonly used in recommender systems, we employ a simple binary approach rather than a continuous scale to calculate examinee similarity, as this study focuses solely on dichotomously scored items where similarity judgments have only two outcomes (correct or incorrect). This approach has lower computational complexity and faster speed. The selected similar examinees apply only to the current item; non-similar examinees may become similar after the next item response. This design expands the reference scope across a complete CAT, providing more usable information for more accurate selection.

After identifying similar examinees, we can modify collaborative filtering's underlying assumptions for the CAT context. One modified assumption: the current examinee can answer the same next item as similar examinees, yielding a direct recommendation strategy that selects items without item parameters. Another assumption: similar examinees share similar ability values with the current examinee, then using item parameters for selection, yielding an indirect recommendation strategy. Both assumptions may identify multiple recommendable items. Given potential exposure imbalance in past data, the final item is selected randomly—a common method for controlling item exposure (Georgiadou, Triantafillou, & Economides, 2007). This produces two recommender strategies: Direct Examinee-Based Recommender (DEBR) selects the intersection of next items answered by all similar examinees and items not yet administered to the current examinee as the candidate pool, then randomly selects one item. Indirect Examinee-Based Recommender (IEBR) identifies the range of current ability estimates among all similar examinees after they answered the current item, then selects items not yet administered whose difficulty parameter b falls within this range as candidates, and randomly chooses one. The operation of matching ability estimates to b parameters draws from stratification methods. Compared to FMI, b-matching has lower computational complexity, faster selection speed, and better exposure control without sacrificing estimation accuracy (Chang & Ying, 1999).

In rare cases, both strategies may fail to find recommendable items, termed selection failure. Since collaborative filtering is used only in the selection process, other CAT procedures continue normally, including ability estimation methods after each item. When no recommendable items are found, the next item is selected by matching the current examinee' s ability estimate to b parameters. Beyond the advantages of b-matching mentioned earlier, if the strategy generating past data did not emphasize uniform bank usage and some items never

appeared in historical data, this method can reactivate unused items, increasing utilization of low-exposure items. In summary, both proposed recommender strategies employ simple, fast operations that prioritize exposure control while maintaining selection accuracy.

3.1 Study Design

Study 1 investigates two common factors affecting CAT item selection and recommender systems. First, does using different traditional strategies to generate past examinee data with distinct characteristics affect recommender strategy performance? The simulation conditions include two selection strategies: FMI emphasizing measurement precision and BAS emphasizing exposure control. Second, does test length, which generates different amounts of past examinee data, affect recommender strategy performance? Simulation conditions include fixed-length tests of 20 and 40 items. Study 1 has $2\times2 = 4$ simulation condition combinations, with 100 replications per combination.

The simulated item bank contains 400 dichotomously scored items based on the three-parameter logistic model (3PLM), with parameters consistent with common strategy comparison settings (Barrada, Olea, & Abad, 2010; Cheng, Patton, & Shao, 2015). Discrimination parameters a follow a normal distribution N(1.2, 0.25), difficulty parameters b follow a standard normal distribution N(0, 1), and guessing parameters c follow N(0.25, 0.02), with moderate positive correlation between a and b parameters ($r = .45$). True ability parameters follow a standard normal distribution N(0, 1). The simulation procedure for Study 1: first, simulate CAT for an initial batch of 1,000 examinees using traditional strategies to generate past examinee data; then, simulate CAT for a second batch of 1,000 examinees with the same ability distribution using recommender strategies with the generated past data. Ability estimation uses Bayesian posterior expectation. With BAS, the item bank is divided into 4 strata of 100 items each, with each examinee answering 5 or 10 items per stratum before advancing. Random item selection is included as a baseline comparison for both test lengths.

3.2 Evaluation Metrics

This study employs seven common CAT evaluation metrics (He, Diao, & Hauser, 2014) to assess measurement precision and item exposure control, plus one new metric to evaluate recommender strategies' utilization of past examinee data. Results for each condition represent means across 100 replications. Metric definitions are as follows:

(1) **Mean Squared Error (MSE):**

$$MSE = \sum(\hat{\theta}_i - \theta_i)^2$$

where $\hat{\theta}_i$ is the final ability estimate for examinee i, θ_i is the true ability, and N is the number of examinees.

(2) **Mean Absolute Error (MAE):**

$$MAE = \sum |(\hat{\theta}_i - \theta_i)|$$

(3) **Correlation between true and estimated ability** $r_{\theta,\hat{\theta}}$:

$$r_{\theta,\hat{\theta}} = \frac{\sum(\theta_i - \bar{\theta})(\hat{\theta}_i - \hat{\bar{\theta}})}{S_\theta S_{\hat{\theta}}}$$

where $\bar{\theta}$ and S_θ are the mean and standard deviation of true abilities, and $\hat{\bar{\theta}}$ and $S_{\hat{\theta}}$ are the mean and standard deviation of final ability estimates.

(4) **Chi-square value** (χ^2) comparing actual to ideal exposure distribution:

$$\chi^2 = \sum(r_i - L/K)^2$$

where r_i is the exposure rate for item i, L is test length, and K is item bank size (Chang & Ying, 1999).

(5) **Overlap Rate (OR)**, defined as the rate at which any two examinees receive identical items:

$$OR = \frac{\sum r_i^2 - \sum r_i}{N(N-1)L}$$

where r_i is item i' s exposure rate and N is the number of examinees.

(6) **Underexposed items**, defined as the number of unused items.

(7) **Overexposed items**, defined as the number of items with exposure rates exceeding 20%.

(8) **Utilization Rate of Examinees**, defined as the proportion of similar examinees called upon during each selection relative to all past examinees.

Among these eight metrics, three evaluate measurement precision (MSE, MAE, and ability correlation), two evaluate item bank usage (χ^2 and underexposed items), two evaluate test security (overlap rate and overexposed items), and utilization rate assesses how much past examinee information recommender strategies can leverage.

3.3 Results

For the 20-item CAT, traditional strategies generated past examinee data with expected characteristics: FMI showed highest measurement precision but uneven bank usage; BAS showed slightly lower precision but better test security and bank usage. When using FMI-generated data, DEBR maintained relatively high precision with only slight decreases compared to FMI, outperforming BAS

and IEBR, while substantially improving uneven item usage. IEBR demonstrated optimal exposure control, surpassing all other strategies on all four bank usage and test security metrics, achieving ideal uniform bank usage while maintaining acceptable precision (higher than random selection). Since utilization rate is unaffected by test stage, we calculated the average across a complete test; DEBR' s utilization rate was substantially higher than IEBR' s. When using BAS-generated data, both recommender strategies maintained precision comparable to BAS while further optimizing test security and bank usage, with nearly identical utilization rates.

For the 40-item CAT, traditional and recommender strategies showed consistent patterns with the 20-item condition. With FMI-generated data, DEBR sacrificed minimal precision while greatly reducing underexposed items; IEBR matched BAS precision and again achieved optimal test security and bank usage for that condition. With BAS-generated data, both recommender strategies preserved precision while further improving bank usage uniformity, with IEBR showing slightly greater improvement than DEBR. In longer tests, both DEBR and IEBR showed increased utilization rates overall, maintaining their relative trends.

Study 1 results show that different traditional strategies generating distinct past examinee data directly influence recommender strategy performance trends. Using FMI-generated data, recommender strategies substantially activate unused items to improve exposure control, creating the common trade-off of sacrificing some precision, with DEBR' s trade-off smaller than IEBR' s. Using BAS-generated data with relatively uniform bank usage, both recommender strategies preserve precision while further improving exposure control, with all metrics including utilization rate being very close. Test length does not alter the trends under specific data conditions but affects absolute values on all metrics, including higher precision and utilization rates and fewer underexposed items.

Within the same test length, the same recommender strategy can show substantial performance differences. This inconsistency arises because test length has dual pathways: it may affect traditional strategy performance and thus past data quality (absolute metric values), or it may affect recommender strategy performance through data quantity. To isolate these effects, we need to control test length and use alternative methods to increase data volume. Additionally, Study 1 generated all past examinee data using traditional strategies, whereas in reality, recommender strategies could use their own generated data after the second batch of examinees complete testing, raising questions about result stability. Study 1 used only simulated item banks, necessitating further investigation with real banks. Study 2 addresses these issues.

4.1 Study Design

Study 2 examines recommender strategy performance in more realistic settings. First, using a real item bank of less-than-ideal quality, will recommender strat-

egy performance be affected? Second, in reality, data accumulation can occur not only through longer tests but also by merging data from two different examinee groups using the same bank. Can recommender strategies maintain good measurement precision and excellent exposure control when using merged data? Here, the ratio of examinees to items increases, equivalent to a significant change in the user-item rating matrix shape in recommender systems. In contrast, increasing test length in Study 1 only changed data sparsity without altering matrix shape. To control this variable, Study 2 uses only the 20-item termination rule.

Study 2 uses 276 items from the TIMSS 2015 eighth-grade science assessment, with 125 items based on 2PLM and 151 on 3PLM. The a parameters concentrate around 1, with few high-discrimination items; b parameters have a narrower range than the simulated bank, particularly lacking low-difficulty items ($b < 0$); 3PLM c parameters are generally larger, indicating lower bank quality than Study 1' s simulated bank. The simulation procedure: first, simulate CAT for an initial batch of 1,000 examinees using traditional strategies to generate past data; then, simulate CAT for a second batch of 1,000 examinees using recommender strategies with this data (same as Study 1); finally, merge data from both batches (2,000 examinees total) as past data and simulate CAT for a third batch of 1,000 examinees using recommender strategies. With BAS, the bank is divided into 4 strata of 69 items each, with 5 items per stratum. Traditional strategies, ability distributions, estimation methods, replications, and evaluation metrics match Study 1.

4.2 Results

Compared to the 20-item condition in Study 1, FMI and BAS generated data with consistent characteristics but lower quality using the real bank. With FMI-generated initial data, both recommender strategies showed the same patterns as Study 1: substantially improving exposure imbalance while DEBR prioritized precision maintenance and IEBR used the bank more uniformly. Both strategies called upon nearly double the number of similar examinees compared to the simulated bank, with DEBR still far exceeding IEBR. After merging FMI and recommender-generated data for the third batch, both strategies showed even greater exposure control improvement. DEBR' s precision remained higher than IEBR and BAS, while IEBR achieved the most ideal exposure control. The number of similar examinees found by DEBR and IEBR remained essentially unchanged from before merging; however, with doubled past examinees, utilization rates halved, approaching Study 1' s 20-item results.

With BAS-generated initial data, both recommender strategies produced similar results, with DEBR showing slight precision improvements and IEBR further reducing χ^2 values and overlap rates, calling upon similar numbers of examinees at lower levels than under FMI conditions. After merging both batches, DEBR also improved test security and bank usage, with IEBR showing more pronounced improvements while measurement precision fluctuations remained

reasonable. Notably, after merging, DEBR found double the similar examinees, keeping utilization rates essentially unchanged, while IEBR called upon the same number of examinees as before, halving its utilization rate.

Examining exposure rate changes across all items during two iterations reveals patterns consistent with utilization rate changes. With FMI-generated initial data (see [Figure 1: see original paper], where the red horizontal line indicates ideal uniform exposure rate $r_{ideal} = 0.072$), DEBR' s precision-exposure trade-off was smaller (Figure 1b), with first-round results closer to FMI (Figure 1a), making it easier to find similar examinees and yielding higher utilization rates. IEBR selected more rarely used items, creating greater exposure improvement (Figure 1d) but drastically reducing similar examinees and utilization rates. With merged data for the second round, both strategies improved exposure control (Figures 1c and 1e), with utilization rates decreasing proportionally and matching final exposure optimization outcomes. The same patterns explain BAS-generated initial data (see [Figure 2: see original paper]): first-round trade-off trends and utilization rates were very close (Figures 2b and 2d), and because BAS has some exposure control (Figure 2a), DEBR and IEBR utilization rates fell between the two FMI conditions. In the second round, DEBR essentially reached its exposure optimization limit (Figure 2c) with minimal utilization rate change, while IEBR continued markedly improving bank usage uniformity (Figure 2e), further reducing utilization rate. Thus, utilization rate can serve as a side measure of recommender strategy characteristics and trade-off trends.

Study 2 results show that using a less-than-ideal real bank does not affect the characteristics and good properties of the two recommender strategies. Merging data from traditional and recommender strategies—thereby increasing data quantity without changing data characteristics—makes DEBR and IEBR' s advantages more pronounced and distinctive.

5 Discussion

This study proposes novel CAT item selection strategies based on collaborative filtering recommendation. Two simulation studies reveal that recommender strategies leveraging past examinee data can ensure good test security and uniform bank usage while achieving precision no lower than stratified methods. In specific CAT scenarios, if past data shows imbalanced bank usage, recommender strategies first activate the entire bank, achieving a good balance between precision and exposure control; when past data lacks extreme exposure imbalance, recommender strategies further optimize exposure control without sacrificing precision. Between the two new strategies, DEBR emphasizes precision maintenance, while IEBR demonstrates stronger exposure control improvement. Both simulation studies show that the characteristics of past examinee data—determined by different traditional strategies—most strongly influence recommender strategy trends, while bank quality, test length, and examinee quantity affect only the absolute metric values by influencing data quality.

This research has two major innovations. First, it identifies the value of past examinee data as prior information for item selection. By bridging current examinee data with extensive historical data, we expand the sources and quantity of information available for CAT selection. Simulation results demonstrate that with sufficient, accurate historical selection data, learning from others' selection experience can identify items matching current examinee abilities while improving past uneven item usage. Compared to data from current examinees, past examinee data is undoubtedly richer with tremendous untapped potential. Second, this research reveals the commonality between recommender systems and CAT item selection, adapting collaborative filtering technology to establish a selection rule set and providing initial evidence that collaborative filtering assumptions apply to CAT contexts. This assumption enables organic integration of recommender system technology with traditional methods to design flexible recommender strategies. For instance, DEBR and IEBR' s excellent uniform bank usage stems from incorporating multiple exposure control operations into user-based recommendation, demonstrating that recommender strategies constitute an improvable framework with room for making adaptive testing more precise and intelligent. As research progresses, particularly with more recommender system integration, dependence on traditional strategies for initial data generation or preventing selection failure may gradually decrease, reducing susceptibility to issues like violated IRT assumptions. This exploration also encourages more psychology and education researchers to consider big data technologies and machine learning algorithms—exemplified by recommender systems—as options to combine with or replace traditional methods.

In both simulation studies, estimation accuracy across ability levels remained dependent on past data quality, generally matching the performance of the strategy that generated the data: higher precision for middle-ability examinees and lower precision for those at distribution extremes, but never below the precision level in the past data. Selection failure probabilities were very small. In the 40-item condition, selecting 40,000 items for 1,000 examinees, DEBR' s average failure rate was 1.15% (462 items) and IEBR' s was 2.03% (812 items)—less than one failure per examinee on average, making the choice of fallback method for failures have minimal impact on precision and exposure. The probability of failing to find similar examinees was even lower, occurring primarily with FMI-generated imbalanced initial data; in other conditions, the probability was less than one in ten thousand. Thus, only a traditional strategy with good exposure control, simulating data from several thousand examinees and serving as a fallback for selection failures, is needed to confidently use recommender strategies for subsequent examinees—whose data can then augment the past examinee pool, continuously increasing reference information diversity while reducing selection failure probability.

As an exploratory attempt at a new method, this study has several limitations warranting further investigation. First, while we examined data quality, characteristics, and quantity—factors most likely affecting recommender performance —we did not deeply analyze examinee ability distribution characteristics and

item bank features that may affect CAT selection strategies. Future research could examine impacts of ability distribution differences between past and new examinees, bank size and parameter distributions, response patterns, and response accuracy on recommender strategy precision and failure rates. Second, as past examinee data volume increased, measurement precision actually decreased, possibly because our design emphasized solving exposure imbalance without additional operations to further improve precision, limiting new strategies' ability to maintain high precision with larger datasets. Future improvements could address this limitation. Third, the proposed strategies apply only to unidimensional, dichotomously scored CAT, whereas many polytomously scored items exist in practice, and more complex multidimensional CAT and cognitive diagnosis CAT based on different IRT models present contemporary research hotspots and challenges (Akbay & Kaplan, 2017; Kaplan, de la Torre, & Barrada, 2015; Zhang & Chang, 2016; Mao & Xin, 2015). Adapting recommender strategies for these complex models represents an important research direction.

Based on our results and these issues, we propose several improvement directions: First, further integrate traditional strategies—for IEBR, replace *b*-matching with more precise selection methods after finding similar examinees. Second, modify similar examinee definitions, such as considering responses from several previous items or borrowing various similarity metrics from recommender systems to find more precise similar examinees and improve selection accuracy. Third, collaborative filtering also includes item-based recommendation, which calculates item similarity applicable to CAT to select unadministered items most similar to answered items, better avoiding selection failure and facilitating selection of newly added, unused items. Fourth, collaborative filtering often struggles with recommendations for new users with scarce data—the "cold start" problem. Methods developed to address this (Lika, Kolomvatsos, & Hadjiefthymiades, 2014) could solve early-test measurement inaccuracy and selection failure issues. Fifth, beyond collaborative filtering, many new recommender system technologies could improve CAT selection, such as model-based recommendation using diverse machine learning algorithms to build complex models from rating data (Ricci, Rokach, & Shapira, 2015), enhancing collaborative filtering' s predictive power and flexibility while providing more options for migrating recommender systems to CAT. In recent years, deep learning has combined with recommender systems to create deep recommendation algorithms addressing massive data and complex recommendation problems (Covington, Adams, & Sargin, 2016; H. Wang, N. Wang, & Yeung, 2015), offering similar promise for large-scale, complex CAT item selection.

6 Conclusion

This study finds: (1) Collaborative filtering recommendation from recommender systems can be adapted to CAT item selection, yielding strategies that ensure certain measurement precision while providing better item exposure control; (2) Past examinee data represents valuable prior information for item selection, with

its characteristics and quality being the primary factors affecting recommender strategy performance.

References

Akbay, L., & Kaplan, M. (2017). Transition to multidimensional and cognitive diagnosis adaptive testing: An overview of cat. *The Online Journal of New Horizons in Education-January*, 7, 206-214.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement*, 34, 438-452.

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80, 1-20.

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

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

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40, 129-145.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation system for adaptive learning. *Applied psychological measurement*, 42, 24-41.

Cheng, Y., Patton, J. M., & Shao, C. (2015). a-stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement*, 75, 260-283.

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191-198). Boston, MA: ACM.

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, 1-39.

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

Jia, Z., Yang, Y., Gao, W., & Chen, X. (2015, February). User-based collaborative filtering for tourist attraction recommendations. In *Computational Intelligence & Communication Technology, 2015 IEEE International Conference on* (pp. 22-25). Ghaziabad, India: IEEE.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied psychological measurement*, 39, 167-188.

Klašnja-Milićević, A., Ivanović, M., & Nanopoulos, A. (2015). Recommender systems in e-learning environments: A survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, 44, 571-604.

Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 77-118). Boston, MA: Springer.

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, 41, 2065-2073.

Liu, Q., Chen, E. H., Zhu, T. Y., Huang, Z. Y., Wu, R. Z., Su, Y., & Hu, G. P. (2018). Research on educational data mining for online intelligent learning. *Pattern Recognition and Artificial Intelligence*, 31, 77-90.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Mao, X. Z., & Xin, T. (2011). Item selection method in computerized adaptive testing. *Advances in Psychological Science*, 19, 1552-1562.

Mao, X. Z., & Xin, T. (2015). Multidimensional computerized adaptive testing: Model, techniques and methods. *Advances in Psychological Science*, 23, 907-918.

Pirasteh, P., Jung, J. J., & Hwang, D. (2014, April). Item-based collaborative filtering with attribute correlation: A case study on movie recommendation. In N. T. Nguyen, B. Attachoo, B. Trawiński, & K. Somboonviwat (Eds.), *Asian Conference on Intelligent Information and Database Systems* (pp. 245-252). Cham, Switzerland: Springer.

Quijano-Sánchez, L., Recio-García, J. A., Díaz-Agudo, B., & Jiménez-Díaz, G. (2011, March). Happy movie: A group recommender application in facebook. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference* (pp. 127-134). Palm Beach, FL: AAAI.

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 1-34). Boston, MA: Springer.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE Internet Computing*, 21, 12-18.

Tan, P. N., Steinbach, M., & Kumar, V. (2007). *Introduction to data mining*. IEEE Transactions on Knowledge & Data Engineering, 22, 1-12.

Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., & Schmidt-Thieme, L. (2010). Recommender system for predicting student performance. *Procedia Computer Science*, 1, 2811–2819.

Wang, H., Wang, N., & Yeung, D. Y. (2015, August). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1235–1244). Sydney, NSW, Australia: ACM.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.

Zhang, S., & Chang, H. H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 1, 67–92.

Zhu, T. Y., Huang, Z. Y., Chen, E. H., Liu, Q., Wu, R. Z., Wu, L., ···Hu, G. P. (2017). Cognitive diagnosis based personalized question recommendation. *Chinese Journal of Computers*, 40, 176–191.

Note: Figure translations are in progress. See original paper for figures.