Item Difficulty Constrained Uniform Adaptive Testing

Wakaba Kishida[1], Kazuma Fuchimoto[1],
Yoshimitsu Miyazawa[2], and Maomi Ueno[1]

[1] The University of Electro-Communications, Tokyo, Japan
{kishida,fuchimoto}@ai.lab.uec.ac.jp
ueno@ai.is.uec.ac.jp
[2] The National Center for University Entrance Examinations, Tokyo, Japan
miyazawa@rd.dnc.ac.jp

Abstract. Computerized adaptive testing tends to select and present items frequently with high discrimination parameters because these items can discriminate examinees' abilities in a wide range. Unfortunately, that tendency leads to bias of item exposure. To address this shortcoming, we propose item difficulty constrained uniform adaptive testing. During the initial stage, an optimal item is selected and presented from a uniform item group generated by a modern uniform test assembly method. The method switches to the secondary stage when the examinee's ability converges. It selects and presents the optimal item with a difficulty parameter value near the examinee's ability estimate from the whole item pool. Empirical experiments demonstrate that the proposed method mitigates the bias of item exposure while maintaining low measurement error by reducing the number of presented items with high discrimination parameters which are likely to be presented frequently by earlier CAT methods.

Keywords: computerized adaptive testing · item response theory · uniform test assembly

1 Introduction

Computerized adaptive testing (CAT) selects and presents an optimal item from an item pool. That process, which is based on item response theory (IRT), maximizes the test information (Fisher information measure) at the examinee's current estimated ability [6, 9]. However, conventional CAT often presents identical items from an item pool to examinees with similar abilities. That extremely increases bias of item exposure distribution. The bias leads to decreasing the reliability of tests because overexposed items are likely to be shared among examinees [6, 9, 11].

To resolve this shortcoming, various countermeasures and alternatives are proposed (e.g. [1, 5, 9, 10]). Kingsbury and Zara [5] proposed a method, of dividing item pools into item groups. Thereafter, from the item group with the smallest value of item exposure among all of them, it selects and presents the optimal item (designated as KZ). Moreover, van der Linden [9] proposed a method

selecting the optimal item from a shadow-test assembled by solving an integer programming problem with several constraints (designated as IP). Choi and Lim [1] proposed another shadow-test approach that minimizes the distance between a test information of a shadow-test and target information (designated as TI). As another approach, van der Linden and Choi [10] proposed a method controlling the item selection probabilistically (designated as Prob). However, these methods increase the bias of measurement accuracies among examinees.

Therefore, Ueno and Miyazawa [7] proposed a method that separates an item pool into several item groups using the test assembly method presented by Ishii et al. [3] in advance. This method was designated as uniform adaptive testing (UAT). The method selects and presents an item from a uniform item group assigned to each examinee. Their results demonstrated that UAT reduced the bias arose for measurement accuracies. However, UAT increases the measurement error through reduction of the item group size.

To overcome this difficulty, Ueno and Miyazawa [8] proposed two-stage uniform adaptive testing (TUAT). Initially, this method selects and presents the optimal item from a uniform item group generated by the method presented by Ishii and Ueno [4]. After the examinee's ability converges, the method switches to the secondary stage to select and present the optimal item from the whole item pool. They demonstrated that item exposure can be reduced by TUAT without any increase in the measurement error. Unfortunately, TUAT shows a marked tendency for frequent selection and presentation of items with high discrimination parameters because these items can discriminate examinee's abilities in a wide range. Consequently, reduction of bias of the item exposure by TUAT can be done only to a limited degree.

Therefore, we propose item difficulty constrained uniform adaptive testing. The proposed method generates numerous item groups in advance using the Hybrid Maximum Clique Algorithm with Parallel Integer Programming presented by Fuchimoto et al. [2], which assembles the greatest quantity of uniform tests. Similarly to TUAT, the algorithm initially selects and presents an optimal item from a uniform item group. When the examinee's ability converges, the proposed method subsequently selects and presents an optimal item with a difficulty parameter value near the examinee's ability estimate from the whole item pool. Empirical experimentation elucidate that the proposed method can mitigate the bias of item exposure while maintaining low measurement error.

2 Item Difficulty Constrained Uniform Adaptive Testing

To resolve the shortcomings presented by a state-of-the-art CAT, TUAT [8], this study proposes a new CAT method, item difficulty constrained uniform adaptive testing, which can reduce bias of item exposure.

2.1 Initial procedure

The method proposed herein generates a large number of uniform item groups using Hybrid Maximum Clique Algorithm with Parallel Integer Programming,

which was demonstrated by Fuchimoto et al. [2]. The uniform item group assembly method differs from that of TUAT [8].

The algorithm of the initial stage is similar to TUAT [8]. At the beginning of the initial stage, the method assigns a different uniform item group to each examinee. During this stage, an optimal item from a uniform item group is selected and presented to maximize Fisher information. This stage provides a rough ability estimate with keeping item exposure distribution as uniform as possible (See [8] for details.).

2.2 Secondary procedure

The secondary procedure provides a more accurate ability estimate with preventing bias of item exposure from increasing. Similarly to TUAT, the method finishes the initial stage when the update difference of the estimate of an examinee's ability is less than a criterion value, which is the Switching Stage Criterion (SSC) [8]. Subsequently, the proposed method starts the secondary procedure. From the whole item pool, the method selects and presents an optimal item with a difficulty parameter value within the neighborhood of the examinee's ability estimate. The neighborhood interval of the examinee's ability estimate $\hat{\theta}$ is defined as

$$\hat{\theta} - \alpha SE(\hat{\theta}) < b < \hat{\theta} + \alpha SE(\hat{\theta}), \tag{1}$$

where $SE(\hat{\theta})$ represents the standard error of the examinee's ability estimate $\hat{\theta}$, and α denotes a hyperparameter. The SSC and the hyperparameter α are optimized to balance low measurement error and low bias of item exposure. More specifically, the selection procedure in this stage is as follows:

1. The difficulty interval is estimated from the current ability estimate $\hat{\theta}$ and its standard error.
2. From items with difficulty parameter values within the estimated difficulty interval, an optimal item that maximizes Fisher information is selected.
3. Based on the examinee's earlier response history, the current ability estimate is updated.
4. Procedures 1–3 are iterated until the update difference of the ability estimate falls to or below a constant value of ϵ.

The proposed method is expected to reduce the quantity of presented items with high discrimination parameters while maintaining low measurement error.

3 Empirical Evaluation

This section evaluates the effectiveness of the proposed method (Proposed) through comparison with earlier computerized adaptive testing methods: conventional CAT (designated as CAT), IP [9], TI [1], KZ [5], and TUAT [8]. We set the total test length as 30. The item group sizes used in KZ, TUAT and Proposed are equal to the test length.

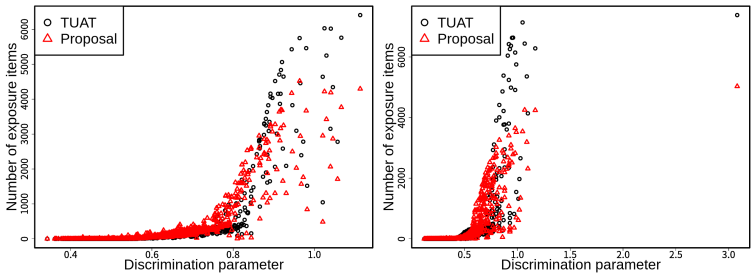
Table 1: Experiment results

Item pool	Method	SD. exposure item	Max. No. exposure items	Number of non-presented items	Measurement error(RMSE)
simulated	CAT	1055.5	10000	832	**0.25**
	IP	984.0	5000	812	**0.25**
	Prob	987.8	5105	819	**0.25**
	TI	1000.4	10000	**0**	0.26
	KZ	918.0	6565	779	0.26
	TUAT(0.100)	864.7	6409	188	0.26
	Proposed (0.100, 0.8)	**682.3**	**4520**	68	0.26
real	CAT	1150.3	10000	836	**0.25**
	IP	1026.4	5000	809	0.26
	Prob	1034.8	5107	812	0.26
	TI	1047.6	10000	**7**	0.26
	KZ	1032.0	7364	792	0.26
	TUAT(0.075)	937.6	7381	274	0.26
	Proposed (0.100,0.6)	**672.7**	**5031**	263	0.27

A simulated item pool including 1000 items and a real item pool including 978 items were used to conduct experiments. For each item included in the simulated item pool, true parameters were generated from $\log a_i \sim N(-0.5, 0.2)$ and $b_i \sim N(0, 1)$, where a_i and b_i respectively signify the discrimination parameter of item i and the difficulty parameter of item i. The examinees' actual abilities are sampled from $\theta \sim N(0, 1)$ 10,000 times. For convergence to the same upper bound exposure counts, we performed our experiments with 5000 as IP upper bound exposure counts and with 0.5 as Prob upper bound exposure rate. A presented in Table 1, the results shown as the values in parentheses for TUAT represent the SSC value. Those for the Proposed represent the SSC value and hyperparameter α. Also, "SD. exposure item" stands for the standard deviation of the numbers of exposure items; "Max. No. exposure item" represents the maximum quantity of exposure items. The quantity of items which have not been presented is signified by the "Number of non-presented items".

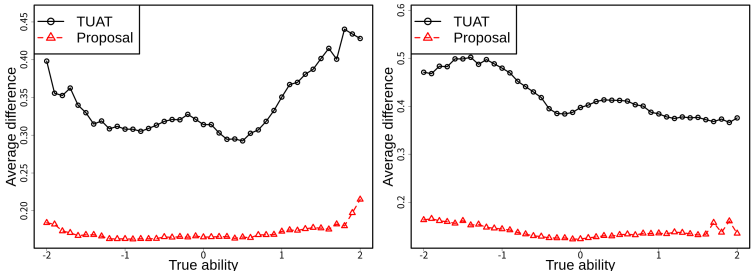
Table 1 shows that TI provides the lowest values of "Number of non-presented items". However, the values of "SD. exposure item" are as large as those of CAT. Moreover, CAT and TI produce equal values of "Max. No. exposure item" as the number of examinees. These findings indicate that one or more items are exposed to all the examinees. Actually, IP, Prob, and KZ all provide lower values of "SD. exposure item" and "Max. No. exposure item" than those of CAT, but "Number of non-presented items" is still large. By contrast, Proposed provides the lowest values of "SD. exposure item" and "Max. No. exposure items" without increasing the measurement error considerably. Furthermore, Proposed has the second lowest values of "No. non-presented items". Next, we analyze the difference between TUAT and Proposed.

Figures 1a and 1b portray scatter plots of the number of exposure items and items' discrimination parameters for TUAT and Proposed. These figures indicate the important tendency of TUAT as able to select and present items with high discrimination parameters because these items can discriminate examinees' abilities in a wide range. A point of marked contrast is that the proposed method



(a) simulated (b) real

Fig. 1: Scatter plots presenting the numbers of presented items and discrimination parameters.



(a) simulated (b) real

Fig. 2: Average difference between a difficulty parameter and an ability estimate in the secondary procedure

yields far fewer presented items with high discrimination parameters. Next, we analyze the reasons underlying this phenomenon.

Figures 2a and 2b depict the average differences between difficulty parameters and ability estimates in the secondary procedure. The difference between a difficulty parameter of the item presented to an examinee in the secondary procedure and the ability estimate is determined as

$$\sqrt{\frac{1}{n-l+1}\sum_{k=l}^{n}(\hat{\theta}_{k-1}-b_k)^2}, \tag{2}$$

where b_k denotes the difficulty parameter of the k-th presented item, and $\hat{\theta}_k$ represents the ability estimate after the k-th item is presented. Items from l-th to n-th are presented in the secondary procedure. Figures 2a and 2b portray an important benefit of TUAT: it often selects items with difficulty parameter values that differ greatly from the ability estimates. In contrast, the proposed method selects items with difficulty parameter values that are approximately equal to the ability estimates. As described previously, a marked tendency of TUAT is the selection items with high discrimination parameters, even when

the difficulty parameter values differ greatly from the ability estimates. This tendency consequently leads to bias of the item exposure. The proposed method avoids selection of items with difficulty parameter values that differ greatly from the ability estimates. As a result, the proposed method relaxes the item exposure bias that is a problem in TUAT.

4 Conclusion

First, the findings presented herein indicate that TUAT [8], which is a state-of-the-art CAT, has a tendency for the selection and presentation of items with high discrimination parameters frequently. To resolve this shortcoming, this study proposed item difficulty constrained uniform adaptive testing. Results of the empirical experiments showed that the proposed method provides a lower bias of item exposure than all comparison methods while maintaining low measurement error. That performance was achieved by reducing the number of presented items with high discrimination parameters, which are presented frequently by the earlier TUAT. Application of the proposed method is limited to IRT models with difficulty parameters. Relaxing the constraints of the proposed method is a goal for our future work.

References

1. Choi, S.W., Lim, S.: Adaptive test assembly with a mix of set-based and discrete items. Behaviormetrika **49**, 231–254 (2022)
2. Fuchimoto, K., Ishii, T., Ueno, M.: Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly. IEEE Transactions on Learning Technologies **15**(2), 252–264 (2022)
3. Ishii, T., Songmuang, P., Ueno, M.: Maximum clique algorithm and its approximation for uniform test form assembly. IEEE Transactions on Learning Technologies **7**(1), 83–95 (2014)
4. Ishii, T., Ueno, M.: Algorithm for uniform test assembly using a maximum clique problem and integer programming. In: Artificial Intelligence in Education. pp. 102–112 (2017)
5. Kingsbury, G.G., Zara, A.R.: Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education **2**(4), 359–375 (1989)
6. Ueno, M., Fuchimoto, K., Tsutsumi, E.: e-testing from artificial intelligence approach. Behaviormetrika **48**(2), 409–424 (2021)
7. Ueno, M., Miyazawa, Y.: Uniform adaptive testing using maximum clique algorithm. In: Artificial Intelligence in Education. pp. 482–493 (2019)
8. Ueno, M., Miyazawa, Y.: Two-Stage uniform adaptive testing to balance measurement accuracy and item exposure. In: Artificial Intelligence in Education. pp. 626–632 (2022)
9. van der Linden, W.J.: Review of the shadow-test approach to adaptive testing. Behaviormetrika pp. 1–22 (2021)
10. van der Linden, W.J., Choi, S.W.: Improving item-exposure control in adaptive testing. Journal of educational measurement **57**(3), 405–422 (2020)
11. Way, W.D.: Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and Practice **17**, 17–27 (1998)