Two-stage uniform adaptive testing to balance measurement accuracy and item exposure

Maomi Ueno[1] and Yoshimitsu Miyazawa[2]

[1] The University of Electro-Communications, Tokyo, Japan
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 (CAT) presents a tradeoff problem involving increasing measurement accuracy vs. decreasing item exposure in an item pool. To address this difficulty, we propose two-stage uniform adaptive testing. In the first stage, the proposed method partitions an item pool into numerous uniform item groups using a state-of-the-art uniform test assembly technique based on the Random Integer Programming Maximum Clique Problem. Then the method selects the optimum item from a uniform item group. In the second stage, when the standard error of an examinee's ability estimate becomes less than a certain value, it switches to selecting and to presenting an optimum item from the whole item pool. Results of numerical experiments underscore the effectiveness of the proposed method.

Keywords: computerized adaptive testing, integer programming, item response theory, maximum clique algorithm, uniform adaptive testing

1 Introduction

Computerized adaptive testing (CAT) selects and presents the optimal item which maximizes the test information (Fisher information measure) at the current estimated ability based on item response theory (IRT) from an item pool. However, in conventional CATs, the same items tend to be presented to examinees who have similar abilities. This tendency leads to bias of the item exposure frequency in an item pool.

To resolve this difficulty, various methods have been proposed (e.g. [1–3]). Recent studies by Songmuang and Ueno [4] and by Ishii and several collaborators [5–7] have explored several techniques using AI technologies to generate numerous uniform test forms from an item pool. Regarding the uniform test forms, each form consists of a different set of items, but the forms have equivalent measurement accuracy (i.e. equivalent test information based on item response theory). Ueno and Miyazawa [8] proposed uniform adaptive testing (UAT) using the Maximum Clique Problem (MCP) described by Ishii et al. [6] to divide an item pool into several equivalent groups of items (uniform item groups) and then select the optimum item from a uniform item group. They demonstrated that the UAT reduced test length, item exposure, and bias of measurement accuracies among examinees although they did not evaluate the measurement accuracy directly. However, the UAT must degrade the measurement accuracy of examinees' abilities because decreasing the test length necessarily increases the measurement error.

To resolve that shortcoming, we propose two-stage uniform adaptive testing. In the first stage, the proposed method partitions an item pool into numerous uniform item groups using the Random Integer Programming Maximum Clique Problem (RIPMCP) presented by Ishii and Ueno [9], which is known to generate the greatest number of uniform tests. Then the method selects the optimum item from a uniform item group. In the second stage, when the examinee's ability estimate error becomes less than a certain value, designated as Switching Stage Criterion (SSC), in the uniform item group, the proposed method switches to the selection and presentation of the optimum item from the whole item pool until the update difference of the examinee's ability estimate becomes less than a constant value. Numerical experiments demonstrate that the proposed method reduces item exposure without increasing the measurement error.

2 Computerized Adaptive Testing Based on Item Response Theory

In CAT, an examinee's ability parameter is estimated based on Item Response Theory (IRT) ([10]) to select the optimum item with the highest information. In the two-parameter logistic model (2PLM), the most popular IRT model, the probability of a correct answer to item i by examinee j with ability $\theta \in (-\infty, \infty)$ is assumed as

$$p(u_i = 1|\theta) = \frac{1}{1 + exp[-1.7a_i(\theta - b_i)]}. \tag{1}$$

Therein, u_i is 1 when an examinee answers item i correctly; it is 0 otherwise. Furthermore, $a_i \in [0, \infty)$ and $b_i \in (\infty, \infty)$ respectively denote the discrimination parameter of item i and the difficulty parameter of item i. The asymptotic variance of estimated ability based on the item response theory was shown by [10] to approach the inverse of Fisher information. Accordingly, item response theory usually employs Fisher information as an index representing the accuracy. In 2PLM, the Fisher information is defined when item i provides an examinee's ability θ using the following equations.

$$I_i(\theta) = \frac{[\frac{\partial}{\partial\theta} p(u_i = 1|\theta)]^2}{p(u_i = 1|\theta)[1 - p(u_i = 1|\theta)]} \tag{2}$$

The results imply that the examinee's ability can be discriminated using an item with high Fisher information $I_i(\theta)$. Accordingly, that ability estimation can be expected to be implemented by selecting items with the highest amount of Fisher information given an examinee's ability estimate $\hat{\theta}$. The test information function $\mathrm{I}_T(\theta)$ of a test form T is defined as $\mathrm{I}_T(\theta) = \sum_{i \in T} \mathrm{I}_i(\theta)$. The asymptotic error of ability estimate $\hat{\theta}$, $\mathrm{SE}(\hat{\theta})$, can be obtained as the inverse of square root of the test information function at a given ability estimate $\hat{\theta}$ as $\mathrm{SE}(\theta) = \frac{1}{\sqrt{\mathrm{I}_T(\theta)}}$.

In conventional CAT, adaptive items are selected from an item pool using the following procedures.

1. An examinee's ability is initialized to $\hat{\theta} = 0$.
2. An item maximizing Fisher information for a given ability is selected from the item pool. It is then presented to the examinee.
3. The examinee's ability estimate is updated from the correct and incorrect response data to the item.
4. Procedures 2 and 3 are subsequently repeated until the update difference of the examinee's ability estimate decreases to a constant value of ϵ or less.

Consequently, CAT can reduce the number of items examined, but it does not reduce the test accuracy in comparison to that of the same fixed test.

3 Two-stage Uniform Adaptive Testing

In a conventional CAT, it is highly likely that the same set of items will be presented to examinees exhibiting similar abilities. Therefore, conventional CAT cannot be used practically in situations where the same examinee can take a test multiple times. Furthermore, because the ability variable follows a standard normal distribution, items with higher information around $\theta = 0$ tend to be exposed frequently. Therefore, bias of the item exposure frequency occurs in an item pool. To resolve the shortcoming, various constrained CATs with item exposure control have been proposed (e.g. [1–3]). Earlier methods have mitigated the bias of item exposure frequency in an item pool. Unfortunately, they also entailed the important difficulty of increased measurement error for examinees. In fact, a tradeoff exists between minimizing item exposure and maximizing the measurement accuracy. Nevertheless, earlier methods did not resolve the tradeoff. For that reason, we propose a new CAT framework that can resolve the tradeoff: two-stage uniform adaptive testing.

3.1 First stage procedure

In the first stage, the proposed method partitions an item pool into numerous uniform item groups similarly to UAT, a method presented by Ueno and Miyazawa [8]. Although UAT employs MCP, which was introduced by Ishii et al. [6], the number of generated uniform item groups remains limited because of its heavy space complexity. In addition, MCP tends to engender a bias of item exposure frequency because it does not consider the bias.

A state-of-the-art uniform test assembly method, Random Integer Programming Maximum Clique Problem (RIPMCP), has been demonstrated by Ishii and Ueno [9] to generate the greatest number of uniform tests. Although the shadow-test method [3] maximizes the test information using integer programing, it increases the difference of measurement accuracies between the first assembled shadow test and the last one. In contrast, the proposed method maximizes the number of uniform item groups with the test constraints, so as not to increase the bias of measurement accuracy for the groups. In the first stage, the proposed method partitions an item pool into numerous uniform item groups using the RIPMCP. The method then selects the optimum item from a uniform item group as described below.

1. An arbitrary uniform item group is selected from a set of unused groups.
2. The optimal item maximizing Fischer information is selected from the group and is presented to an examinee in Procedure 1.
3. The examinee's ability estimate is updated from the examinee's response.
4. Procedures 2 and 3 are repeated until the asymptotic error of ability estimate $\mathrm{SE}(\hat{\theta})$ reaches a constant value of ε or less.

If a set of unused groups is empty in Procedure 1, then the algorithm resets it as a universal set of uniform item groups. The number of groups is optimized by comparing the respective performances of several numbers of groups. Item selection from a uniform item group accelerates convergence of the ability estimate to the neighborhood of the true ability value because the item difficulties in each group are distributed sparsely and uniformly over all the examinees' abilities.

3.2 Second stage procedure

The first stage rapidly provided a roughly approximated ability estimate of an examinee. The second stage reaches a more accurate ability estimate of the examinee. More specifically, when the examinee's ability estimate error becomes less than the determined value, designated as Switching Stage Criterion (SSC), in the first stage, it switches to the second stage, which selects and presents the optimum item from the whole item pool. The second stage is conducted until the update difference of the examinee's ability estimate becomes less than a constant value or less, just as traditional CATs do. The SSC is optimized by changing the value to compare performance. For this study, we use the Fischer information measure as an item selection criterion that becomes accurate for the second stage because it is an asymptotic approximation. Therefore, the second stage is expected to approach the true ability value efficiently and rapidly without greatly increasing the item exposure.

4 Numerical Evaluation

This section presents a comparison of the performances of the proposed method (designated as Proposal) to those of other computerized adaptive testing methods (conventional adaptive testing in 2 (designated as CAT), Kingsbury and Zara [1] CAT (designated as KZ), van der Linden's IP-based CAT [3] (designated as IP), Linden and Choi's item-eligibility probability method [2] (designated as Prob) and the method described by Ueno and Miyazawa [8] (designated as UAT). Additionally, we evaluate the performances of the UAT employing RIPMCP to

Table 1. Experiment results obtained using an actual item pool

Test length	Method	No. item-groups	Avg. exposure item	Measurement error (RMSE)	No. non-presented items
30	CAT	-	227.27 (227.99)	0.24	846
	KZ(20)	48	131.58 (140.35)	0.29	750
	IP	-	80.86 (33.28)	0.33	607
	Prob.	-	95.85 (40.83)	0.34	665
	UAT(20)	215	20.94 (12.05)	0.50	23
	UAT-RIPMCP(20)	342	20.47 (8.91)	0.54	1
	Proposal(20, 0.225)	342	80.21 (163.75)	0.24 (0.69)	604
50	CAT	-	243.90 (233.59)	0.20	773
	KZ(25)	39	165.56 (198.94)	0.23	676
	IP	-	83.61 (31.66)	0.29	380
	Prob.	-	104.60 (39.98)	0.27	500
	UAT(20)	215	20.94 (12.06)	0.48	23
	UAT-RIPMCP(20)	342	20.47 (8.91)	0.52	1
	Proposal(20, 0.075)	342	69.83 (151.16)	0.20 (0.57)	284

generate uniform item groups designated as UAT-RIPMCP. Furthermore, we employ OC=5 for proposal, UAT, and UAT-RIPMCP.

An experiment was conducted using the item pool of real data, with 978 items, and a test constraint. Table 1 presents the results. In Table 1, the values in parentheses for KZ, UAT, and UAT-RIPMCP denote the group sizes. Those for Proposal represent the uniform item group sizes and SSC values. "Avg. exposure item" expresses the average exposure count of an item (the standard error of numbers of exposure items in parentheses), and "No. non-presented items" represents the number of items that have not been presented. The average test lengths (the standard error in parentheses) in the first stage for the total test lengths 30 and 50 are, respectively, 3.83 (1.11) and 9.65 (2.30). Those in the second stage for the total test lengths 30 and 50 are the remaining test lengths, respectively, 26.17 and 40.35. The average test lengths for the total test lengths 30 and 50 show large differences when compared to those in the simulation experiments because of their large difference of the optimum SSC values. Otherwise, the table lays out results that are almost identical to those obtained from the simulation experiment. The RMSEs in the first stage for the total test lengths 30 and 50 are, respectively, 0.69 and 0.57 and those in the second stage for the total test lengths 30 and 50 are, respectively, 0.24 and 0.20. In fact, results indicate that the proposed method reduces item exposure without increasing the measurement error. The results demonstrate that only the proposed method resolves the tradeoff problem between increasing measurement accuracy and decreasing item exposure.

5 Conclusion

The discussion and results presented herein have demonstrated that CAT entails tradeoff difficulties between increasing measurement accuracy and decreasing item exposure in an item pool. To address this difficulty, we proposed two-stage uniform adaptive testing. Experiments were conducted to compare the performance of the proposed method with that demonstrated by conventional methods. Results of those experiments demonstrated that, among all methods, only the proposed method resolved the tradeoff. We expect to apply the proposed uniform adaptive testing method to adaptive learning systems [11, 12] and Deep IRT [13, 14] in future studies.

References

1. Kingsbury, G.G. and Zara, A.R.: Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, **2**(4), 359–375 (1989)
2. van der Linden, W. J., and Choi, S. W. Improving Item - Exposure Control in Adaptive Testing. Journal of educational measurement, **57**(3), 405–422 (2020)
3. van der Linden, W.J.: Review of the shadow-test approach to adaptive testing. Behaviormetrika (2021). https://doi.org/10.1007/s41237-021-00150-y
4. Songmuang, P. and Ueno, M.: Bees algorithm for construction of multiple test forms in e-testing. IEEE Transactions on Learning Technologies, **4**(3), 209–221 (2011)
5. Ishii, T., Songmuang, P., and Ueno, M.: Maximum Clique Algorithm for Uniform Test Forms Assembly. International Conference on Artificial Intelligence in Education (AIED), LNAI 7926, 451–462, (2013)
6. Ishii, T., Songmuang, P., and Ueno, M.: Maximum clique algorithm and its approximation for uniform test form assembly. IEEE Transactions on Learning Technologies **7**(1), 83–95 (2014)
7. Ishii, T. and Ueno, M.: Clique Algorithm to Minimize Item Exposure for Uniform Test Forms Assembly. International Conference on Artificial Intelligence in Education (AIED), LNCS 9112, 638–641 (2015)
8. Ueno, M. and Miyazawa, M.: Uniform Adaptive Testing Using Maximum Clique Algorithm. International Conference on Artificial Intelligence in Education (AIED), LNAI 11625, 482–493 (2019)
9. Ishii, T. and Ueno, M.: Algorithm for Uniform Test Assembly Using a Maximum Clique Problem and Integer Programming. International Conference on Artificial Intelligence in Education (AIED), LNAI 10331, 102–112 (2017)
10. Lord, F. and Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley (1968)
11. Ueno, M. and Miyazawa,Y.:Pobability based scaffolding system with fading. Artificial Intelligence in Education (AIED), LNAI 9112, 492–503 (2015)
12. Ueno, M. and Miyazawa,Y.:IRT-based adaptive hints to scaffold learning in programming. IEEE Transactions on Learning Technologies, **11**(4), 415–428 (2018)
13. Tsutsumi,E., Kinoshita,R. and Ueno,M.: Deep item response theory as a novel test theory based on deep learning. Electronics, 10(9), 2021.
14. Tsutsumi,E., Kinoshita,R. and Ueno,M.: Deep-irt with independent student and item networks. In Proceedings of the 14th International Conference on Educational Data Mining (EDM), 2021.