Uniform Adaptive Testing Using Maximum Clique Algorithm

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

Abstract. Computerized adaptive testing (CAT) presents a tradeoff problem between increasing measurement accuracy and decreasing item exposure in an item pool. To address this difficulty, we propose a new CAT that partitions an item pool to numerous uniform item groups using a maximum clique algorithm and then selects the optimum item with the highest Fischer information from a uniform item group. Numerical experiments underscore the effectiveness of the proposed method.

Keywords: Computerized adaptive testing · e-testing · Item response theory · Maximum clique algorithm · Uniform test form assembly

1 Introduction

Computerized Adaptive Testing (CAT) selects and presents the optimal item that maximizes the test information (Fisher information measure) at the current estimated ability based on item response theory (IRT) from an item pool. After each response, the examinee's ability estimate is updated. Then the subsequent item is selected to have optimal properties at the new estimate. Adaptive item selection to each examinee can reduce the number of examined items so as not to decrease the test accuracy in comparison with the same fixed test. However, in conventional CATs, the same items tend to be presented to examinees who have similar ability. This property causes bias of the item exposure frequency in an item pool. Earlier studies [1] demonstrated that frequently exposed items deteriorate rapidly. To resolve this difficulty, Kingsbury and Zara (1989) proposed partitioning of an item pool into several groups of items and then selected the optimal item that maximizes Fisher information from the group minimizing item exposure [1]. Furthermore, van der Linden and et al. (1998,2004,2016) proposed a shadow-test approach that maximizes Fisher information under several constraints (e.g., test area and item exposure frequency) using integer programming [2–4]. Earlier methods mitigated the bias of item exposure frequency from an item pool. However, they encountered the difficulty that increases bias of measurement accuracy for examinees. In addition, this problem necessarily engenders a bias of examinees' required test lengths in CAT. Thus, a tradeoff exists between decreasing

S. Isotani et al. (Eds.): AIED 2019, LNAI 11625, pp. 482–493, 2019.
https://doi.org/10.1007/978-3-030-23204-7_40

item exposure and increasing measurement accuracy. Nevertheless, earlier methods do not address the tradeoff. To resolve that shortcoming, we propose a new framework that can control the balance between item exposure and measurement accuracy. More specifically, we use a state-of-the-art uniform test assembly technique to divide an item pool into several equivalent groups of items and thereby adjust the degree of item exposure. 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). Recent studies explored several techniques using AI technologies to generate numerous uniform test forms from an item pool [5–9]. Especially, among all methods, uniform test assembly using the maximum clique algorithm is known to generate the greatest number of uniform test forms [6–9]. This method formalized the uniform test assembly with overlapping items conditions as a maximum clique problem (MCP), where overlapping items represent common items among multiple test forms. Here it is noteworthy that the determined number of overlapping items increases the item exposure frequency. Determination of the number of overlapping items for the MCP method can therefore control the degree of item exposure.

The MCP has never been utilized for CATs. Therefore, this study proposes a new CAT method which reduces the degree of item exposure using the MCP. The proposed method partitions an item pool into numerous uniform item groups using the MCP method. Then, from a uniform item group, we select the optimum item with the highest Fischer information, which reflects the measurement accuracy.

Salient benefits of using the method are the following.

1. The proposed method solves the tradeoff between item exposure and the measurement accuracy (test length).
2. The proposed method decreases the bias of measurement accuracy (test length) for examinees, that's a uniform adaptive testing.

Experiments were conducted to compare the performances of the proposed method with conventional methods. The results show that the proposed method dynamically improves the measurement accuracy (reduces test length) without largely increasing item exposure. A particularly surprising finding is that increasing the item size of the uniform item group does not necessarily improve the measurement accuracy. Rather, an optimum item size exists for accuracy. This is the main reason why the proposed partition improves the measurement accuracy (reduces test length) without greatly increasing item exposure.

2 Computerized Adaptive Testing Based on Item Response Theory

2.1 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, and 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.

2.2 Fisher Information

The asymptotic variance of estimated ability based on the item response theory is known to approach the inverse of Fisher information [10]. 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.

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

where

$$p'(u_i = 1|\theta) = \frac{\partial}{\partial\theta} p(u_i = 1|\theta).$$

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 $\mathrm{SE}_T(\theta) = \frac{1}{\sqrt{\mathrm{I}_T(\theta)}}$.

2.3 Computerized Adaptive Testing

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 given ability is selected from the item pool. It is presented to the examinee.
3. The examinee's ability estimate is updated from the correct/wrong response data to the item.
4. Procedures 2 and 3 are repeated until the update difference of the examinee's ability estimate decreases to a constant value ϵ or less.

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

2.4 Constrained CAT with Item Exposure Control

In 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 the standard normal distribution, items with higher information around $\theta = 0$ tend to be exposed frequently. Therefore, bias of item exposure frequency occurs in an item pool. An earlier report [1] described that frequently exposed items tend to deteriorate rapidly.

To resolve this difficulty, Kingsbury and Zara (1989) proposed the partitioning of an item pool into several groups of items and then selected the optimal item maximizing Fisher information from the group minimizing item exposure [1,2,4]. The item-pool partitioning procedure is the following.

1. An item pool is partitioned into several groups of items.
2. The estimated ability of an examinee is initialized to $\hat{\theta} = 0$.
3. The group minimizing the number of exposure items is selected from an item pool.
4. The item maximizing Fischer information is selected from the group and is presented to the examinee.
5 After each response, the examinee's ability estimate is updated.
6. Procedures 2, 3, and 4 are repeated until the update difference of the estimated ability decreases to ϵ or less. ϵ is set to 0.01, which is used conventionally for actual computerized adaptive testing.

The number of groups was ascertained by comparing the respective performances of several numbers of groups.

Actually, van der Linden and et al. (1998, 2004, 2016) proposed a shadow-test approach that maximizes Fisher information under several constraints (e.g., test area and item exposure frequency) using integer programming [2–4]. The procedure used for constrained computerized adaptive testing is the following.

1. The estimated ability of an examinee is initialized to $\hat{\theta} = 0$.
2. The item set (shadow test with I items) maximizing Fischer information is then assembled using the integer programming shown below.

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta) x_i \tag{3}$$

subject to

$$\sum_{i=1}^{I} x_i = n; (\text{test length}),$$

$x_i = 1$ if item i is included in the shadow test, $x_i = 0$ otherwise.

If the exposure count of item i is greater than R, then $x_i = 0$,

where R text is the upper bound of the exposure count by the user.

3. The item maximizing Fischer information is selected from the shadow test and is presented to an examinee.
4. After each response, the examinee's ability estimate is updated.
5. Procedures 2, 3, and 4 are repeated until the update difference of the estimated ability decreases to $\epsilon = 0.01$ or less.

Earlier methods mitigated the bias of item exposure frequency in an item pool. However, they led to the important difficulty of increased bias of measurement accuracies (errors) for examinees. Furthermore, this difficulty necessarily engenders a bias of examinees' required test lengths in CAT. In fact, a tradeoff exists between minimizing item exposure and maximizing the measurement accuracy (test information). Nevertheless, earlier methods do not adjust the tradeoff. For that reason, we propose a new CAT framework that can control the tradeoff.

3 Uniform Adaptive Testing Using Maximum Clique Algorithm

The proposed method partitions an item pool to numerous equivalent groups of items to adjust the degree of item exposure using the MCP method. The method then selects the optimum item with the highest Fischer information from a uniform partition of the item pool.

3.1 Uniform Partitioning of the Item Pool

To maximize the number of uniform tests with an overlap condition, Ishii et al. proposed the maximum clique problem for uniform test assembly [7]. The clique problem is a combinational optimization problem in graph theory. We apply this method to uniform partitioning of the item pool as described below.

Letting V be a finite set of vertexes, and letting E be a set of edges, the graph is represented as a pair $G = \{V, E\}$. The maximum clique problem seeks the clique which has the maximum number of vertexes in the given graph. Letting $G = \{V, E\}$ be a finite graph, and letting $C \subseteq V$ be a clique, then the maximum clique problem is formally defined as shown below:

$$\begin{aligned} &\textbf{maximize } |C| \\ &\textbf{subject to} \\ &\qquad \forall v, \forall w \in C, \{v, w\} \in E \\ &\qquad \text{(clique constraint)}. \end{aligned} \tag{4}$$

Here, uniform partitioning of item pool has the following specifications:

1. Any item group in the uniform partition satisfies all partition constraints.
2. Any two item groups in a uniform partition comprise a different set of items (i.e., any two groups have fewer overlapping items than the number allowed in the overlapping constraint).

Accordingly, uniform partitioning of the item pool can be described as the maximum clique extraction from a graph:

$$V = \left\{ \begin{array}{l} s : s \in S, \text{Feasible item-group } s \\ \quad \text{satisfies all constraints} \\ \quad \text{excepting the overlapping} \\ \quad \text{constraint from a given} \\ \quad \text{item pool} \end{array} \right\}$$

$$E = \left\{ \begin{array}{c} \{s', s''\} : \text{The pair of } s' \text{ and } s'' \\ \text{satisfies the} \\ \text{overlapping constraint} \end{array} \right\}.$$

The test constraints include a constraint for the number of items, and the test information of the item group. Letting L_{θ_k} be a lower bound, and letting U_{θ_k} be an upper bound for test information related to $\mathrm{I}_T(\theta_k)$, then a constraint for test information function is written as the following equation.

$$L_{\theta_k} \leq \mathrm{I}_T(\theta_k) \leq U_{\theta_k} \tag{5}$$

Letting OC be the allowed number in the overlapping constraint and letting both s and s' be item groups, then the overlapping constraint is defined as the following equation:

$$\forall s, \forall s' \in V, \tag{6}$$

$$|s \cap s'| \leq \mathrm{OC} \tag{7}$$

This maximum clique problem seeks the maximum set of feasible item groups in which any two groups satisfy the overlapping constraint. Therefore, this optimization problem theoretically maximizes the number of equivalent item groups. We apply the approximated MCP algorithm [9], which is a state-of-the-art algorithm, to obtain numerous uniform item groups.

3.2 Adaptive Item Selection from an Item Group

Using the obtained item groups, the proposed method selects and presents the optimal items to an examinee as explained 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 presented to an examinee in Procedure 1.
3. The examinee's ability estimate is updated from his/her response.
4. Procedures 2 and 3 are repeated until the update difference of the estimated ability of the examinee reaches a constant value of $\epsilon = 0.01$ 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.

4 Numerical Evaluation

This section presents a comparison of the performance of the proposed method (designated as Proposal) to those of other computerized adaptive testing methods (conventional adaptive testing in 2.3 (designated as CAT), Kingsbury and Zara (1989) CAT in 2.4 (designated as KZ), and van der Linden's IP based CAT in 2.4 (designated as IP). For the proposed method and KZ, we construct item groups with 50 items (We write Proposal(50) and KZ(50)) and item groups with 100 items (We write Proposal(100) and KZ(100)) to investigate the effects of item sizes for the measurement accuracy and item exposure. Furthermore, we use two numbers of overlapping items for the proposed methods OC = 0 and OC = 10 to investigate the effects of OC for the measurement accuracy and item exposure. Additionally, we conduct experiments with $R = 50, 80, 90, 100,$ and 150 as the upper bound exposure counts of IP. Also, L_{θ_k} and U_{θ_k} of the proposed method are determined as described in an earlier report [7]. We conducted the following two experiments using simulation data and actual data.

4.1 Simulation Experiment

We conducted a simulation experiment as described hereinafter.

1. Item pools with 500 and 1000 items are generated. The true parameters of each item are generated from $a_i \sim U(0, 1)$ and $b_i \sim N(0, 1)$.
2. The true abilities of examinees are sampled from $\theta \sim N(0, 1)$.
3. Each adaptive testing method is conducted using each item pool. The examinees' response data are generated from $p(u_i \mid \widehat{\theta})$. Correct response data are generated if $p(u_i \mid \widehat{\theta}) > 0.5$. Incorrect response data are generated otherwise. The convergence (Stopping) criterion is $\epsilon = 0.01$, which is used conventionally for actual computerized adaptive testing [14].
4. Procedures 2–3 are repeated 1000 times.

Table 1 presents the results. In Table 1, "overlapping items" represents the number of overlaps, "No. Item groups" denotes the number of generated (uniform for the proposal) item groups, "Avg. test length" stands for the average test length which reflects the measurement accuracy of the test (the standard error of test lengths in parenthesis), and "S.D. estimates error" expresses the standard deviation of asymptotic errors of estimates $\widehat{\theta}$. When the value of "S.D. estimates error" approaches to zero, the tests presented to the different examinees by the CAT have the same estimation accuracy.

Also, "Max. No. exposure item" shows the maximum number of exposure items. "Avg. exposure item" expresses the average exposure count of an item (the standard error of numbers of exposure items in parenthesis). For IP results, Table 1 shows only those results of R with the best accuracy (the smallest value of "Avg. test length") because of limitations of space. Consequently, $R = 150$ for 500 item pool size and $R = 80$ for 1000 item pool were selected.

Results obtained for "No. item groups" demonstrate that the proposed method generated numerous uniform item groups. "Avg. test length", which

Table 1. Results using simulated data

Item pool size	Methods	Over lapping items	No. item-groups	Avg. test length	S.D. estimates error	Max.No. exposure item	Avg. exposure item
500	AT	-	-	65.1 (7.29)	0.022	1000	130.1 (223.5)
	IP	-	-	73.2 (16.07)	0.588	150	146.4 (19.4)
	KZ(50)	0	10	57.6 (9.40)	0.031	649	115.2 (182.5)
	KZ(100)	0	5	62.2 (8.51)	0.029	751	124.4 (206.8)
	Proposal(50)	0	5	33.9 (7.82)	0.03	200	67.9 (84.7)
		10	136	34.3 (8.28)	0.032	227	68.6 (40.1)
	Proposal(100)	0	3	39.5 (7.35)	0.047	334	79.1 (123.1)
		10	99983	40.2 (6.86)	0.017	326	80.3 (78.5)
1000	AT	-	-	70.4 (7.11)	0.022	1000	70.4 (158.1)
	IP	-	-	79.5 (26.93)	0.48	80	79.5 (6.2)
	KZ(50)	0	20	62.4 (9.94)	0.034	450	62.4 (112.5)
	KZ(100)	0	10	66.1 (9.28)	0.031	593	66.1 (133.4)
	Proposal(50)	0	9	31.9 (7.13)	0.038	112	31.9 (46.3)
		10	8758	35.1 (8.19)	0.025	154	35.1 (21.8)
	Proposal(100)	0	4	39.4 (9.44)	0.063	200	39.4 (70.5)
		10	100000	42.5 (8.05)	0.015	227	42.9 (41.9)

Table 2. Result using actual data

Item pool size	Methods	Over lapping items	No. item-groups	Avg. test length	S.D. estimates error	Max.No. exposure item	Avg. exposure item
978	AT	-	-	65.5 (10.72)	0.016	1000	67 (163.2)
	IP	-	-	87.9 (18.22)	0.323	90	89.9 (1.2)
	KZ(50)	0	20	63.5 (12.30)	0.042	382	45.1 (88.3)
	KZ(100)	0	10	61.7 (13.11)	0.044	556	44.66 (108.6)
	Proposal(50)	0	7	41.2 (4.51)	0.043	143	42.1 (60.9)
		10	8669	40.7 (4.70)	0.043	136	41.6 (19.3)
	Proposal(100)	0	2	54.8 (8.84)	0.033	500	56 (139.3)
		10	7088	54.3 (7.67)	0.031	179	55.5 (42.6)

reflects the measurement accuracy of the corresponding CAT, is one of the most important indexes in this study because reducing the number of examined items so as not to increase item exposure solves the tradeoff between item exposure and the measurement accuracy. The results of "Avg. test length," which reflects the measurement accuracy of CAT, surprisingly show that the proposed method provides the best performance among all the CATs, although the item alternatives were constrained using the uniform item groups. Presenting items with extremely high information at the early stage of CAT is known to adversely cause the local solution for ability estimation and to interrupt the convergence to the true estimate because the estimate at the early stage is often far from the true one [2]. The proposed method has a uniform distribution of item characteristics over the whole ability area. It constrains the number of items with the uniform conditions so as not to select items with extremely high information for a specific ability area. This property shortens the test length. It is noteworthy that it mitigates item exposure because the proposed method presents less non-informative

items to examinees. Additionally, it is notable that it yields the smallest standard deviation of test length. Therefore, the proposed method decreases the bias of measurement accuracy for examinees. No significant differences were found in the values of "S.D. estimates error" among the methods because all methods employ the same convergence criterion $\epsilon = 0.01$. Although the proposed method did not provide the lowest values of "Max. No. exposure item," the best values of IP result from the constraint of the upper bound R. The proposed method shows the lowest values of "Avg. exposure items" among all methods. Comparing the performances of the uniform item group sizes 50 and 100 for the proposed method and KZ, the performances with item size 50 are better than those with item size 100. The reason is that even the uniform item group is affected by the local solution problems when the item size is large. This result emphasizes that partitioning an item pool is effective for CAT. This result also suggests that there might be an optimal size of the uniform item group. In addition, comparing the performances of overlapping item sizes $\mathrm{OC} = 0$ and $\mathrm{OC} = 10$ for the proposed method, the number of generated uniform item groups with $\mathrm{OC} = 50$ is much larger than that with $\mathrm{OC} = 0$. However, contrary to expectations, performances with $\mathrm{OC} = 0$ outperform those with $\mathrm{OC} = 10$ for all criteria except for "No. item groups." The uniform item groups generated with $\mathrm{OC} = 0$ have higher quality for uniform measurement accuracy and item exposure than those with $\mathrm{OC} = 10$, although CAT must repeatedly use the same uniform item groups with $\mathrm{OC} = 0$. Therefore, a tradeoff must exist between the quality of uniform item groups and the number of generated uniform item groups. Moreover, as the item pool size increases, the performances of CATs do not necessarily increase. In fact, the proposed method increases the performances as the item pool size increases only when the item group size is 50 because the item-groups can gather high quality items when the item size is large. From these results, it is recommended to develop a large size item pool and then assemble uniform item-groups with an optimum item size.

Conventional CAT shows the lowest values of "Avg. test length" and "S.D. estimates error" among all CATs because it repeatedly selects and presents the same items to different examinees. In addition, the values of "Max. No exposure items" and "Max. No. exposure item" of Conventional CAT have the worst results among all methods. Particularly "Max. No exposure items" demonstrates that conventional CAT presented the same item to all examinees.

Although IP decreased the values of "Max. No. exposure item" because of the upper bound R, it did not provide good performance for other criteria. However, a naive method, KZ, provided better performance than IP did, except for "Max. No. exposure item." The reason is that partitioning the item pool to several groups is highly effective for the same reason as that for the proposed method, although it uses naive random sampling.

The results demonstrate that partitioning an item pool to several groups with an appropriate number of items is highly effective for CAT.

4.2 Experiment Conducted Using Actual Data

This section presents evaluation of the effectiveness of the proposed method using actual data. An experiment was conducted using the item pool of real data, with 978 items, and a test constraint used in the synthetic personality inventory (SPI) examination (actual CAT style), which is a popular aptitude test in Japan [15].

Table 2 presents the results. For IP results, Table 2 presents only the results of $R = 90$ with the best accuracy (the smallest value of "Avg. test length") because of space limitations. The table shows almost identical results to those of the simulation experiment. Namely, the proposed method provides the best performances among all the methods. The generated uniform item groups with $OC = 10$ are much more numerous than those with $OC = 0$. In this case, performances with $OC = 10$ slightly outperform those with $OC = 0$ for all criteria. As described in Sect. 4.1, a tradeoff must exist between the quality of uniform item groups and the number of generated uniform item groups. In this experiment, the proposed method provides the best performance for item size of 50 and $OC = 10$.

5 Conclusions

This paper has demonstrated that CAT has a tradeoff problem between increasing measurement accuracy and decreasing item exposure in an item pool. To address this difficulty, we proposed a new CAT that partitions an item pool to numerous uniform item groups using a uniform test assembly based on the maximum clique algorithm. Then we select the optimum item with the highest Fischer information from a uniform item group.

Experiments were conducted to compare the performance of the proposed method with those of conventional methods. Results show that the proposed method dynamically improves the measurement accuracy (reduces test length) without greatly increasing item exposure. Contrary to our expectations, the results did not show that the proposed method using numerous uniform item groups necessarily outperforms that using a few groups. Results suggest that a tradeoff for the proposed method must exist between the quality of uniform item groups and the number of generated uniform item groups. We expect to investigate the means of solving this tradeoff problem as a subject of future work.

For this study, we used Fischer information measure as an item selection criterion. Although the Fischer information measure becomes accurate for late stage of CAT because it is an asymptotic approximation, recent studies have proposed more accurate information measures and the item selection algorithms [16–18]. We expect to apply the proposed uniform partition of item pool technique to the information measure and its applications [19,20] in future studies.

References

1. Kingsbury, G.G., Zara, A.R.: Procedures for selecting items for computerized adaptive tests. Appl. Measur. Educ. **2**(4), 359–375 (1989)
2. van der Linden, W.J., Reese, L.: A model for optimal constrained adaptive testing. Appl. Psychol. Measur. **22**(3), 259–270 (1998)
3. van der Linden, W.J., Veldkamp, B.: Constraining item exposure in computerized adaptive testing with shadow tests. J. Educ. Behav. Stat. **29**(3), 273–291 (2004)
4. Choi, S.W., Moellering, K.T., Li, J., van der Linden, W.J.: Optimal reassembly of shadow tests in CAT. Appl. Psychol. Measur. **40**(7), 469–485 (2016)
5. Songmuang, P., Ueno, M.: Bees algorithm for construction of multiple test forms in e-testing. IEEE Trans. Learn. Technol. **4**(3), 209–221 (2011)
6. Ishii, T., Songmuang, P., Ueno, M.: Maximum clique algorithm for uniform test forms assembly. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 451–462. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39112-5_46
7. Ishii, T., Songmuang, P., Ueno, M.: Maximum clique algorithm and its approximation for uniform test form assembly. IEEE Trans. Learn. Technol. **7**(1), 83–95 (2014)
8. Ishii, T., Ueno, M.: Clique algorithm to minimize item exposure for uniform test forms assembly. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 638–641. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_80
9. Ishii, T., Ueno, M.: Algorithm for uniform test assembly using a maximum clique problem and integer programming. In: André, E., Baker, R., Hu, X., Rodrigo, M.M.T., du Boulay, B. (eds.) AIED 2017. LNCS (LNAI), vol. 10331, pp. 102–112. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61425-0_9
10. Lord, F., Novick, M.R.: Statistical Theories of Mental Test Scores. Addison-Wesley, M.R. (1968)
11. Sun, K.T., Chen, Y.J., Tsai, S.Y., Cheng, C.F.: Creating IRT-based parallel test forms using the genetic algorithm method. Appl. Measur. Educ. **21**(2), 141–161 (2008)
12. Belov, D.I., Armstrong, R.D.: A constraint programming approach to extract the maximum number of non-overlapping test forms. Comput. Optim. Appl. **33**(2), 319–332 (2006)
13. van der Linden, W.J.: Linear Models for Optimal Test Design. Springer, New York (2005). https://doi.org/10.1007/0-387-29054-0
14. van der Linden, W.J., Glas, C.A.W.: Elements of Adaptive Testing. Springer, New York (2010). https://doi.org/10.1007/978-0-387-85461-8
15. Recruit, Synthetic Personality Inventory (SPI) (2014). http://www.spi.recruit.co.jp/
16. Ueno, M.: An extension of the IRT to a network model. Behaviormetrika **29**(1), 59–79 (2002)
17. Ueno, M. and Songmuang, P.: Computerized adaptive testing based on decision tree. In: The Tenth IEEE International Conference on Advanced Learning Technologies (ICALT), pp. 191–193 (2010)
18. Ueno, M.: Adaptive testing based on bayesian decision theory. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 712–716. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39112-5_95

19. Ueno, M., Miyasawa, Y.: Probability based scaffolding system with fading. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS (LNAI), vol. 9112, pp. 492–503. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19773-9_49
20. Ueno, M., Miyazawa, Y.: IRT-based adaptive hints to scaffold learning in programming. IEEE Trans. Learn. Technol. **11**(4), 415–428 (2018)