IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Computerized Adaptive Testing to Balance Exposure Bias and Measurement Accuracy using Zero-suppressed Binary Decision Diagrams

MAOMI UENO[1], (Member, IEEE) KAZUMA FUCHIMOTO[1], WAKABA KISHIDA[1], and YOSHIMITSU MIYAZAWA[2]

[1]Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan (e-mail: ueno@ai.lab.uec.ac.jp, fuchimoto@ai.lab.uec.ac.jp).
[2]Research and Development, The National Center for University Entrance Examinations

Corresponding author: Maomi Ueno (e-mail: ueno@ai.lab.uec.ac.jp)

This work was supported by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research: Nos. JP24H00739 and 24KJ1124.

ABSTRACT Computerized adaptive testing (CAT) presents a tradeoff dilemma involving item exposure bias and measurement accuracy. To resolve this dilemma, we propose a new two-step CAT mechanism to balance exposure bias and measurement accuracy. Using zero-suppressed binary decision diagram (ZDD), the proposed method first selects and presents an optimal item from an equivalent item pool, which uniformly divides the whole item pool. The first step rapidly provides a roughly approximated ability estimate of an examinee. The second step produces a more accurate ability estimate of the examinee. Specifically, the second step selects an optimal item from the whole item pool and presents an item with a difficulty parameter value that is approximately equal to the examinee ability estimate. Numerical experiment results underscore the effectiveness of the proposed method.

INDEX TERMS Computerized adaptive testing, item response theory, parallel test assembly, zero-suppressed binary decision diagrams

I. INTRODUCTION

Computerized Adaptive Testing (CAT) (e.g. [1]) selects and presents an optimal item from an item pool to maximize the information for an examinee's current ability as estimated based on item response theory (IRT). After the examinee responds to the item, the examinee's ability is estimated according to the history of response data. Subsequently, the next item is selected to have the maximum information at the current ability estimate. Adaptive item selection for each examinee can reduce the number of presented items without decreasing the measurement accuracy of the examinee's ability compared to a fixed (non-adaptive) test. In fact, CAT has already been applied for many high-stakes assessments such as the National Assessment of Educational Progress (NAEP) [2], Trends in International Mathematics and Science Study (TIMSS) [3], Progress in International Reading Literacy Study (PIRLS) [4], the Program for the International Assessment of Adult Competencies (PIAAC) [5], in addition to others.

However, conventional CAT tends to present the same items to examinees with similar abilities. Therefore, it is inadequate for situations in which the same examinee takes a test multiple times. Additionally, items with greater information around $\theta = 0$ tend to be exposed frequently because IRT requires the assumption that the ability variable follows a standard normal distribution. It entails bias of item exposure frequency in an item pool. This bias consequently decreases the test reliability because the contents of overexposed items might be known to future examinees with high probability [6]–[8].

To resolve this difficulty, many researchers have proposed alternative CATs of various kinds specifically to alleviate the bias associated with item exposure (e.g. [7], [9]–[11]). Van der Linden [7] proposed a CAT that selects the optimal item from a subset of an item pool, which is designated as a "shadow test." The shadow test is assembled from an item pool using integer programming (IP) to satisfy test constraints. Choi and Lim [9] developed another mechanism to minimize the difference between test information of a shadow test and target information (TI). Using a probabilistic approach, van der Linden and Choi [11] proposed a CAT (designated as Prob) method that controls the item exposure using probabilistic item selection. As the most recent approach, Lim and Choi described a hybrid item exposure control method [12] using the combination of an a-stratification method [13], [14] and an eligibility probabilistic method [11]. Nevertheless, these methods led to the difficulty that they increase the bias of measurement accuracies among examinees. In addition, this difficulty necessarily engenders bias of the required test lengths for CAT examinees.

To overcome these difficulties posed by earlier methods, we propose a new two-step CAT algorithm. The first step of the proposed method, using zero suppressed binary decision diagram (ZDD), selects and presents an optimal item from an equivalent item pool which uniformly divides a whole item pool. The equivalent item pools are constructed before running CAT by uniformly dividing the whole item pool so that each has equivalent measurement accuracy but with a different set of items. For this study, we use a state-of-the-art parallel test assembly technique to divide an item pool into equivalent item pools. The parallel test forms have the same test properties (number of test items, test area, test information, etc.), but each form consists of different test items. Recent studies have explored several techniques using AI technologies to generate numerous parallel test forms from an item pool [15]–[20]. Especially among all methods, parallel test assembly using a zero-suppressed binary decision diagram (ZDD) [21] is known to generate the greatest number of parallel test forms [20]. However, when we apply the ZDD directly to generate equivalent item pools, the ZDD method often leads to computer memory overflow. To resolve this shortcoming, this study proposes a novel ZDD compilation algorithm implemented to address this particular difficulty. (1) A ZDD is constructed with approximated measurement accuracies of shared nodes. Specifically, during the breadth-first search, nodes are shared when the difference in measurement accuracies between two nodes at the same depth is smaller than a determined threshold parameter value. Then, measurement accuracy of the shared node uses an approximated value by averaging the two nodes' measurement accuracies. (2) Paths from the constructed ZDD are randomly sampled to search and enumerate paths that exactly satisfy the constraints of measurement accuracies.

Consequently, each equivalent item pool includes a different set of items, but each has equivalent measurement accuracy. Because the first step selects the optimal item from an equivalent item pool, it presents a different set of items to each examinee until the examinee's ability estimate converges. The first step rapidly provides a roughly approximated ability estimate of an examinee because item difficulties in each equivalent item pool are distributed uniformly and sparsely. The second step reaches a more accurate ability estimate of the examinee. Specifically, after the examinee's ability estimate converges, the method switches to select the optimal item from the whole item pool in the second step. For this study, because the Fisher information measure is an asymptotic approximation, we use it as an item selection criterion that is sufficiently accurate for the second step.

However, improvement of the bias of the item exposure might be inadequate because the second step exhibits a tendency to select and present items with high Fisher information (examinee's measurement accuracy) for widely various examinee abilities rather than with a difficulty parameter that is approximately equal to the current ability estimate. To relax this tendency, the second step of the proposed method selects an optimal item from items which satisfy the item difficulty interval (IDI) condition based on the standard error of the estimated ability. Therefore, the second step selects an optimal item with a difficulty parameter value that is approximately equal to the examinee ability estimate. Similar techniques to the IDI condition have been proposed for multi-objective optimization problem studies (e.g., [22], [23]).

Findings obtained from numerical experiments demonstrate that the proposed method can mitigate bias of item exposure while maintaining low measurement error.

II. CONVENTIONAL CAT

A. ITEM RESPONSE THEORY

To select the optimum item with the highest Fisher information, conventional CAT estimates an examinee's ability based on Item Response Theory (IRT) [24]. For the three-parameter logistic model (3PLM) [24], the most widely recognized IRT model, the probability of a correct answer to item $i(=1,2,\ldots,n)$ by examinee $j(=1,2,\ldots,J)$ with ability $\theta_j \in (-\infty,\infty)$ is assumed as

$$p(u_i=1|\theta_j)=c_i+\frac{1-c_i}{1+\exp(-Da_i(\theta_j-b_i))}. \quad (1)$$

Therein, u_i takes a value of 1 when an examinee answers item i correctly. It is 0 otherwise. In addition, $a_i \in [0,\infty)$, $b_i \in (\infty,\infty)$, and $c_i \in [0,1]$ respectively represent the discrimination parameter of item i, the difficulty parameter of item i, and the guessing parameter of item i. Furthermore, D is a scale factor used to approximate the cumulative distribution function of the standard normal distribution to 3PLM. Actually, the widely used scale factor is $D=1.701$. Especially when $c_i=0$, the model in Eq. (1) is designated as a two-parameter logistic model (2PLM).

In addition, a widely used IRT model is the generalized partial credit model (GPCM) [25]. With GPCM, the probability of receiving a score on item i in ordered category $k(=0,1,\ldots,K-1)$ is defined as

$$p_i(k|\theta_j)=\frac{\exp(\sum_{s=1}^{k}\alpha_i(\theta_j-\beta_{i,s}))}{\sum_{k=0}^{K-1}[\exp(\sum_{s=1}^{k}\alpha_i(\theta_j-\beta_{i,s}))]}, \quad (2)$$

where $K-1$ represents the number of response categories and where $\beta_{i,s} \in (\infty,\infty)$ is the step difficulty parameter for receiving a score on item i in category s. It is noteworthy that $\beta_{i,0}$ is 0.

The ability Expected A Posteriori (EAP) estimate $\hat{\theta}$ [26] is calculated as

$$\hat{\theta}=\int_{-\infty}^{\infty}\theta f(\theta|\mathbf{u})d\theta, \quad (3)$$

where

$$f(\theta|\mathbf{u}) = \frac{L(\mathbf{u}|\theta)f(\theta)}{\int_{-\infty}^{\infty} L(\mathbf{u}|\theta)f(\theta)d\theta}.$$

Therein, $L(\mathbf{u}|\theta)$ denotes a likelihood function of θ and $f(\theta)$ represents a prior distribution of θ, the standard normal distribution $N(0, 1^2)$.

The asymptotic variance of an examinee's ability estimate for IRT is known to approach the inverse of Fisher information [27]. The Fisher information function for 3PLM [27] is defined when item i provides an examinee's ability θ from the following equations:

$$I_i(\theta) = \frac{[p'(u_i = 1|\theta)]^2}{p(u_i = 1|\theta)[1 - p(u_i = 1|\theta)]}, \tag{4}$$

where

$$p'(u_i = 1|\theta) = \frac{\partial}{\partial\theta} p(u_i = 1|\theta). \tag{5}$$

In addition, the Fisher information function for GPCM [25] is defined when item i provides an examinee's ability θ using the following equations as

$$I_i(\theta) = \sum_{k=0}^{K} \frac{[p_i'(k|\theta)]^2}{p_i(k|\theta)}, \tag{6}$$

where

$$p_i'(k|\theta) = \frac{\partial}{\partial\theta} p_i(k|\theta). \tag{7}$$

An item with high Fisher information $I_i(\theta)$ can estimate the examinee's ability accurately. Therefore, conventional CAT usually implements an item selection method with the highest amount of Fisher information given an examinee's ability estimate $\hat{\theta}$.

The test information $I_T(\theta)$ of a test form T is defined as $I_T(\theta) = \sum_{i \in T} I_i(\theta)$. As a result, the asymptotic error of ability estimate $\hat{\theta}$ is obtained as the inverse of square root of the test information function at a given ability estimate $\hat{\theta}$.

B. ALGORITHM OF CONVENTIONAL CAT

Conventional CAT selects optimal items from an item pool, as described hereinafter.

1. Procedure 1 initializes an examinee's ability estimate to $\hat{\theta} = 0$.
2. Procedure 2 selects an item with the highest Fisher information for the examinee's ability estimate from an item pool and presents the item.
3. Procedure 3 updates the examinee's ability estimate from the examinee's response history data using Eq. (3).
4. Procedures 2 and 3 are repeated until the ability estimate converges.

Consequently, CAT sequentially selects and presents an optimal item to an examinee's ability estimate. As a result, CAT can reduce the test length without reducing the measurement accuracy of the examinee's ability compared to the fixed test.

III. CAT METHODS WITH ITEM EXPOSURE CONTROL

As described earlier, conventional CATs tend to present the same items to examinees who have similar abilities. This property engenders bias of the item exposure frequency in an item pool and consequently decreases the test reliability.

To overcome this difficulty, various CAT methods incorporating item exposure have been proposed.

A. METHOD USING IP

As a well known approach, van der Linden proposed a method selecting the optimal item from a shadow test assembled by solving IP with constraints to control the item exposure [7]. This method selects and then presents the optimal item as described below.

1. Procedure 1 initializes an examinee's estimated ability to $\hat{\theta} = 0$.
2. The shadow test is assembled along with solution of the IP, as shown below.
   maximize
   $$\sum_{i=1}^{n} I_i(\hat{\theta})x_i. \tag{8}$$
   subject to
   $$r_i x_i \quad \leq \quad R(i = 1, 2, \cdots, n), \tag{9}$$
   $$\sum_{i=1}^{n} x_i \quad = \quad L, \tag{10}$$
   $$x_i = \begin{cases} 1, & \text{if item } i \text{ is included} \\ 0, & \text{otherwise} \end{cases}$$
   Therein, r_i denotes the number of times item i is presented. In addition, R expresses the maximum number of times it is presented. L represents the test length.
3. The item which maximizes Fisher information is selected from the shadow test.
4. The current ability estimate is updated based on the examinee's response history.
5. Procedures 2–4 are repeated until the ability estimate converges.

This method can keep the number of times each item is presented under the upper bound R. However, this method avoids presenting only items that reach the maximum number of times to present. Therefore, the method has limited effectiveness at mitigating bias of item exposure.

B. SHADOW-TEST APPROACH WITH PROBABILISTIC ELIGIBILITY (PROB)

Van der Linden and Choi proposed a method of controlling item exposure probabilistically [11]. More specifically, this method selects the optimal item and presents it as described below.

1. Procedure 1 initializes an examinee's estimated ability to $\hat{\theta} = 0$.

2. Then the eligibility probability of item i for examinee j is calculated as

$$EP^{(i,j)} = \min\{\frac{r^{max}}{IER_{i,(j-1)}} EP^{(i,j-1)}, 1\}, \quad (11)$$

$$IER_{i,j} = \frac{1}{j}\sum_{j'=1}^{j} IE_{i,j'}, \quad (12)$$

$$IE_{i,j'} = \begin{cases} 1 \quad \text{if the } i\text{-th item is exposed} \\ \text{to the } j'\text{-th examinee, and} \\ 0 \quad \text{otherwise.} \end{cases} \quad (13)$$

   Therein, $IER_{i,j}$ denotes the item exposure rate when examinees $1, 2, ..., j$ finish their tests, and where r^{max} represents the upper bound of the item exposure rate.
3. Following eligibility probability computation, some items become ineligible.
4. The item which maximizes Fisher information is selected from the item pool without ineligible items.
5. Based on the examinee's earlier response history, the estimated examinee ability $\hat{\theta}$ is updated.
6. Procedures 2–5 are repeated until the estimated examinee ability converges.

This method avoids presenting only items that are above the upper bound of the item exposure rate r^{max}. However, this method does not guarantee that items with low item exposure will be selected for examinees. Therefore, the method can provide only limited effectiveness in decreasing the bias of item exposure.

C. SHADOW-TEST APPROACH USING TARGET INFORMATION (TI)

As an approach that is similar to IP, Choi and Lim proposed a method [9] of selecting an optimal item from a shadow test assembled by solving IP.

1. Procedure 1 initializes an examinee's estimated ability to $\hat{\theta} = 0$.
2. Then the shadow test is assembled by solving the IP presented below:
   minimize y.
   subject to

$$\begin{aligned} \sum_{i=1}^{n} I_i(\hat{\theta})x_i &\leq T + y, \\ \sum_{i-1}^{n} I_i(\hat{\theta})x_i &\geq T - y, \\ y &\geq 0, \\ \sum_{i=1}^{n} x_i &= L, \end{aligned}$$

   where T denotes the target value of test information when examinees $1, 2, \ldots, J-1$ finish tests.
3. The item which maximizes Fisher information is selected from the assembled shadow test.
4. Based on the examinee's earlier response history, the current ability estimate is updated.
5. Procedures 2–5 are repeated until the estimated examinee ability converges.

Conventional CAT selects greedily to maximize information at each item presentation. This item selection tends to cause overexposure of items with high Fisher information during the early step and tends to leave items with low Fisher information during the later step. By contrast, TI can assemble shadow tests directly with a determined target value of test information using the minimax approach, which minimizes the deviation between the test information value of the assembled shadow test and the target value of test information. Therefore, TI might avoid presentation of only those items with high Fisher information from the whole item pool in the early step. Consequently, TI is expected to alleviate the bias of item exposure.

However, because TI cannot control the bias of item exposure directly, TI can provide only limited effectiveness for decreasing the bias of item exposure.

D. SHADOW-TEST APPROACH WITH A-STRATIFICATION AND PROBABILISTIC ELIGIBILITY (HYBRID)

Lim and Choi proposed a hybrid item exposure control method [12] incorporating the a-stratification method [13], [14] and the eligibility probabilistic method (Section. III-B). The a-stratification method divides the item pool into multiple sets of active items based on values of the discrimination parameter according to Chang and van der Linden [14]. Actually, during the early steps of CAT, a-stratification uses a set of active items with the lowest discrimination parameters. By contrast, during the last steps of CAT, a-stratification uses a set of active items with the highest a-parameters. More specifically, the hybrid method selects the optimal item and presents it as described below.

1. Procedure 1 initializes an examinee's estimated ability to $\hat{\theta} = 0$.
2. Procedure 2 calculates the eligibility probability of item i for examinee j as Eq. (11).
3. Procedure 3 determines the set of ineligible items V_{inel} according to eligibility probability $EP^{(i,j)}$.
4. Procedure 4 determines the set of active items V_a based on the a-stratification method [13], [14] with item position p, where p represents the number of presented items within CAT.
5. The item which maximizes Fisher information is selected from the item pool without the set of ineligible items V_{inel} and the set of active items V_a.
6. Based on the examinee's earlier response history, the estimated examinee ability $\hat{\theta}$ is updated.
7. Procedures 2–4 are repeated until the estimated examinee ability converges.

The a-stratification method tends to select the item with the value of the difficulty parameter closest to the estimated examinee's ability because active items have the equivalent discrimination parameter. As a consequence, the a-stratification method is expected to decrease the bias of item exposure. In addition, the eligibility probabilistic method avoids presenting only items that are above the upper bound of the item exposure rate r^{max}. Therefore, the hybrid item exposure control method can decrease the bias of item exposure.

IV. CAT METHOD USING ZDD

Earlier item exposure control methods for CAT have not resolved the tradeoff dilemma between decreasing item exposure and increasing measurement accuracy.

To balance this tradeoff, we propose a new two-step CAT framework using ZDD. The first step divides an item pool into as many equivalent item pools as possible using ZDD, as described by Fuchimoto et al. [20]. That process of division is known to lead to assembly of the greatest number of parallel test forms with the highest measurement accuracy. In fact, ZDD is an efficient graphical representation of a set of item combinations [21]. It can decrease the calculation time and the computer memory which are used.

Subsequently, CAT selects an item from an equivalent item pool assigned to each examinee. After the examinee's ability estimate converges in the first step, it switches to the second step, which selects and presents the optimal item from the whole item pool. As a result, the proposed method can reduce both the test length and the item exposure without decreasing the measurement accuracy of an examinee's ability. However, improvement of the bias of the item exposure is constrained because the second step tends toward frequent selection of items with high Fisher information for widely various examinees' abilities, rather than those with difficulty parameters that are approximately equal to the current estimated examinee ability. To address this shortcoming, the second step of the proposed method selects an optimal item from items that satisfy the item difficulty interval (IDI) condition based on the standard error of the estimated ability. The use of the IDI condition thereby avoids biased selection of items with high Fisher information for widely various examinees' abilities.

Details of the proposed CAT are presented in the following subsections.

A. ZERO-SUPPRESSED BINARY DECISION DIAGRAM (ZDD)

A ZDD is constructed from a binary decision tree (BDT) by the application of two reduction rules that eliminate redundancy. These reduction rules provide the ZDD with the advantages of reducing computation time and memory usage. As a result, the ZDD achieves compactness and efficiency by representing subsets using binary variables, as explained below.

Given a finite set $I = \{x_1, x_2, \ldots, x_n\}$ with ordered binary variables, a family of sets $\mathcal{F} \subseteq 2^I$ exists, where each subset $S \subseteq I$ is a set of binary variables x_i from the finite set I. Each binary variable x_i represents whether $x_i \in S$ or $x_i \notin S$ for each subset S, defined as

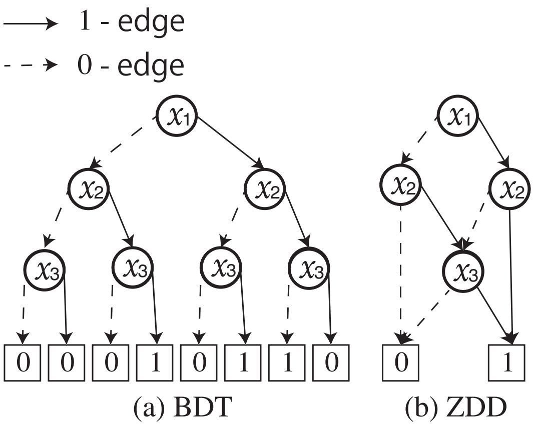


FIGURE 1. BDT and ZDD.

$$x_i = \begin{cases} 1 & \text{if } x_i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, each subset $S \in \mathcal{F}$ can be represented as a unique combination of binary values for each binary variable x_i in the finite set I.

A ZDD is a directed acyclic graph (DAG) that represents the family of sets $\mathcal{F}$ compactly and which has terminal nodes of two types: a 1-terminal node representing valid subsets in $\mathcal{F}$; and a 0-terminal node representing subsets not included in $\mathcal{F}$. Consequently, each path from the root to the 1-terminal node corresponds to a unique subset $S \in \mathcal{F}$.

Figures 1(a) and 1(b) respectively represent examples of a BDT and a ZDD, where the finite set is given as $I = \{x_1, x_2, x_3\}$. In addition, these graph structures have two terminal nodes, which are shown as rectangles in Figure 1: 1-terminal and 0-terminal. A path from the root node to the 1-terminal node in these graph structures corresponds to a unique subset $S \in \mathcal{F}$. Every non-terminal node is presented as a circle in Figure 1. Each non-terminal node is labeled using a binary variable x_i. Moreover, each node has two outgoing edges: a 1-edge and a 0-edge. The 1-edge and the 0-edge respectively signify that the parent node is an element of each subset S, and not. For instance, in Figure 1, the subset $\{x_1, x_3\}$ is represented in both the BDT and the ZDD by following the 1-edge at x_1 (indicating $x_1 \in S$), the 0-edge at x_2 (indicating $x_2 \notin S$), and the 1-edge at x_3 (indicating $x_3 \in S$) before reaching the 1-terminal node. This traversal ensures that x_1 and x_3 are included in the subset S and that x_2 is excluded. Accordingly, in Figure 1, both BDT and ZDD correspond to the same family of sets $\mathcal{F}$, which consists of $\{\{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}\}$. This comparison demonstrates that the ZDD can represent the same family of sets with fewer nodes than BDT can.

The ZDD is obtained by applying the two reduction rules by Minato [21] to the BDT. Specifically, the two reduction rules are defined as presented hereinafter.

1. In reduction rule 1, when two non-terminal nodes represent the identical binary variable x_i and their 1-edge and 0-edge point to nodes that represent identical subtrees, these two nodes are shared into a single node. Reduction rule 1 eliminates duplicate nodes representing the same subtrees, thereby reducing redundancy in the graph structure.
2. In reduction rule 2, nodes with a 1-edge pointing to the 0-terminal node are removed because these nodes are not elements of any valid subset S in the family of sets $\mathcal{F}$. Reduction rule 2 simplifies the graph structure by eliminating redundant nodes that cannot lead to the 1-terminal node.

By applying these two reduction rules, a canonical ZDD representing the family of sets $\mathcal{F}$ is obtained. The canonical ZDD provides a unique and compact representation of the family of sets $\mathcal{F}$, ensuring that redundant nodes and subtrees are fully eliminated.

B. EQUIVALENT ITEM POOL CONSTRUCTION USING ZDD

The proposed method, inspired by the work of Fuchimoto et al. [20], divides an item pool into several equivalent item pools using zero-suppressed binary decision diagram. For the proposed method, we define a finite set $I = \{x_1, x_2, \ldots, x_n\}$ with ordered binary variables, where n represents the number of items in the item pool. Each binary variable x_i is defined as presented below.

$$x_i = \begin{cases} 1, & \text{if the item } i \text{ is included for} \\ & \text{equivalent item pool, and} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Additionally, we define the family of sets $\mathcal{F} \subseteq 2^I$ as the set of equivalent item pools, where each subset $S \in \mathcal{F}$ is defined as a set of binary variables which satisfies the following constraints.

$$\sum_{i=1}^{n} x_i = M, \tag{11}$$

$$\forall \gamma \in \{1, 2, \ldots, \mathit{\Gamma}\}, LB_{\theta_\gamma} \leq \sum_{i=1}^{n} I_i(\theta_\gamma) x_i \leq UB_{\theta_\gamma}. \tag{12}$$

Therein, M represents the number of items in each equivalent item pool. The number of items M is optimized to balance the bias of item exposure and measurement error. In addition, LB_{θ_γ} and UB_{θ_γ} respectively represent a lower bound and an upper bound of the test information at point θ_γ on the ability level. For earlier studies of automated parallel test forms assembly, the upper and lower bounds of test information have been determined arbitrarily based on the desired measurement error (e.g., [19], [20], [28], [29]). By contrast, the present study determines the upper and lower bounds of test information based on characteristics of item information in the item pool. For the proposed method, these bounds are set as presented below.

$$LB_{\theta_\gamma} = I_{\mu,\theta_\gamma} n, \tag{13}$$

$$UB_{\theta_\gamma} = (I_{\mu,\theta_\gamma} + I_{\sigma,\theta_\gamma}) n, \tag{14}$$

$$I_{\mu,\theta_\gamma} = \frac{1}{n} \sum_{i=1}^{n} I_i(\theta_\gamma), \tag{15}$$

$$I_{\sigma,\theta_\gamma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (I_i(\theta_\gamma) - I_{\mu,\theta_\gamma})}. \tag{16}$$

In those equations, I_{μ,θ_γ} and I_{σ,θ_γ} respectively represent the average and the standard deviation of the Fisher information (Eq. 4) at the ability level θ_γ over all items. Eq. (13) and Eq. (14) mean that each equivalent item pool is guaranteed to assemble items with above-average test information.

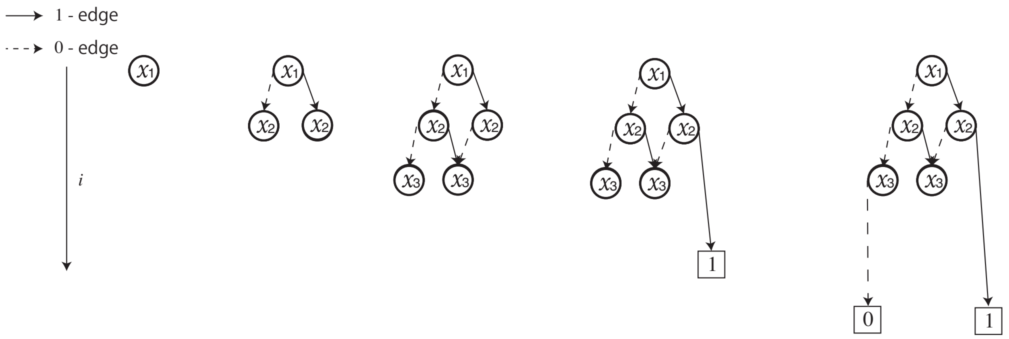
When we apply the conventional ZDD [21] to divide an item pool, the BDD has high space complexity of $\mathcal{O}(2^n)$, where n represents the number of items in the item pool. To mitigate this space complexity, the proposed method implements a breadth-first search [30] to compress the BDT into the ZDD. Frontier-based search constructs a ZDD directly using top-down and breadth-first approaches without increasing the computer memory usage or computation time. The conventional ZDD compilation shares two nodes with identical test information values at all test score levels during the top-down and breadth-first approaches. However, it leads to insufficient sharing of nodes because, from a vast search space, finding nodes with exactly identical measurement accuracies is extremely difficult. Consequently, when we apply the ZDD directly to generate equivalent item pools, this method often causes computer memory overflow.

To solve this problem, this study proposes a novel ZDD compilation algorithm. First, the proposed method constructs a ZDD that approximates the test information value of the merged two nodes by their average value. In the ZDD, during breadth-first search using frontier-based search, nodes are shared with an approximated value by averaging the two nodes' measurement accuracies when the difference in measurement accuracy between two nodes at the same depth is less than a threshold value.

Next, paths that satisfy the measurement accuracy constraints are searched and enumerated from the constructed ZDD. The exact measurement accuracy for each of the enumerated paths is recalculated to enumerate paths that exactly satisfy the measurement accuracy constraints.

1) Approximated ZDD construction

For frontier-based search, we designate a number of item variables as tl and designate a test information array as $tis (tis = [ti_1, ti_2, \ldots, ti_\Gamma])$, where Γ represents the number of discretized points for the test information function. The numbers of the item variable value and each element value in the test information array respectively correspond to $\sum_{i=1}^{n} x_i$ in Eq. (11) and $\sum_{i=1}^{n} I_i(\theta_\gamma) x_i$ in Eq. (12). Frontier-based search calculates the values of these variables for each node.



(1) root vertex (2) child vertices (3) share vertices (4) connect 1-terminal vertex (5) connect 0-terminal vertex

FIGURE 2. Outline of the approximated ZDD construction.

Specifically, 1) Approximated ZDD construction algorithm comprises five procedures, as presented in Figure 2.

1. Procedure 1 creates a root node. Then, Procedure 1 sets zero to the number of items variable *tl* value and zero to each element *tis*$[\gamma]$ value in the test information array.
2. Procedure 2 creates a 0-child node with a 0-edge and a 1-child node with a 1-edge. Then, Procedure 2 adds one to the number of items variable *tl* value for 1-child nodes. Subsequently, Procedure 2 adds Fisher information $I_i(\theta_\gamma)$ of depth i to every element $tis[\gamma]$ value in the test information array for 1-child nodes.
3. Procedure 3 merges two nodes into a single node when the difference of $tis[\gamma]$ values for the two nodes is less than the threshold value I_{th}. Then, each element *tis*$[\gamma]$ value in the test information array for the merged node is approximated by the average of the corresponding values from the two original nodes.
4. Procedure 4 connects a 1-edge to the 1-terminal node when the number of items variable *tl* value and each element *tis*$[\gamma]$ value in the test information array satisfy the following constraints, which correspond to Eq. (11) and Eq. (12) as
   Condition 1. $M = tl$,
   Condition 2. $\forall\gamma \in \{1, 2, \ldots, \Gamma\}, I_{\mathrm{LB}}(\theta_\gamma) \leq tis[\gamma] \leq I_{\mathrm{UB}}(\theta_\gamma)$.
5. In Procedure 5, a 1-edge and a 0-edge are connected to the 0-terminal node when one of the following constraints is satisfied because the test constraints are not satisfied.
   Condition 1. $M < tl$,
   Condition 2. $\exists\gamma \in \{1, 2, \ldots, \Gamma\}$ s.t. $I_{\mathrm{UB}}(\theta_\gamma) < tis[\gamma]$,
   Condition 3. $\exists\gamma \in \{1, 2, \ldots, \Gamma\}$ s.t. $M = tl$ and $tis[\gamma] < I_{\mathrm{LB}}(\theta_\gamma)$.
6. Procedure 6 executes Procedures 2–5 sequentially for all items in the finite set I, resulting in a ZDD representing the family of equivalent item pools $\mathcal{F}$. Then, the two reduction rules are applied to the constructed ZDD to remove redundant nodes and identical subtrees because frontier-based search does not guarantee a canonical graph structure [30]. Consequently, a canonical ZDD representing the family of equivalent item pools $\mathcal{F}$ is obtained by application of the two reduction rules.

The specific algorithms for Procedure 1 through Procedure 6 are presented in Appendix A.

2) Exact search method from approximated ZDD

In 1) Approximated ZDD construction, each element in the test information array of a shared node is approximated by the average of the test information values of two nodes. Therefore, paths that include the shared node are not guaranteed to satisfy test information constraints in Eq. (12) exactly. Additionally, the proposed ZDD construction is unable to control overlapping items. Allowing some overlapping items enables repeated use of items, thereby increasing the total number of equivalent item pools. However, when the maximum number of overlapping items is too large, all equivalent item pools might become nearly identical. (In earlier studies of parallel test assembly, the maximum number of overlapping items is usually set as 20% of the test length (the number of items of each parallel test) (e.g., [19], [20], [28], [29])).

To address these limitations, the proposed method enumerates paths that exactly satisfy the test information constraints and overlapping items constraints from the constructed ZDD.

For the proposed method, equivalent item pools are defined as the families of sets $\mathcal{P} \subseteq \mathcal{F}$ which satisfy all test constraints. The proposed method searches equivalent item pools using the following procedures.

1. Procedure 1 sets $\mathcal{P}$ to the empty set.
2. Procedure 2 searches a subset S from the constructed canonical ZDD $\mathcal{F}$ using random sampling.
3. Procedure 3 proceeds to Procedure 4 when the binary variables of the sampled subset S satisfy the test information constraints; otherwise, it returns to Procedure 2.
4. Procedure 4 proceeds to Procedure 5 when the binary variables of the sampled subset S satisfy the following overlapping item constraint; otherwise, it returns to

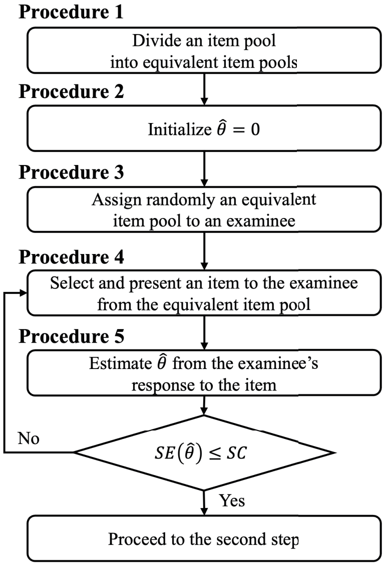

Procedure 1
Divide an item pool into equivalent item pools
Procedure 2
Initialize $\hat{\theta} = 0$
Procedure 3
Assign randomly an equivalent item pool to an examinee
Procedure 4
Select and present an item to the examinee from the equivalent item pool
Procedure 5
Estimate $\hat{\theta}$ from the examinee's response to the item
No
$SE(\hat{\theta}) \leq SC$
Yes
Proceed to the second step

FIGURE 3. Flowchart of the first step.

Procedure 2.

$$\forall P \in \mathcal{P}, \sum_{i \in I} x_i^S x_i^P \leq OC, \tag{17}$$

Therein, x_i^S denotes a binary variable x_i in subset S and x_i^P represents a binary variable x_i in subset P. In addition, OC is defined as the maximum number of common items between any pair of equivalent item pools.

5. In Procedure 5, the sampled subset S is added to the family of sets $\mathcal{P}$ ($\mathcal{P} \leftarrow \mathcal{P} \cup \{S\}$).
6. Procedure 6 repeats Procedures 2–5 until a determined computation time is reached.

Consequently, the proposed method enumerates equivalent item pools that exactly satisfy all test constraints. The specific algorithms for Procedure 1 through Procedure 6 are presented in Appendix B.

C. FIRST STEP WITH EQUIVALENT ITEM POOLS

The first step selects and presents an item from an equivalent item pool assigned to each examinee.

Subsequently, it estimates an examinee's ability, as described in the six procedures below, as presented in Figure 3.

1. Procedure 1 divides an item pool into equivalent item pools using ZDD in Section IV-B.
2. Procedure 2 initializes an examinee's estimated ability to $\hat{\theta} = 0$.
3. Procedure 3 assigns a randomly equivalent item pool from a set of unused item pools to an examinee.
4. Procedure 4 selects an item for the examinee with the maximizing item information for the examinee's ability from the equivalent item pool selected in Procedure 2.
5. Procedure 5 estimates the examinee ability $\hat{\theta}$ based on the examinee's response to the presented item.
6. Procedures 2 and 3 are repeated until the asymptotic error of ability estimate $SE(\hat{\theta})$ reaches a threshold value or less. The threshold value is designated as the switching criterion SC.

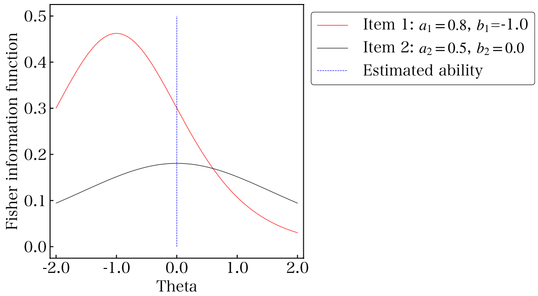


FIGURE 4. Example of an item with high Fisher information for widely various examinee ability value parameters.

Here, if a set of unused item pools is empty in Procedure 1, then the proposed method resets it as a universal set of equivalent item pools.

The first step switches to the second step, which selects the optimum item from the whole item pool and then presents it when the examinee's ability estimate error becomes less than the switching criterion SC. The switching criterion SC is optimized to balance the bias of item exposure and measurement error.

Item selection from an equivalent item pool accelerates the ability estimation to approach the true ability value because the item difficulties in each equivalent item pool are distributed sparsely and uniformly over the abilities of all examinees. At the same time, item selection from equivalent item pools can decrease the bias of item exposure.

D. SECOND STEP WITH THE IDI CONDITION

The first step rapidly approaches a roughly approximated ability estimate of an examinee. The second step reaches a more accurate ability estimate of the examinee because it selects the optimum item from the whole item pool and then presents it. The second step proceeds until the update difference of the estimated examinee ability becomes less than a constant value or less, similarly to traditional CATs. An item selection criterion employing the Fisher information measure becomes more accurate for the second step than for the first step because it is an asymptotic approximation of the inverse estimate variance (e.g. [31]). Therefore, the second step is expected to approach the true ability value efficiently and rapidly without greatly increasing the item exposure.

The second step assumes that an item with a difficulty parameter approximately equal to the ability estimate is selected and presented to an examinee. However, the proposed and traditional CAT methods tend to select and present items with high Fisher information for widely various examinees' abilities. Fig. 4 depicts an example of items 1 and 2 with different

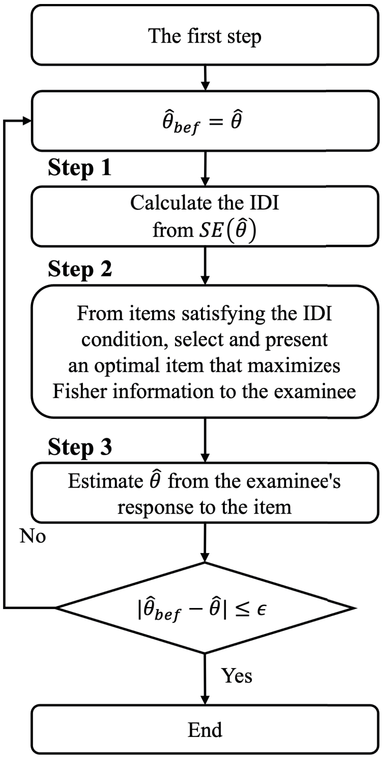

The first step
$\hat{\theta}_{bef} = \hat{\theta}$
Step 1
Calculate the IDI from $SE(\hat{\theta})$
Step 2
From items satisfying the IDI condition, select and present an optimal item that maximizes Fisher information to the examinee
Step 3
Estimate $\hat{\theta}$ from the examinee's response to the item
No
$|\hat{\theta}_{bef} - \hat{\theta}| \leq \epsilon$
Yes
End

FIGURE 5. **Flowchart of the second step.**

discrimination parameter a_i values. The vertical axis and the horizontal axis respectively show the Fisher information and ability. When the ability value is 0.0, item 2 has a lower Fisher information than item 1 has, although the value of item 2 has reached its peak. This phenomenon causes overexposure of items with high Fisher information for widely various abilities of the examinees.

To relax this tendency, in the second step, the proposed method restricts the items which satisfy the IDI condition based on the estimation error for that examinee's ability. Actually, the IDI condition for 3PLM is defined as

$$\hat{\theta} - \delta SE(\hat{\theta}) < b_i < \hat{\theta} + \delta SE(\hat{\theta}), \tag{18}$$

where δ stands for a tuning parameter that is optimized to balance the measurement error and the bias of item exposure. Moreover, $SE(\hat{\theta})$ represents the standard error of the examinee's ability estimate $\hat{\theta}$ as

$$SE(\hat{\theta}) = \sqrt{\int_{-\infty}^{\infty} (\theta - \hat{\theta})^2 f(\theta \mid \mathbf{u})}. \tag{19}$$

In addition, the IDI condition for GPCM is defined as

$$\hat{\theta} - \delta SE(\hat{\theta}) < \beta_{i,s} < \hat{\theta} + \delta SE(\hat{\theta}). \tag{20}$$

More specifically, the algorithm in the second step can be presented as shown below, as illustrated in Figure 5.

1. Procedure 1 calculates the IDI from $SE(\widehat{\theta})$, the current examinee's ability $\hat{\theta}$, and the standard error.
2. From items satisfying the IDI condition, Procedure 2 selects an optimal item that maximizes Fisher information.
3. Procedure 3 updates the estimated examinee ability based on the examinee's responses.
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 item exposure with high discrimination parameters while retaining low measurement error.

V. EXPERIMENTATION

As presented in this section, we optimize the tuning parameters for the proposed method, which presents a tradeoff between decreasing item exposure and increasing measurement accuracy among the values of parameters. Therefore, we evaluate the tradeoff by changing the parameter values to ascertain optimal values and to maximize the method's performance. Subsequently, we compare the performance achieved using the proposed method with the performance achieved using earlier methods.

We use two simulated item pools with 1000 items and two actual item pools for experiments. Items in the simulated item pools have discrimination parameters a_i and difficulty parameters b_i of IRT. In addition, guessing parameters c_i of all items are 0 because it is assumed for this study that the item cannot be answered correctly by guessing. Specifically, we generated two simulated item pools according to the most commonly used approaches applied in earlir studies (e.g., [26], [31]) as

$$\text{Simulation1} : \log a_i \sim N(-0.5, 0.2), b_i \sim N(0,1), \tag{21}$$

$$\text{Simulation2} : \log a_i \sim N(-0.75, 0.2), b_i \sim N(0,1). \tag{22}$$

Table 1 presents details of the actual item pool used for the synthetic personality inventory examination. It is a widely used aptitude test in Japan [32]. For this study, the actual item pool is designated as SPI. The values of guessing parameters c_i in all items from SPI are 0. Additionally, we use another actual item pool (designated as Science), which was used in earlier studies (e.g. [33]). Science includes 918 items with 3PLM and 82 items with GPCM.

We sampled the examinees' actual abilities from $\theta \sim N(0,1)$ 10,000 times. We set the total test length as 30.

A. COMPARISON OF THE NUMBER OF EQUIVALENT ITEM POOLS

This experiment demonstrates the benefits provided by the proposed ZDD method through comparison of the number

TABLE 1. **Details of the actual item pool**

Item Pool Size	Item discrimination Parameter a			Item difficulty parameter b		
	Range	Mean	SD	Range	Mean	SD
978	$0.12 - -3.08$	0.46	0.19	$-4.00 - 4.55$	-0.22	1.57

of equivalent item pools with those assembled using an earlier method, specifically Hybrid Maximum Clique Algorithm using Parallel Integer Programming (HMCAPIP) [19] with simulated and actual item pools. Actually, HMCAPIP [19] is known to assemble the maximum number of parallel test forms among conventional test form assembly methods. Additionally, this study determined the hyperparameters as presented below.

1. The value of M is determined as the value which generates the maximum number of equivalent item pools by changing the value from 5 to 100 in increments of 5.
2. The maximum number of common items OC is set to 20% of M items in each equivalent item pool according to earlier studies ([19], [20], [28], [29]).
3. The lower bound LB_{θ_γ} and the upper bound UB_{θ_γ} of the test information at point θ_γ on the ability level are calculated based on Eq. (13) and Eq. (14).
4. The time limitation for all methods is 24 hr.

The parameter values for HMCAPIP were set based on the explanation presented by [19]. For this study, CPLEX 12.9 [34] was applied to the IP for HMCAPIP.

Table 2 for the simulated item pool and Table 3 for the actual item pool present the number of equivalent item pools produced using the proposed ZDD method and using HMCAPIP by modifying M and OC.

As presented in the tables, the proposed ZDD method assembles more equivalent item pools than the earlier methods do when the number of equivalent item pools becomes large. In fact, the proposed ZDD method assembles a maximum of over 200,000 equivalent item pools, whereas HMCAPIP is limited to assembling only around 100,000 equivalent item pools. The HMCAPIP method is incapable of increasing the number of equivalent item pools to more than around 100.000 because of its high time and space complexity.

By contrast, when the number of equivalent item pools generated by the proposed method is less than 70,000, the proposed method assembles fewer equivalent item pools than HMCAPIP does. The reason for that outcome is that the true number of possible equivalent item pools is small. Therefore, the number of assembled equivalent item pools is also small because the rate of valid paths which satisfy the test information constraints is small in the approximated ZDD.

To confirm that rationale, we calculate $P_{\text{valid, info}}$, which represents the rate of valid paths which satisfy the test information constraints, as

$$P_{\text{valid,info}} = \frac{N_{\text{valid,info}}}{N_{\text{sampled}}},$$

where $N_{\text{valid,info}}$ represents the number of valid paths which satisfy the test information constraints, and where N_{sampled} stands for the total number of random sampling iterations.

Furthermore, the proposed ZDD method assembles equivalent item pools by seeking paths that satisfy the overlapping item constraint among those that satisfy the test information constraints. To assess the rate, we calculate $P_{\text{valid,oc}}$, which represents the rate of valid paths which satisfy the overlapping item constraint as

$$P_{\text{valid,oc}} = \frac{N_{\text{valid,oc}}}{N_{\text{valid,info}}},$$

where $N_{\text{valid,oc}}$ represents the number of valid paths which satisfy the test information constraints and the overlapping item constraint.

Table 4 for the simulated item pool and Table 5 for the actual item pool respectively present $P_{\text{valid,info}}$ and $P_{\text{valid,oc}}$. In Tables 4 and 5, the values of $P_{\text{valid,info}}$ vary from 0.01 to 0.17 depending on the value of M because the upper and lower bounds of test information for equivalent item pools differ according to the value of M.

By contrast, the value of $P_{\text{valid,oc}}$ remains nearly unchanged under all conditions because the value of OC is set to 0.20 of the value of M.

Despite these extremely low values of $P_{\text{valid,info}}$ and $P_{\text{valid,oc}}$, the proposed ZDD method can assemble more equivalent item pools than HMCAPIP can for a large number of possible equivalent item pools.

B. OPTIMIZATION OF EQUIVALENT ITEM POOL AND THE SWITCHING CRITERION

The proposed method entails a tradeoff between the bias of item exposure and the measurement accuracy of an examinee's ability, as affected by the value of the number of items M in each equivalent item pool and the threshold value SC of the switching criterion. Therefore, to infer the optimal value to balance the bias of item exposure and the examinee ability measurement accuracy, we tune the tradeoff by changing values of M and SC via grid search. Actually, the values of M and the threshold value SC are changed respectively as 5 to 100 in 5 steps and 0.05 to 0.50 in 0.05 steps.

Fig. 6 presents the percentage difference of the root mean squared error ($RMSE$) between the students' ability estimates and the true values (red line) on the right vertical axis and the standard deviation of item exposure rate $SD.IER$ (blue line) on the left vertical axis for the value of SC shown on the horizontal axis. Here, the $RMSE$, the $IER_{i,J}$, and the $SD.IER$ are defined respectively as presented below.

$$RMSE = \sqrt{\sum_{j=1}^{J}(\theta_j - \hat{\theta_j})^2}, \tag{23}$$

$$IER_{i,J} = \frac{1}{J}\sum_{j=1}^{J} IE_{i,j}, \tag{24}$$

$$IE_{i,j} = \begin{cases} 1 & \text{if the } i\text{-th item is exposed} \\ \text{to the } j\text{-th examinee, and} & \\ 0 & \text{otherwise.} \end{cases}, \tag{25}$$

$$SD.IER = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(IER_{i,J} - IER_{\mu})^2} \tag{26}$$

TABLE 2. Numbers of equivalent item pools generated from each method by changing the value *M* (simulated item pool).

Item Pool	*M*	HMCAPIP	Proposal
simulation 1	5	**27401**	16307
	10	**97829**	62526
	15	112172	**119658**
	20	114286	**176625**
	25	106048	**235317**
	30	107321	**284231**
	35	100433	**314048**
	40	100175	**319819**
	45	94671	**321774**
	50	92429	**293114**
	55	87720	**270989**
	60	83021	**231472**
	65	80562	**195432**
	70	74354	**123919**
	75	71381	**122775**
	80	65900	**79012**
	85	61288	**66068**
	90	**56682**	48552
	95	**51061**	34227
	100	**42238**	23343
simulation 2	5	**29449**	18916
	10	**101052**	64832
	15	115459	**122190**
	20	117170	**185894**
	25	108152	**252701**
	30	108958	**296333**
	35	101997	**339295**
	40	101620	**340870**
	45	95825	**330727**
	50	93430	**301038**
	55	88578	**270989**
	60	83846	**245029**
	65	80366	**200429**
	70	74069	**162538**
	75	70817	**123870**
	80	65248	**91094**
	85	60953	**72272**
	90	**56479**	50857
	95	**51710**	36256
	100	**43190**	24895

Bold numbers in the table signify the best performances.

TABLE 3. Numbers of equivalent item pools generated using the respective methods by changing the value *M* (Actual item pool).

Item Pool	*M*	HMCAPIP	Proposal
SPI	5	**25024**	18627
	10	**100430**	69156
	15	112052	**128596**
	20	114057	**200869**
	25	105722	**258038**
	30	106570	**311875**
	35	99046	**322250**
	40	98346	**327259**
	45	92143	**313079**
	50	89339	**286742**
	55	84860	**263949**
	60	79806	**174305**
	65	75711	**181155**
	70	70290	**123919**
	75	66591	**97334**
	80	61521	**75823**
	85	**57146**	52225
	90	**50957**	32324
	95	**44360**	23304
	100	**35326**	17754
Science	5	**19896**	17293
	10	**97651**	66782
	15	111745	**115331**
	20	114080	**200411**
	25	105690	**259012**
	30	106811	**301127**
	35	99786	**323312**
	40	99774	**331631**
	45	94193	**310294**
	50	91779	**282901**
	55	87032	**264189**
	60	82271	**181921**
	65	79052	**168590**
	70	73128	**133911**
	75	69757	**101289**
	80	64386	**77518**
	85	**60031**	59823
	90	**54944**	45282
	95	**50162**	41234
	100	**41554**	24822

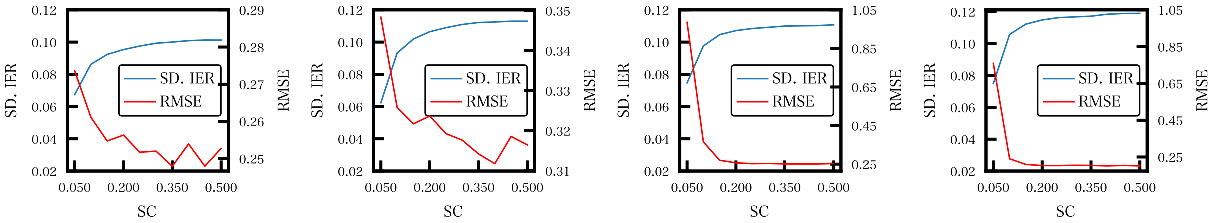
Bold numbers in the table signify the best performances.

TABLE 4. Rates of valid paths that satisfy the test information constraints and overlapping item constraint (Simulated item pool).

Item Pool	M	$P_{\text{valid,info}}$	$P_{\text{valid,oc}}$
simulation 1	5	0.008	1.35×10^{-3}
	10	0.027	1.54×10^{-3}
	15	0.070	1.12×10^{-3}
	20	0.086	1.42×10^{-3}
	25	0.104	1.52×10^{-3}
	30	0.170	1.12×10^{-3}
	35	0.128	1.53×10^{-3}
	40	0.107	1.86×10^{-3}
	45	0.170	1.12×10^{-3}
	50	0.100	1.75×10^{-3}
	55	0.131	1.23×10^{-3}
	60	0.068	1.56×10^{-3}
	65	0.068	1.62×10^{-3}
	70	0.067	1.12×10^{-3}
	75	0.048	1.23×10^{-3}
	80	0.025	1.86×10^{-3}
	85	0.026	1.23×10^{-3}
	90	0.011	1.72×10^{-3}
	95	0.011	1.25×10^{-3}
	100	0.009	1.26×10^{-3}
simulation 2	5	0.008	1.37×10^{-3}
	10	0.031	1.33×10^{-3}
	15	0.054	1.31×10^{-3}
	20	0.089	1.37×10^{-3}
	25	0.121	1.31×10^{-3}
	30	0.140	1.31×10^{-3}
	35	0.143	1.38×10^{-3}
	40	0.147	1.38×10^{-3}
	45	0.132	1.43×10^{-3}
	50	0.105	1.65×10^{-3}
	55	0.099	1.63×10^{-3}
	60	0.076	1.45×10^{-3}
	65	0.066	1.56×10^{-3}
	70	0.065	1.26×10^{-3}
	75	0.048	1.30×10^{-3}
	80	0.027	1.74×10^{-3}
	85	0.025	1.45×10^{-3}
	90	0.022	1.24×10^{-3}
	95	0.019	1.34×10^{-3}
	100	0.010	1.54×10^{-3}

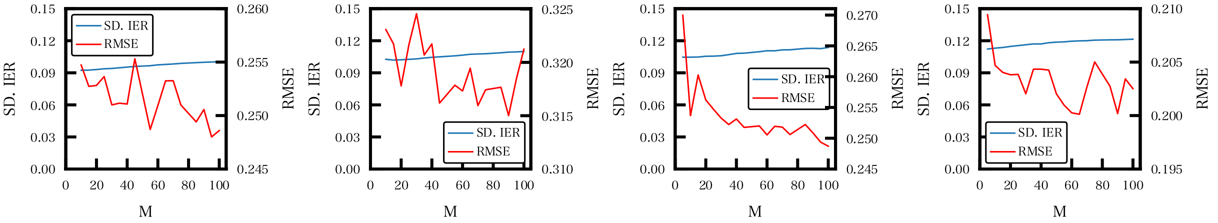
TABLE 5. Rates of valid paths that satisfy the test information constraints and overlapping item constraint (Actual item pool).

Item Pool	M	$P_{\text{valid,info}}$	$P_{\text{valid,oc}}$
SPI	5	0.008	1.43×10^{-3}
	10	0.029	1.44×10^{-3}
	15	0.060	1.31×10^{-3}
	20	0.091	1.34×10^{-3}
	25	0.126	1.25×10^{-3}
	30	0.150	1.27×10^{-3}
	35	0.136	1.45×10^{-3}
	40	0.149	1.34×10^{-3}
	45	0.144	1.33×10^{-3}
	50	0.100	1.75×10^{-3}
	55	0.113	1.42×10^{-3}
	60	0.074	1.44×10^{-3}
	65	0.090	1.23×10^{-3}
	70	0.049	1.54×10^{-3}
	75	0.039	1.54×10^{-3}
	80	0.032	1.45×10^{-3}
	85	0.024	1.33×10^{-3}
	90	0.014	1.44×10^{-3}
	95	0.010	1.42×10^{-3}
	100	0.008	1.41×10^{-3}
Science	5	0.007	1.54×10^{-3}
	10	0.028	1.43×10^{-3}
	15	0.048	1.45×10^{-3}
	20	0.079	1.54×10^{-3}
	25	0.110	1.44×10^{-3}
	30	0.136	1.35×10^{-3}
	35	0.120	1.64×10^{-3}
	40	0.131	1.54×10^{-3}
	45	0.115	1.64×10^{-3}
	50	0.120	1.44×10^{-3}
	55	0.113	1.43×10^{-3}
	60	0.084	1.32×10^{-3}
	65	0.084	1.23×10^{-3}
	70	0.053	1.54×10^{-3}
	75	0.035	1.75×10^{-3}
	80	0.031	1.54×10^{-3}
	85	0.023	1.61×10^{-3}
	90	0.021	1.29×10^{-3}
	95	0.017	1.47×10^{-3}
	100	0.011	1.37×10^{-3}



(a) Simulation 1 (b) Simulation 2 (c) SPI (d) Science

FIGURE 6. RMSE of the measurement accuracy and the standard deviation of item exposure for each value of *SC*.



(a) Simulation 1 (b) Simulation 2 (c) SPI (c) Science

FIGURE 7. RMSE of measurement accuracy and the standard deviation of item exposure for each value of *M*.

Therein, $IER_{i,J}$ and IER_μ respectively represent the item exposure rate of item i for all examinees and the average of item exposure rate over all items for all examinees ($IER_\mu = \frac{1}{n}\sum_{i=1}^{n} IER_{i,J}$). The left vertical axis, which shows $SD.IER$ given each value of SC in Fig. 6, depends on the value of M. Therefore, the left vertical axis shown in Fig. 6 represents the minimum $SD.IER$ by changing $M = 5$ to $M = 100$ in five steps for each value of SC. Fig. 6 shows that the $SD.IER$ becomes large when SC becomes large, but the $RMSE$ becomes small. To resolve this tradeoff, we determined each optimal value of SC for an item pool to minimize the value of $SD.IER$ among a set of items with the lowest $RMSE$ within a three-significant-digit range.

Fig. 7 portrays the tradeoff between the RMSE and the $SD.IER$ by changing M. As presented in Fig. 7, when the value of M becomes small, the percentage standard deviation of item exposure rate $SD.IER$ decreases, but the $RMSE$ of measurement accuracy tends to increase. Specifically, when $M < 50$, the $RMSE$ of the measurement accuracy tends to increase. By contrast, improvement of the percentage standard deviation of item exposure rate $SD.IER$ is limited by changing the value of M. Therefore, for this study, we found the optimal value of M for an item pool to minimize $SD.IER$ given a condition to minimize the $RMSE$ within a three-significant-digit range.

C. EFFECTIVENESS OF THE IDI CONDITION

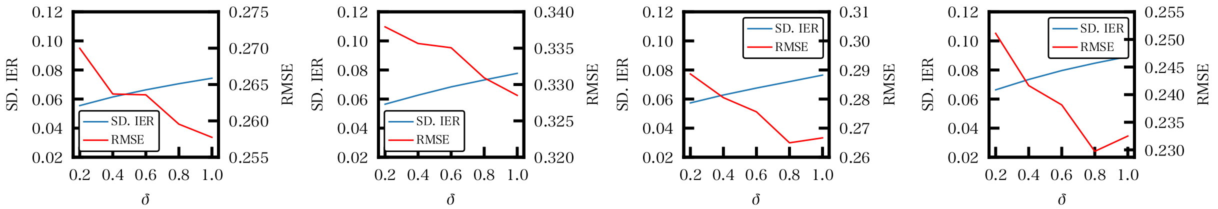
For item selection of the second step, the proposed method restricts the items which satisfy the Item Difficulty Interval (IDI) condition. The IDI condition is expected to mitigate the tradeoff between the bias of item exposure and the $RMSE$ by tuning parameter δ. Increasing δ increases the interval length of the IDI condition to mitigate the degree of restriction for item selection. Specifically, the value of the tuning parameter δ was changed 0.2 to 1.0 in 0.2 increments to ascertain the optimal value to balance the bias of item exposure and the $RMSE$.

TABLE 6. Determined values of parameter δ.

Item pool	δ
Simulation 1	1.0
Simulation 2	1.0
SPI	0.8
Science	0.8

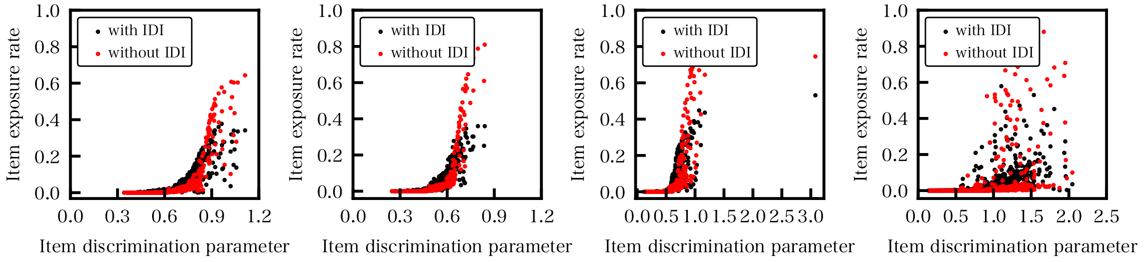
Fig. 8 presents the $RMSE$ (red line) on the right vertical axis and shows the percentage standard deviation of the item exposure rate $SD.IER$ (blue line) on the left vertical axis for the value of parameter δ. As shown in Fig. 8, $SD.IER$ becomes large when δ becomes large. However, the $RMSE$ becomes small because increasing the value of δ mitigates the degree of restriction for item selection. Accordingly, as presented in Table 6, the values of the parameter δ were determined to minimize $SD.IER$, given the condition to minimize $RMSE$ within a three-significant-digit range for an item pool.

Fig. 9 depicts a scatter plot of the i-th item's discrimination parameter value a_i and the i-th item's item exposure rate $IER_{i,J}$ for the proposed method with and without the IDI condition. When the discrimination parameter value becomes large, the item exposure of the proposed method without the IDI condition tends to become large because these items tend to have high Fisher information for widely various examinee abilities. By contrast, the proposed method using the IDI condition selects and presents items with a lower bias of item exposure than that obtained using the proposed method without the IDI condition. In addition, Fig. 10 depicts a scatter



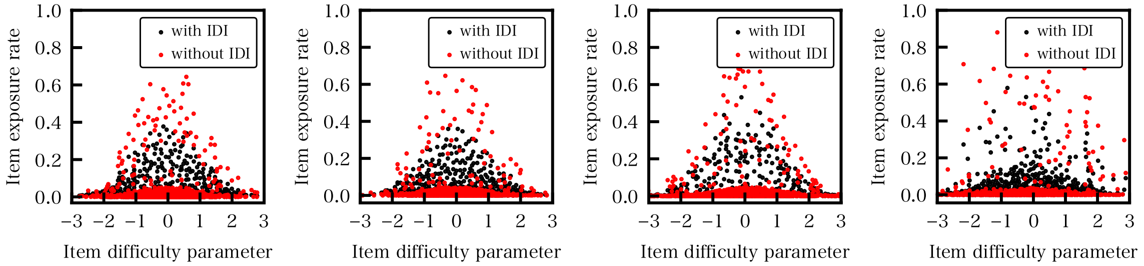
(a) Simulation 1 (b) Simulation 2 (c) SPI (d) Science

FIGURE 8. Tradeoff between the examinee ability measurement accuracy and the bias of item exposure by changing parameter δ.



(a) Simulation 1 (b) Simulation 2 (c) SPI (d) Science

FIGURE 9. Scatter plot of the *i*-th item's discrimination parameter value a_i and the *i*-th item's item exposure rate IER_i for the proposed method with and without the IDI condition using each item pool.



(a) Simulation 1 (b) Simulation 2 (c) SPI (d) Science

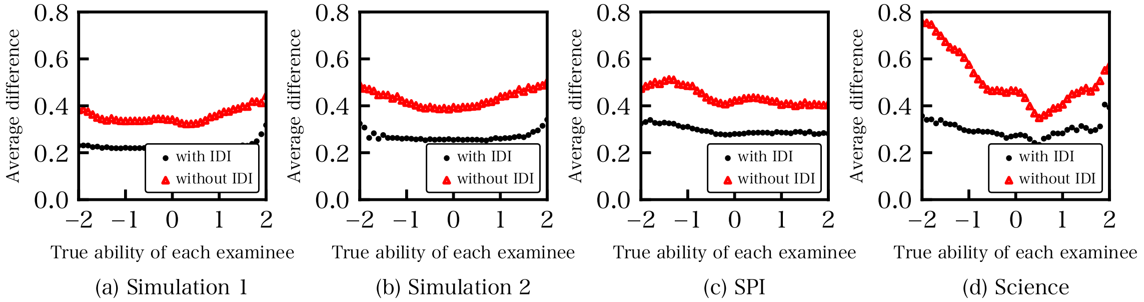
FIGURE 10. Scatter plot of the *i*-th item's difficulty parameter value b_i and the *i*-th item's item exposure rate IER_i for the proposed method with and without the IDI condition using each item pool.

plot of the i-th item's difficulty parameter value and the i-th item's item exposure rate $IER_{i,J}$ for the proposed method with and without the IDI condition. Each difficulty parameter of GPCM in Fig. 10 presents only the closest value of step difficulty parameters to the examinee ability estimate. When the difficulty parameter value is approximately $b_i = 0.0$, the $IER_{i,J}$ of the proposed method without the IDI condition tends to become large because the number of examinees for whom ability estimates are almost zero is the largest, as described previously. It is noteworthy that the proposed method with the IDI condition can select and present items with lower item exposure than the proposed method without the IDI condition can.

Regarded in greater detail, Fig. 11 depicts the average difference between the presented item difficulty parameter and the examinee ability estimate for each ability value in the second step. The horizontal axis shows the true ability of each examinee. The vertical axis shows the average difference between the presented item difficulty parameter and the examinee ability estimate in the second step given as

$$\frac{1}{n-l'+1}\sum_{l=l'}^{L}(\hat{\theta}_{l-1}-b_l)^2, \tag{27}$$

where b_l is the difficulty parameter of the l-th presented item, and where $\hat{\theta}_l$ represents the ability estimate after the examinee answered the l-th item. Each difficulty parameter b_l of GPCM in Eq. (27) represents only the closest value of step difficulty parameters to the examinee ability estimate. Items from l'-th to L-th are presented in the second step. Fig. 11 shows that the proposed method without the IDI condition often selects items with difficulty parameter values that differ greatly from the ability estimates. In contrast, the proposed method with the IDI condition selects items with difficulty parameter values that more closely approximate the ability estimates.



(a) Simulation 1 (b) Simulation 2 (c) SPI (d) Science

FIGURE 11. Average difference between the presented item difficulty parameter and the estimated ability for each ability value in the second step.

TABLE 7. Details of threshold values for item pools in Situation 3

Item pool	Threshold value
Simulation 1	0.25
Simulation 2	0.31
SPI	0.25
Science	0.20

D. COMPARING THE PROPOSED METHOD TO THE CONVENTIONALLY USED METHOD

This section presents a comparison of the performance results of the proposed method using ZDD (designated as Proposed (ZDD)), the proposed method using approximated ZDD in section IV-B1) (designated as Proposed (Approximated ZDD)), only the first step of Proposed (ZDD) (designated as Proposed (ZDD first-step)), the proposed method using Hybrid Maximum Clique Algorithm using Parallel Integer Programming [20] (designated as Proposed (HMCAPIP) [29]), and only the first step of Proposed (HMCAPIP) (designated as Proposed (HMCAPIP first-step), which is the same method as that presented by Ueno and Miyazawa [28]) to the performance results of other computerized adaptive testing methods (conventional adaptive testing in Section II-B (designated as CAT), van der Linden's IP-based method [7] in Section III-A (designated as IP), Linden and Choi's item-eligibility probability method [11] in Section III-B (designated as Prob)), Choi and Lim's target information method [9] (designated as TI), the a-stratification method [13], [14] (designated as a-stratification), and Lim and Choi's hybrid method [12] in Section III-D (designated as Hybrid).

Comparisons of the performances of the proposed methods to those of other methods were conducted under the following two experiment conditions.

1. Experiment condition 1 has fixed test length. Linden [35] described that this condition is in agreement with practice in nearly all reported adaptive testing studies (e.g. [9], [13], [14], [35]). It fairly compares the performance of the proposed methods to those of earlier methods. Particularly, experiment condition 1 has two test constraints: the test length is 30; and the number of examinees is 10,000. These test constraints are the same as those described in Sections V-B and V-C. In addition, experiment condition 1 restricts the upper bound to be imposed on the exposure rates of the items. Specifically, experiment condition 1 restricts the maximum number of item exposures to 50% of the total number of examinees according to [28], [29], [36]. For instance, for Prob and Hybrid, we set the upper bound of exposure rate as $r^{max} = 0.5$ according to [11], [12]. For IP, we set the maximum number of times as $R = 5,000$ according to [7]. Additionally, we apply a restriction of the upper-bound of item exposure [37] to CAT (designated as CAT (Restrict)) and TI (designated as TI (Restrict)). By the restriction method, when an item is used by more than 50% of the total number of examinees, the item is removed from the item pool.
2. Experiment condition 2 applies a stopping rule based on a predetermined level of accuracy for ability estimation according to Wainer et al. [38]. The stopping rule, which is expected to reduce the test length without decreasing the measurement accuracy, is used frequently in actual settings of adaptive tests. By the stopping rule, each method repeats to select and present items until the standard error of the estimated examinee ability decreases to a threshold value or less. Actually, we set the threshold values for each item pool as shown in Table 7. These thresholds are obtained as the average of the standard error of the examinee's ability estimate (Eq. 19) using conventional CAT in experiment condition 1. In addition, when the standard error of the examinee's ability estimate does not converge, each method finished CAT when 60 items were presented. It is noteworthy that CAT and the proposed method can apply the stopping rule to reduce the test length without decreasing the measurement accuracy. By contrast, IP, Prob, TI, and Hybrid cannot apply the stopping rule because these methods require the test length in the constraint of IP. Therefore, these methods are inapplicable when using experiment condition 2.

Table 8 presents the standard deviation of item exposure rate $SD.IER$ (Eq. 26), the maximum number of item exposure rate $Max.IER$, $RMSE$ (Eq. 23), and the number of non-presented items in experiment condition 1. Here, the non-presented items are those items which have not been pre-

TABLE 8. Comparing the proposed method to earlier methods for experiment condition 1

Item pool	Method	$SD.IER$	$Max.IER$	Number of non-presented items	$RMSE$
simulation 1	CAT	0.105	1.000	843	**0.26**
	CAT (Restrict)	0.097	0.500	826	**0.26**
	IP	0.098	0.500	829	0.25
	Prob	0.098	0.556	837	0.25
	TI	0.101	1.000	268	**0.26**
	TI (Restrict)	0.094	0.500	758	**0.26**
	a-stratification	0.092	1.000	476	**0.26**
	Hybrid	0.085	0.495	542	0.26
	Proposed (HMCAPIP first-step)	0.029	0.163	15	0.33
	Proposed (HMCAPIP)	0.066	0.434	219	**0.26**
	Proposed (Approximated ZDD)	0.088	0.503	182	0.33
	Proposed (ZDD first-step)	**0.023**	**0.154**	**3**	0.33
	Proposed (ZDD)	0.061	0.372	60	**0.26**
simulation 2	CAT	0.116	1.000	874	**0.32**
	CAT (Restrict)	0.100	0.500	845	**0.32**
	IP	0.102	0.500	841	0.33
	Prob	0.103	0.541	864	**0.32**
	TI	0.110	1.000	306	**0.32**
	TI (Restrict)	0.101	0.500	748	**0.32**
	a-stratification	0.086	1.000	321	0.33
	Hybrid	0.095	0.500	689	0.33
	Proposed (HMCAPIP first-step)	0.029	0.183	68	0.39
	Proposed (HMCAPIP)	0.071	0.557	281	0.34
	Proposed (Approximated ZDD)	0.083	0.611	123	0.39
	Proposed (ZDD first-step)	**0.026**	**0.172**	56	0.39
	Proposed (ZDD)	0.056	0.365	**52**	0.34
SPI	CAT	0.114	1.000	850	**0.26**
	CAT (Restrict)	0.100	0.500	822	**0.26**
	IP	0.102	0.500	819	**0.26**
	Prob	0.103	0.545	840	**0.26**
	TI	0.106	1.000	482	**0.26**
	TI (Restrict)	0.100	0.500	798	**0.26**
	a-stratification	0.093	1.000	522	**0.26**
	Hybrid	0.098	0.500	678	0.27
	Proposed (HMCAPIP first-step)	0.034	0.278	135	0.37
	Proposed (HMCAPIP)	0.072	0.593	398	**0.26**
	Proposed (Approximated ZDD)	0.092	0.551	332	0.36
	Proposed (ZDD first-step)	**0.029**	**0.246**	**111**	0.37
	Proposed (ZDD)	0.071	0.531	281	**0.26**
Science	CAT	0.122	1.000	893	**0.21**
	CAT (Restrict)	0.104	0.500	855	**0.21**
	IP	0.104	0.500	854	**0.21**
	Prob	0.107	0.531	878	**0.21**
	TI	0.107	1.000	754	**0.21**
	TI (Restrict)	0.102	0.500	759	**0.21**
	a-stratification	0.092	1.000	543	0.22
	Hybrid	0.087	0.496	689	0.22
	Proposed (HMCAPIP first-step)	0.061	0.292	267	0.36
	Proposed (HMCAPIP)	0.071	0.529	399	0.24
	Proposed (Approximated ZDD)	0.088	0.533	253	0.36
	Proposed (ZDD first-step)	**0.056**	**0.282**	**216**	0.36
	Proposed (ZDD)	0.066	0.487	236	0.24

Bold numbers in the table signify the best performances.

sented by any examinee.

Proposed (HMCAPIP) and Proposed (ZDD) provide lower values of $SD.IER$, $Max.IER$, and "Number of non-presented items" than earlier methods provide without greatly increasing the $RMSE$. Especially, Proposed (ZDD) provides the lowest values of $SD.IER$, $Max.IER$, and "Number of non-presented items" without greatly increasing the associated $RMSE$. The results demonstrate that Proposed (ZDD) provides the best performance of tradeoff control between decreasing item exposure and increasing measurement accuracy using the standard CAT's condition. Furthermore, Proposed (ZDD first-step) and Proposed (ZDD) respectively have lower values of $SD.IER$, $Max.IER$, and "Number of non-presented items" than either Proposed (HMCAPIP first-step) or Proposed (HMCAPIP) has. Results indicate that the proposed method using ZDD is more effective at mitigating the tradeoff than the proposed method using the maximum clique algorithm and IP is. By contrast, Proposed (HMCAPIP first-step) and Proposed (ZDD first-step) provide lower values of $SD.IER$ and $Max.IER$ than other methods provide. In addition, Proposed (ZDD first-step) provides the very lowest values of "Number of non-presented items" in almost all

TABLE 9. Comparing the proposed method to earlier methods for experiment condition 2

Item pool	Method	$SD.IER$	$Max.IER$	Number of non-presented items	$RMSE$	$Avg.TL$
simulation 1	CAT	0.104	1.000	773	**0.25**	**30.45**
	CAT (Restrict)	0.098	0.500	723	**0.25**	31.25
	Proposed (HMCAPIP first-step)	0.110	0.601	35	0.27	59.81
	Proposed (HMCAPIP)	0.071	0.508	97	0.27	36.16
	Proposed (Approximated ZDD)	0.115	0.592	55	0.27	50.17
	Proposed (ZDD first-step)	0.104	0.567	**2**	0.27	59.67
	Proposed (ZDD)	**0.066**	**0.478**	75	0.26	35.67
simulation 2	CAT	0.117	1.000	810	**0.31**	**31.72**
	CAT (Restrict)	0.108	0.500	768	**0.31**	33.72
	Proposed (HMCAPIP first-step)	0.109	0.641	24	0.35	58.23
	Proposed (HMCAPIP)	0.078	0.558	135	0.33	37.35
	Proposed (Approximated ZDD)	0.115	0.583	123	0.35	52.17
	Proposed (ZDD first-step)	0.102	0.630	**3**	0.35	57.12
	Proposed (ZDD)	**0.071**	**0.506**	103	0.32	36.25
SPI	CAT	0.108	1.000	757	**0.24**	**30.30**
	CAT (Restrict)	0.101	0.500	702	**0.24**	31.35
	Proposed (HMCAPIP first-step)	0.091	0.644	14	0.26	59.51
	Proposed (HMCAPIP)	0.075	0.580	173	**0.24**	38.16
	Proposed (Approximated ZDD)	0.094	0.592	111	0.26	48.53
	Proposed (ZDD first-step)	0.082	0.623	**3**	0.26	59.21
	Proposed (ZDD)	**0.069**	**0.531**	87	**0.24**	37.67
Science	CAT	0.121	1.000	801	**0.20**	**31.90**
	CAT (Restrict)	0.111	0.500	786	**0.20**	33.11
	Proposed (HMCAPIP first-step)	0.098	0.684	13	0.23	59.61
	Proposed (HMCAPIP)	0.081	0.535	40	0.21	54.41
	Proposed (Approximated ZDD)	0.105	0.632	60	0.23	49.02
	Proposed (ZDD first-step)	0.095	0.682	**3**	0.23	59.21
	Proposed (ZDD)	**0.075**	**0.502**	26	0.21	38.12

Bold numbers in the table signify the best performances.

cases. Proposed (HMCAPIP first-step) and Proposed (ZDD first-step) provide the largest values of $RMSE$. These methods present a large tradeoff between decreasing $SD.IER$ and decreasing $RMSE$.

By contrast, Proposed (Approximated ZDD) enumerates equivalent item pools based on the ZDD that approximates test information value of the merged two nodes by their average value. Therefore, Proposed (Approximated ZDD) does not guarantee to satisfy the test information constraints. As a result, Proposed (Approximated ZDD) provides higher values of $RMSE$ than Proposed (ZDD) does.

Additionally, Proposed (Approximated ZDD) does not control the overlapping items among equivalent item pools. In fact, as demonstrated in the experiment described in Section V-A, the rates of valid from the approximated ZDD-based item pools pasts which satisfy the overlapping item constraint are less than 2×10^{-3}. As a result, Proposed (Approximated ZDD) provides higher values of $SD.IER$, $MAX.IER$, and the 'number of non-presented items' than Proposed (ZDD) provides.

When using earlier methods, TI provides the lowest values of RMSE, but the values of $SD.IER$ are nearly identical to those of CAT. Additionally, TI, CAT, and a-stratification yield the same values of $Max.IER$: 1.000. This finding implies that one or more items are presented to all examinees. IP, Prob, and TI (Restrict) provide values of $Max.IER$ as around 0.500, but the values of $SD.IER$ are as large as those of CAT (Restrict). A-stratification and Hybrid provide lower values of $SD.IER$ than earlier methods provide, but the "Number of non-presented items" is still large.

Table 9 presents the $SD.IER$, the $Max.IER$, the $RMSE$, the number of non-presented items, and the average of test length $Avg.TL$ of the methods in experiment condition 2.

The results indicate that Proposed (HMCAPIP) and Proposed (ZDD) provide lower values of $SD.IER$ and $Max.IER$ than other methods provide, but without greatly increasing the test length or the RMSE. Moreover, Proposed (HMCAPIP) and Proposed (ZDD) provide lower values of "Number of non-presented items" than any method except for Proposed (HMCAPIP first-step) and Proposed (ZDD first-step). Especially, Proposed (ZDD) provides lower values of $SD.IER$, $Max.IER$, and "Number of non-presented items" than Proposed (HMCAPIP) provides. The results demonstrate that, in actual settings, Proposed (ZDD) provides the best performance of tradeoff control between decreasing item exposure and increasing measurement accuracy without greatly increasing the test length.

Proposed (HMCAPIP first-step) and Proposed (ZDD first-step) provide lower values of "Number of non-presented items" than other methods provide, but these methods provide the highest values of $Avg.TL$ because of their high $RMSE$.

Similar to experiment condition 1, Proposed (Approximated ZDD) provides higher values of $SD.IER$, $Max.IER$, "Number of non-presented items", $RMSE$ than Proposed (ZDD) does. Additionally, Proposed (Approximated ZDD) provides higher values of $Avg.TL$ because of their high $RMSE$.

By contrast, CAT provides the lowest values of $Avg.TL$, but $SD.IER$ is extremely large. Additionally, CAT produces the values of $Max.IER$ as 1.000. This case implies that one or more items are exposed to all examinees. In addition, CAT (Restrict) provides values of $Max.IER$ of approximately 0.500, but the values of $SD.IER$ are as large as those of CAT.

VI. CONCLUSION

Conventional CAT entails the difficulties posed by a tradeoff between increased measurement accuracy and decreased item exposure in an item pool. To resolve this tradeoff dilemma, we proposed two-step adaptive testing using zero-suppressed binary decision diagrams. During the initial step, an optimal item is selected and presented from an equivalent item pool divided using automated parallel test form assembly with zero-suppressed binary decision diagrams. The proposed method switches to the second step when the examinee's ability estimate converges. The second step selects and presents the optimal item with a difficulty parameter value approximating the examinee's ability estimate from the whole item pool. The first step rapidly provides a roughly approximated ability estimate of an examinee. The second step reaches a more accurate ability estimate of the examinee.

Experiments were conducted for comparison of the performance achieved using the proposed method with the performance achieved using conventional methods. Results of empirical experimentation demonstrated that the proposed method provides a lower bias of item exposure than the compared methods did, but while maintaining low measurement error. Especially, the results demonstrated that partitioning an item pool to several equivalent item pools with an appropriate number of items is extremely effective to address the tradeoff difficulty associated with CAT. Based on these results, we recommend development of a large item pool, with subsequent assembly of equivalent item pools.

Recently, through the rapid progress achieved in the study of artificial intelligence, several CAT methods [39], [40] using deep learning have been proposed to improve the measurement accuracy of an examinee's ability. These CAT methods are applicable to the idea of the proposed adaptive testing method. In addition, various knowledge tracing (KT) methods (e.g. Deep-IRT [41]–[43]) have been proposed for adaptive learning systems using deep learning to discover concepts that the student has not mastered. That discovery is achieved by tracing a student's evolving knowledge state. As future work, we expect to apply the proposed adaptive testing method to CAT methods using deep learning, KT methods, and adaptive learning systems [44], [45].

ACKNOWLEDGMENTS

The authors thank Dr. Wim J. van der Linden and Dr. Seung W. Choi for helpful comments on this work. Parts of this research were reported in an earlier conference paper, published as presented in references [29], [36]. This work was supported by JSPS KAKENHI Grants: Nos. JP24H00739 and JP24KJ1124.

REFERENCES

[1] W. Linden and C. Glas, *Computerized Adaptive Testing: Theory and Practice*, 1st ed. Dordrecht, Netherlands: Springer, 2000.

[2] P. A. Jewsbury and P. W. van Rijn, "IRT and MIRT models for item parameter estimation with multidimensional multistage tests," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 4, pp. 383–402, Oct. 2020.

[3] I. V. Mullis, M. O. Martin, and M. von Davier, "Timss 2023 assessment frameworks," 2021. [Online]. Available: https://timssandpirls.bc.edu/timss2023/frameworks/index.html

[4] I. V. Mullis and M. O. Martin, "Pirls 2021 assessment frameworks," 2019. [Online]. Available: https://timssandpirls.bc.edu/pirls2021/frameworks/

[5] K. Yamamoto, H. J. Shin, and L. Khorramdel, "Multistage adaptive testing design in international large-scale assessments," *Educational Measurement: Issues and Practice*, vol. 37, no. 4, pp. 16–27, Oct. 2018.

[6] M. Ueno, K. Fuchimoto, and E. Tsutsumi, "e-testing from artificial intelligence approach," *Behaviormetrika*, vol. 48, no. 2, pp. 409–424, Jul. 2021.

[7] W. J. van der Linden, "Review of the shadow-test approach to adaptive testing," *Behaviormetrika*, vol. 49, no. 2, pp. 169–190, Sep. 2021.

[8] W. D. Way, "Protecting the integrity of computerized testing item pools," *Educational Measurement: Issues and Practice*, vol. 17, no. 4, pp. 17–27, 1998.

[9] S. W. Choi and S. Lim, "Adaptive test assembly with a mix of set-based and discrete items," *Behaviormetrika*, vol. 49, no. 1, pp. 231–254, Aug. 2022.

[10] G. G. Kingsbury and A. R. Zara, "Procedures for selecting items for computerized adaptive tests," *Applied Measurement in Education*, vol. 2, no. 4, pp. 359–375, 1989.

[11] W. J. van der Linden and S. W. Choi, "Improving item-exposure control in adaptive testing," *Journal of Educational Measurement*, vol. 57, no. 3, pp. 405–422, Sep. 2019.

[12] S. Lim and S. W. Choi, "Item exposure and utilization control methods for optimal test assembly," *Behaviormetrika*, vol. 51, no. 1, pp. 124–156, Dec. 2023.

[13] H.-H. Chang and Z. Ying, "A-stratified multistage computerized adaptive testing," *Applied Psychological Measurement*, vol. 23, no. 3, pp. 211–222, 1999.

[14] H.-H. Chang and W. J. Van der Linden, "Optimal stratification of item pools in α-stratified computerized adaptive testing," *Applied Psychological Measurement*, vol. 27, no. 4, pp. 262–274, 2003.

[15] T. Ishii, P. Songmuang, and M. Ueno, "Maximum clique algorithm for uniform test forms assembly," in *Artificial Intelligence in Education*. Berlin, Heidelberg, Germany: Springer, 2013, pp. 451–462.

[16] ——, "Maximum clique algorithm and its approximation for uniform test form assembly," *IEEE Transactions on Learning Technologies*, vol. 7, no. 1, pp. 83–95, Jan. 2014.

[17] T. Ishii and M. Ueno, "Clique algorithm to minimize item exposure for uniform test forms assembly," in *Artificial Intelligence in Education*, Cham, Switzerland, Jan. 2015, pp. 638–641.

[18] ——, "Algorithm for uniform test assembly using a maximum clique problem and integer programming," in *Artificial Intelligence in Education*, Cham, Switzerland, Jun. 2017, pp. 102–112.

[19] K. Fuchimoto, T. Ishii, and M. Ueno, "Hybrid maximum clique algorithm using parallel integer programming for uniform test assembly," *IEEE Transactions on Learning Technologies*, vol. 15, no. 2, pp. 252–264, 2022.

[20] K. Fuchimoto, S.-I. Minato, and M. Ueno, "Automated parallel test forms assembly using zero-suppressed binary decision diagrams," *IEEE Access*, vol. 11, no. 1, pp. 112 804–112 813, Oct. 2023.

[21] S.-i. Minato, "Zero-suppressed bdds for set manipulation in combinatorial problems," in *Proceedings of the 30th International Design Automation Conference*, Dallas, TX, USA, Mar. 1993, pp. 272–277.

[22] Y. Hou, Y. Wu, and H. Han, "Multistate-constrained multiobjective differential evolution algorithm with variable neighborhood strategy," *IEEE Transactions on Cybernetics*, vol. 53, no. 7, pp. 4459–4472, Aug. 2022.

[23] Y. Qin, J. Ren, D. Yang, H. Zhou, H. Zhou, and C. Ma, "Decomposition-based multiobjective evolutionary algorithm with density estimation-based dynamical neighborhood strategy," *Applied Intelligence*, vol. 53, no. 24, pp. 29 863–29 901, Nov 2023.

[24] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick, Eds. Reading: Addison-Wesley, 1968, pp. 397–479.

[25] E. Muraki, "A generalized partial credit model: Application of an em algorithm," *Applied Psychological Measurement*, vol. 16, no. 2, pp. 159–176, Jun. 1992.

[26] F. B. Baker and S.-H. Kim, *Item response theory: Parameter estimation techniques*, 2nd ed. Boca Raton, FL, USA: CRC press, 2004.

[27] F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores*. Addison-Wesley, Washington, WA, USA, 1968.

[28] M. Ueno and Y. Miyazawa, "Uniform adaptive testing using maximum clique algorithm," in *Artificial Intelligence in Education*. Springer, Cham, Switzerland, 2019, pp. 482–493.

[29] W. Kishida, K. Fuchimoto, Y. Miyazawa, and M. Ueno, "Item difficulty constrained uniform adaptive testing," in *International Conference on Artificial Intelligence in Education*. Springer, Cham, Switzerland, 2023, pp. 568–573.

[30] D. E. Knuth, *The art of computer programming: Bitwise tricks & techniques; Binary Decision Diagrams*, 1st ed. Boston, MA, USA: Addison-Wesley, 2009, vol. 4, no. 1.

[31] W. J. van der Linden and others, *Handbook of Item Response Theory, Volume One: Models*, 1st ed. Boca Raton, FL, USA: Chapman and Hall/CRC, 2016.

[32] Recruit, "Synthetic Personality Inventory," 2024. [Online]. Available: https://www.spi.recruit.co.jp/

[33] S. W. Choi, S. Lim, and W. J. van der Linden, "TestDesign: an optimal test design approach to constructing fixed and adaptive tests in R," *Behaviormetrika*, vol. 49, no. 2, pp. 191–229, Aug. 2022.

[34] IBM, "Ilog cplex optimization studio cplex 12.9," 2019. [Online]. Available: https://www.ibm.com/jp-ja/products/ilog-cplex-optimization-studio

[35] W. J. Van der Linden and C. A. W. Glas, *Elements of adaptive testing*. New York, NY, USA: Springer, 2010, vol. 10.

[36] M. Ueno and Y. Miyazawa, "Two-Stage uniform adaptive testing to balance measurement accuracy and item exposure," in *Artificial Intelligence in Education*. Springer, Cham, Switzerland, 2022, pp. 626–632.

[37] J. Revuelta and V. Ponsoda, "A comparison of item exposure control methods in computerized adaptive testing," *Journal of Educational Measurement*, vol. 35, no. 4, pp. 311–327, Dec. 1998.

[38] H. Wainer, N. J. Dorans, B. F. Green, L. Steinberg, R. Flaugher, R. J. Mislevy, and D. Thissen, *Computerized adaptive testing: A primer.* Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc, 1990.

[39] Y. Zhuang, Q. Liu, Z. Huang, Z. Li, S. Shen, and H. Ma, "Fully adaptive framework: Neural computerized adaptive testing for online education," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, pp. 4734–4742, Jun. 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/20399

[40] W. Feng, A. Ghosh, S. Sireci, and A. S. Lan, "Balancing test accuracy and security in computerized adaptive testing," in *International Conference on Artificial Intelligence in Education*. Cham, Switzerland: Springer, Jun. 2023, pp. 708–713.

[41] E. Tsutsumi, Y. Guo, and M. Ueno, "DeepIRT with a hypernetwork to optimize the degree of forgetting of past data," in *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. Durham, United Kingdom: International Educational Data Mining Society, July 2022, pp. 543–548.

[42] E. Tsutsumi, Y. Guo, R. Kinoshita, and M. Ueno, "Deep knowledge tracing incorporating a hypernetwork with independent student and item networks," *IEEE Transactions on Learning Technologies*, vol. 17, no. 1, pp. 951–965, Dec. 2023.

[43] E. Tsutsumi, T. Nishio, and M. Ueno, "Deep-IRT with a temporal convolutional network for reflecting students' long-term history of ability data," in *International Conference on Artificial Intelligence in Education*. Cham, Switzerland: Springer, Jul. 2024, pp. 250–264.

[44] M. Ueno and Y. Miyazawa, "Probability based scaffolding system with fading," in *Artificial Intelligence in Education*. Cham, Switzerland: Springer, 2015, pp. 492–503.

[45] ——, "IRT-based adaptive hints to scaffold learning in programming," *IEEE Transactions on Learning Technologies*, vol. 11, no. 4, pp. 415–428, Aug. 2017.

APPENDIX A DESCRIPTION OF THE PROPOSED ZDD CONSTRUCTION

The proposed ZDD construction requires the following inputs.

- Tuning parameter I_{th} represents the threshold value in Procedure 3.
- Finite set I represents a set of items in an item pool.
- Constant value parameter M denotes the number of items in equivalent item pool.
- Constant value parameter n stands for the number of items in the item pool.
- Constant value parameter Γ represents the number of discretized points for the test information function.
- Constant value parameters $I_{\text{LB}}(\theta_\gamma)$ and $I_{\text{UB}}(\theta_\gamma)$ respectively denote the lower and upper bounds of the test information function at the test score level θ_γ.

Using these inputs, Algorithm A provides a description of the proposed ZDD construction. The output of Algorithm A is the set $\mathcal{F}$, which represents equivalent item pools that satisfy Eq. (11) and Eq. (12).

```
procedure First stage
    Input: $I_{th}, I, n, M, \Gamma, I_{\text{LB}}(\theta_\gamma), I_{\text{UB}}(\theta_\gamma)$
    Output: $\mathcal{F}$
    Create a new node $v_{root}$
▷ root node
    $v_{\text{root}}.state.tl \leftarrow 0$
    $v_{\text{root}}.state.tis \leftarrow Array[\Gamma]$
▷ Declare an array of size $\Gamma$
    for $\gamma \leftarrow 1$ to $\Gamma$ do
        $v_{\text{root}}.state.tis[\gamma] \leftarrow 0$
    end for
    $V_1 \leftarrow \{v_{root}\}$
▷ $V_i$ is a set of nodes of depth $i$
    for $i \leftarrow 2$ to $n$ do
        $V_i \leftarrow \emptyset$
    end for
    $V_{n+1} \leftarrow$ {0-terminal node, 1-terminal node}
    for $i \leftarrow 1$ to $n$ do
        for each $v \in V_i$ do
            for each $x_i \in \{0, 1\}$ do
▷ 0-edge, 1-edge
                $\{i', state'\} \leftarrow$ Child$(i, M, v.state, x_i)$
▷ $i'$ is the depth of the child node. $state'$ is $tl$ and $tis$ of the child node.
                $v' \leftarrow$create a new node
▷ child node
                if $\{i', state'\}$ is $\{n+1, 0\}$ then
                    $v' \leftarrow$ 0-terminal node
                else if $\{i', state'\}$ is $\{n+1, 1\}$ then
                    $v' \leftarrow$ 1-terminal node
                else
                    $v'.state \leftarrow state'$
                    share_node $\leftarrow$ False
                    for each $w \in V_{i+1}$ do
                        if $v'.state.tl = w.state.tl$ then
                            for $\gamma \leftarrow 1$ to $\Gamma$ do
                                if $I_{th} \le |v'.state.tis[\gamma] - w.state.tis[\gamma]|$ then
```

```
                           next $w$
                        end if
                     end for
                     UpdateState($v'$, $w$)
                     $v' \leftarrow w$
▷ share node
                     share_node ← True
                     break
                  end if
               end for each
               if share_node is False then
                  $V_{i+1} \leftarrow V_{i+1} \cup v'$
               end if
            end if
            $v$.child[$x_i$]← $v'$
         end for each
      end for each
   end for
   $\mathcal{F} \leftarrow \mathrm{ReductionRule}(v_{root})$
▷ ReductionRule applies the two reduction rules to the constructed ZDD.
   Output $\mathcal{F}$
end procedure
procedure Child($i, M$, state, $x_i$)
   if $x_i = 1$ then
      $state'.tl \leftarrow state.tl + 1$
      for $\gamma \leftarrow 1$ to $\Gamma$ do
         $state'.tis[\gamma] \leftarrow state.tis[\gamma] + I_i(\theta_\gamma)$
▷ $I_i(\theta_\gamma)$ in eq(4)
      end for
   end if
   if $state'.tl = M$ then
      for $\gamma \leftarrow 1$ to $\Gamma$ do
         if not $I_{\mathrm{LB}}(\theta_\gamma) < state'.tis[\gamma] < I_{\mathrm{UB}}(\theta_\gamma)$ then
            return $\{n+1, 0\}$
▷ 0-terminal node
         end if
      end for
      Return $\{n+1, 1\}$
▷ 1-terminal node
   end if
   if $state'.tl + n - i < M$ then
      for $\gamma \leftarrow 1$ to $\Gamma$ do
         if $I_{\mathrm{UB}}(\theta_\gamma) < state'.tis[\gamma]$ then
            Return $\{n+1, 0\}$
▷ 0-terminal node
         end if
      end for
   end if
   Return $\{i+1, state'\}$
end procedure
procedure UpdateState($v'$, $w$)
   for $\gamma \leftarrow 1$ to $\Gamma$ do
      $w.state.tis[\gamma] \leftarrow (v'.state.tis[\gamma] + w.state.tis[\gamma])/2$
   end for
end procedure
```

APPENDIX B DESCRIPTION OF THE PROPOSED ZDD SAMPLING METHOD

The proposed ZDD sampling method requires the following inputs.

- Constant value time LT stands for the algorithm's total computation time limit.
- Finite set I represents a set of items in an item pool.
- Constant value parameter n stands for the number of items in the item pool.
- Constant value parameter Γ represents the number of discretized points for the test information function.
- Constant value parameters $I_{\mathrm{LB}}(\theta_\gamma)$ and $I_{\mathrm{UB}}(\theta_\gamma)$ respectively denote the lower and upper bounds of the test information function at the test score level θ_γ.
- Constant value parameter OC is the maximum number of common items between any pair of equivalent item pools.

Algorithm 1 provides a description of the proposed ZDD sampling method. The output of Algorithm 1 is the family of sets $\mathcal{P}$, which represents parallel test forms that satisfy all test constraints.

Algorithm 1: Second stage

```
procedure Second stage
   Input: $I_{th}, I, n, \Gamma, I_{\mathrm{LB}}(\theta_\gamma), I_{\mathrm{UB}}(\theta_\gamma), OC$
   Output: $\mathcal{P}$
   $st \leftarrow now()$
▷ $now()$ retrieves the current timestamp to track the elapsed computation time.
   $\mathcal{F} \leftarrow$ First stage $(I_{th}, I, n, \Gamma, I_{\mathrm{LB}}(\theta_\gamma), I_{\mathrm{UB}}(\theta_\gamma))$
   $\mathcal{P} \leftarrow \emptyset$
   while $(now() - st) < LT$ do
      $S \leftarrow RandomSampling(\mathcal{F})$
▷ random sampling subset $S$ from the constructed canonical ZDD $\mathcal{F}$
      for $\gamma \leftarrow 1$ to $\Gamma$ do
         $TI(\theta_\gamma) \leftarrow TestInfo(S, \theta_\gamma)$
▷ TestInfo calculates the test information $\sum_{i=1}^{n} I_i(\theta_\gamma)x_i, x_i \in S$ at test score level $\theta_\gamma$ from the binary variables in subset $S$.
         if $TI(\theta_\gamma) < I_{\mathrm{LB}}(\theta_\gamma)$ or $I_{\mathrm{UB}}(\theta_\gamma) < TI(\theta_\gamma)$ then
            Next while
         end if
      end for
      for each $P \in \mathcal{P}$ do
         if $\sum_{i \in I} x_i^S x_i^P > OC$ then
            Next while
         end if
      end for
      $\mathcal{P} \leftarrow \mathcal{P} \cup \{S\}$
   end while
   Output $\mathcal{P}$
```

```
23: end procedure
```

MAOMI UENO received a Ph.D. degree in computer science from the Tokyo Institute of Technology in 1994. He has served as a professor of the Graduate School of Information Systems at the University of Electro-Communications since 2013. He was conferred the Best Paper award at IEEE ICTAI2008. His interests include machine learning, data mining, Bayesian statistics, Bayesian networks, and educational technology. He is an IEEE member.

KAZUMA FUCHIMOTO received B.E. and M.E. degrees from the University of ElectroCommunications in 2020 and 2022, where he is currently pursuing a D.E. degree. His research interests include educational technology and computer science.

WAKABA KISHIDA received a B.E. degree from the University of ElectroCommunications in 2023, where she is currently pursuing an M.E. degree. her research interests include educational technology and computer science.

YOSHIMITSU MIYAZAWA received a Ph.D. degree in computer science from the University of Electro-Communications, in 2014. He has served as an associate professor of Research and Development at the National Center for University Entrance Examinations since 2019. His research interests include educational technology and computer science.

• • •