Probability-based scaffolding system using Sliding Hidden Markov IRT for longitudinal learning

Maomi Ueno$^{1[0000-0003-3598-8867]}$, Yoshimitsu Miyazawa$^{2[0009-0007-0773-0275]}$, and Emiko Tsutsumi$^{3[0000-0003-3338-8892]}$

[1] The University of Electro-Communications, Tokyo, Japan
ueno@ai.lab.uec.ac.jp
[2] The National Center for University Entrance Examinations, Tokyo, Japan
miyazawa@rd.dnc.ac.jp
[3] Hosei University, Tokyo, Japan
tsutsumi@hosei.ac.jp

Abstract. This study proposes a scaffolding system that provides adaptive hints using a new dynamic assessment probabilistic model, i.e.,Sliding Hidden Markov Item Response Theory (SHMIRT). The SHMIRT optimizes the degree of forgetting past data for the prediction of a student's performance by adjusting the student's ability change throughout the learning process. Using the SHMIRT, the system provides hints so that the student's correct response probability approaches 0.5 to each task even during long-term learning. We assess the causal effects of the proposed scaffolding mechanism using Inverse Probability Weighting (IPW) for long-term learning data in actual classes. The results demonstrate that the proposed system is effective.

Keywords: adaptive learning · scaffolding · dynamic assessment · learning analytics · causal inference · hint · Item Response Theory

1 Introduction

Vygotsky (1978) introduced the Zone of Proximal Development (ZPD), where a learner cannot solve difficulties alone, but can do so with an expert's help, to promote student development[28]. Bruner (1978), like Vygotsky, emphasized the social nature of learning, reporting that other people should help a child develop skills through a process designated as scaffolding[3]. He defined scaffolding as steps taken to reduce the degrees of freedom in carrying out some task so that children can concentrate on difficult skills. The term of scaffolding first appeared in the literature when Wood et al. (1976) described how tutors interacted with preschoolers to help them solve a block reconstruction problem [32]. Brown and Ferrara [2] and Campione [4] examined a ZPD-based assessment method, "dynamic assessment", by which a cascading sequence of hints (so-called "graded hints") is provided to enable dynamic assessment of how much support students need for completing various benchmark tasks. Each hint is staged in a graded fashion known as a cascading sequence of hints. A student is given a task to solve. If the student is unable to solve the task independently, then the student is given a series of graded hints, one after another, until the achievement is successful. The graded hints become increasingly concrete as the sequence is followed. Results demonstrated that students needing only a minimum number of hints to solve the tasks tended to achieve the greatest learning gain ([32, 2, 4]). Consequently, to scaffold a student efficiently, a teacher should predict how much support a student needs to complete a task. Then a teacher must determine the optimal degree of assistance which should be given to support the student's development [31]. To ascertain the optimal degree of assistance for student development, Ueno and Miyazawa proposed probability-based scaffolding, which assumes that optimal scaffolding is based on a probabilistic decision rule: given that, for a teacher's assistance, there exists an optimal probability of a student's correct answer to facilitate the student's development [24, 25]. Specifically, to predict a student's performance (correct answer probability) given hints, they first proposed an Item Response Theory (IRT) model for dynamic assessment, by which students are tested when given dynamic conditions of providing a series of graded hints. They then developed a scaffolding system that presented adaptive hints using the ability which was estimated using the IRT from the student's response data. To ascertain the optimal probability, they used the scaffolding system to compare the learning performance by changing the predictive probability. Results indicated that scaffolding to make the students' success probability 0.5 provided the best learning performance. Nevertheless, their experiments did not conclusively confirm the actual effectiveness for long-term learning because they used only seven tasks during a few hours. Especially, their system did not incorporate consideration of unique features in which the estimated ability was changed dynamically by learning. In short-term learning of a single skill as in this experiment, the improvement in a learner's ability is limited, and the learning system would still work effectively even if it assumed that learners ability do not change with their learning. However, in actual long-term learning, where a learner progresses while learning new various skills, a learner's ability changes significantly, and the system would incorrectly predict the learner's probability of answering a new task correctly and then can not provide an optimal scaffolding to the learner. Therefore, their system has only limited application for practical use. More recently, Bernd and Chounta(2024) employed Additive Factor Model (AFM) instead of IRT to predict a student's correct answer probability and analyze the relationship between his/her hint request behaviors and ZPD[19]. However, their analyses also did not take into account changes in a student's ability.

To realize a probability-based scaffolding for long-term learning, this study proposes a new IRT, Sliding Hidden Markov Item Response Theory (SHMIRT) that adapts to long-term learner ability changes, to predict a student's correct answer probability given hints. SHMIRT incorporates the sliding window Hidden Markov into IRT. Thereby, it optimizes the degree of forgetting past data for the prediction of a student's performance by adjusting the student's ability change throughout the learning process. With the proposed method, a window of specified length moves over the student's past response data. Subsequently, the ability is estimated solely from data within the window. The window size can be optimized easily using cross-validation because it can be searched from several discrete values. We developed a scaffolding system using SHMIRT to provide adaptive hints. The system provides hints so that the correct response probability of the student approaches 0.5 to each task using SHMIRT. The performance of the proposed system was assessed in an actual university course of "Discrete mathematics" for one semester. We assess the causal effects of the proposed scaffolding mechanism using Inverse Probability Weighting (IPW) by comparing those obtained using prior systems for the same course.

Results demonstrate the following. 1) SHMIRT improves the student performance prediction accuracy and enables the system to provide hints such that the predictive probabilities of a student's correct answer for tasks are approximately 0.5 throughout long-term learning. 2) The proposed system improves students' learning performance compared to that obtained using the earlier systems. Although many excellent scaffolding systems[1, 7, 11, 20] have been developed in recent years, there is no system with a fading function based on probabilistic-based scaffolding which adjusts a student's ability change throughout the learning process.

2 Data from Dynamic Assessment System

We developed the dynamic assessment system to obtain students' response data from tasks using a series of graded hints to apply IRT to dynamic assessment data.

Let $\{k\}$,$(k = 1, 2, \ldots, K-1)$ be a series of graded hints for task j. For that series, $k = 0$ when the task is presented without a hint. First, the dynamic assessment system in a computer presents task j without a hint to student i. If the student responds incorrectly, then the system presents hint $k = 1$. Otherwise, the system stores the student's response and presents the next task: $j+1$. If the student responds incorrectly to task j with hint $k = 1$, then the system presents hint $k = 2$. Alternatively, the system stores the student's response and presents the next task: $j + 2$. Consequently, the system presents hints from $k = 1$ to $k = K - 1$ until the student answers correctly. This procedure is repeated until $j = M$. After applying this procedure for N students, one obtains dynamic assessment data as

$$X = \{x_{ijk}\}, (i = 1, \cdots, N, j = 1, \cdots, M, k = 0, \cdots K),$$

where

$$x_{ijk} = \begin{cases} 1 : \text{student } i \text{ answered correctly to task } j \text{ when} \\ \quad\; k\text{-th hint or the previous hint before } k \text{ was presented} \\ 0 : \text{else other.} \end{cases}$$

Therein, x_{ijK} denotes the response data when student i cannot answer correctly with hint $K - 1$.

3 Item Response Theory for Dynamic Assessment

3.1 Sliding Hidden Markov Item Response Theory

This section proposes a new IRT for application to data X obtained in dynamic assessment: Sliding Hidden Markov Item Response Theory (SHMIRT). Various Hidden Markov IRT (HMIRT) models without consideration of hints have been proposed to adjust to the changing abilities of a student (e.g.[5, 6, 29, 30]). For accurate prediction throughout the longitudinal process, a key issue is the optimal degree of forgetting past response data. Actually, past response data might not reflect the current ability of the student accurately because the student ability has changed by learning. However, earlier HMIRTs [6, 29] did not consider the degree of forgetting past data. Although Wilson et al. (2016) [30] introduced a new HMIRT incorporating a new method of forgetting past data, it can not be used directly for dynamic assessment data.

To optimize the degree of forgetting past data, this study proposes Sliding Hidden Markov Item Response Theory (SHMIRT), by which a student's ability changes to follow the Sliding Hidden Markov process (Fig. 1). The sliding window method is used mainly in the fields of information technology, image processing, and voice recognition (e.g. [21]). When using this method, a window of specified length is moved over the student's past response data. Subsequently, the ability is estimated solely based on data within the window. In actuality, SHMIRT has a window size parameter L, which determines the degree of forgetting of past data. Window size L can be optimized easily using cross-validation because the optimal window size can be searched greedily from several discrete values. Furthermore, SHMIRT incorporates hint parameters into the model. Specifically, in SHMIRT, a random variable u_{jk} denotes the response of a student to task $j(1, \cdots, M)$ with k-th hint as

$$u_{jk} = \begin{cases} 1 : \text{a student answers correctly to task } j \text{ with } k\text{-th hint} \\ 0 : \text{else other.} \end{cases}$$

Inspired by Samejima's graded response model [18], we propose the probability $p(u_{jk} = 1 \mid \theta_{it})$ that student i answers task j with k-th hint correctly at time t

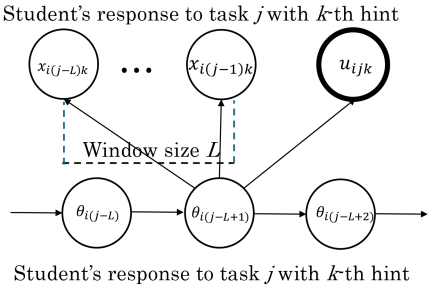


Fig. 1. Outline of SHMIRT.

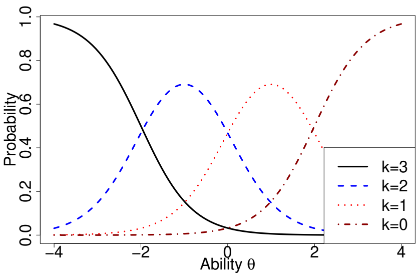


Fig. 2. Examples of IRCs for hints.

as

$$p(u_{jk} = 1 \mid \theta_{it}) = \frac{1}{1 + \exp(-a_j(\theta_{it} - b_{jk})} - \frac{1}{1 + \exp(-a_j(\theta_{it} - b_{j(k-1)})}. \quad (1)$$

Therein, $a_j \in (0, \infty), (j = 1, \cdots, M)$ stands for the j-th task's discrimination parameter expressing the discriminatory power for students' abilities of task j, $b_{jk} \in (-\infty, \infty), (k = 1, \cdots, K)$, which satisfies $b_{j1} > b_{j2} > \cdots, b_{jK}$, a difficulty parameter expressing the degree of difficulty of task j after the k-th hint is presented. Furthermore, $\theta_{it} \in (-\infty, \infty), (i = 1 \cdots, N, t = 1, \cdots, M - L)$ represents student i's ability at time point t. The prior distribution of θ_{it} is a normal distribution defined as

$$\theta_{it} \sim N(\theta_{it-1}, \sigma), \quad (2)$$

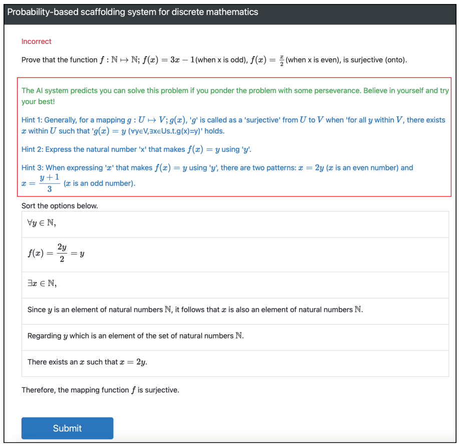
where σ is a variance parameter that is included to regulate ability changes. Consequently, this parameter avoids overfitting and thereby raises the student performance prediction accuracy. In addition, $\frac{1}{(1+\exp(-a_j(\theta_{it}-b_{j(-1)})))} = 0.0$, $\frac{1}{(1+\exp(-a_j(\theta_{it}-b_{jK})))} = 1.0$ and the initial value θ_{i0} for the prior distribution is zero.

Here, we simply assume a unidimensional ability variable which reflects the student development for a domain. In the case of no hints for a task and a fixed t, i.e., $K = 1$ and $k = 0$, equation (1) is identical to that of conventional IRT(2-parameters logistic model) [10]. In addition, when $K = 1$, $k = 0$, and $L = 1$, SHMIRT is equivalent to an earlier HMIRT model [29, 6].

As presented in Fig. 1, when the student has answered more than L tasks, the predictive correct answer probability of $p(u_{ijk} = 1 \mid \theta_{it})$ is estimated using the estimated ability $\widehat{\theta_{it}}$ and hint parameters a_j , b_{jk} stored in the database. Ability θ_{it} is estimated from the student's responses $\{x_{i(j-L)k}, \cdots, x_{i(j-1)k}\}$. The system provides hints so that the correct response probability of the student is 0.5 to each task using SHMIRT. Each time the student answers to the task, SHMIRT slides the window by one task and repeats the process presented above. Therefore, the ability vector of student i is represented as $\boldsymbol{\theta}_i = \{\theta_{i1}, \cdots, \theta_{i(M-L)}\}$. When the window size L is small, the ability depends only on the most recent response data. When the window size L is large, the ability depends on response data which are much farther in the past.

Fig. 2 depicts an example of item response curves (IRCs) in (1) for a task with two hints. The horizontal axis shows the student's abilities. The vertical axis shows the probability $p(u_{jk} = 1|\theta_{it})$ that student i will respond correctly to task j after the k-th hint is presented. The response curve with $k = 0$ represents the correct response probability given an ability for a task with no hint. The curves for $k = 1, 2$ represent the correct response probability given an ability after the k-th hint is presented. The curve for $k = 3$ shows the wrong response probability after all hints (k=1, 2) are presented.

For this study, we estimate student ability θ_{it}, variance σ, an item discrimination parameter a_j, and an item difficulty parameter b_j as the expected a posterior (EAP) using the Metropolis–Hastings method in the MCMC algorithm [13]. The prior distributions of respective parameters are set as $\theta_{i0} \sim N(0.0, 1.0)$,



Probability-based scaffolding system for discrete mathematics

Incorrect

Prove that the function $f : \mathbb{N} \mapsto \mathbb{N}$; $f(x) = 3x - 1$(when x is odd), $f(x) = \frac{x}{2}$(when x is even), is surjective (onto).

The AI system predicts you can solve this problem if you ponder the problem with some perseverance. Believe in yourself and try your best!

Hint 1: Generally, for a mapping $g : U \mapsto V; g(x)$, 'g' is called as a 'surjective' from U to V when 'for all y within V, there exists x within U such that '$g(x) = y$ (∀y∈V,∃x∈Us.t.g(x)=y)' holds.

Hint 2: Express the natural number 'x' that makes $f(x) = y$ using 'y'.

Hint 3: When expressing 'x' that makes $f(x) = y$ using 'y', there are two patterns: $x = 2y$ (x is an even number) and $x = \dfrac{y+1}{3}$ (x is an odd number).

Sort the options below.

$\forall y \in \mathbb{N},$

$f(x) = \dfrac{2y}{2} = y$

$\exists x \in \mathbb{N},$

Since y is an element of natural numbers $\mathbb{N}$, it follows that x is also an element of natural numbers $\mathbb{N}$.

Regarding y which is an element of the set of natural numbers $\mathbb{N}$.

There exists an x such that $x = 2y$.

Therefore, the mapping function f is surjective.

Fig. 3. Example of adaptive hints presented by the probability-based scaffolding system

$\theta_{it} \sim N(\theta_{it-1}, \sigma)$, $\sigma \sim IG(1.0, 1.0)$, $\log a_j \sim N(0.0, 0.2)$, $b_{jk} \sim N(\mu_{jk}, 0.4^2)$, and $\mu_{jk} = \frac{4}{K \times k} - 2$, with $IG(\alpha, \beta)$ representing the inverse gamma distribution with shape parameter α and scale parameter β, and $N(\mu, \sigma)$ denoting a normal distribution with expected value μ and variance σ. The posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N)$, $\mathbf{a} = (a_1, \cdots, a_M)$,$\mathbf{b} = (b_1, \cdots, b_M)$ is given as presented below.

$$p(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}, \sigma \mid \boldsymbol{X}, L) \propto L(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) p(\boldsymbol{a}) p(\boldsymbol{b}) p(\boldsymbol{\theta}) p(\sigma) \tag{3}$$
$$= \prod_{i=1}^{N} \left[\left[\prod_{t=0}^{M-L} \prod_{j=t+1}^{t+1+L} \prod_{k=0}^{K} p(u_{jk} = 1 \mid \theta_{it})^{x_{ijk}} \right] \left[\prod_{j=1}^{M} p(a_j) \cdot p(\boldsymbol{b}_j) \right] \left[\prod_{t=0}^{M-L} \prod_{j=1}^{N} p(\theta_{it} \mid \theta_{it-1}, \sigma) \right] p(\sigma) \right].$$

4 Probability-based Scaffolding System for Longitudinal Learning

4.1 Dynamic assessment in a discrete mathematics course

For learning SHMIRT, we obtained dynamic assessment data of a discrete mathematics course comprising 15 lectures of 90 min each. The participants were 426 first-year technical college students who took a discrete mathematics course during 2019–2021. The participants solved the presented tasks in the dynamic assessment system for discrete mathematics once a week after they took a course lecture. The system has 123 tasks. It presents 7–9 items with graded hints to a student after each lecture. Each task has 4–12 hints.

An example of a task with a hint is presented in Fig. 3. The tasks are proof problems. Students complete the proof by sorting the given options. First, the system presents a task for a student to solve. If the student is unable to solve the task independently, then the student is provided one of a series of graded hints, one after another, until the achievement is successful. The first hint presented the necessary prior knowledge to solve the task. For these tasks, the graded hints are designed to approach the final answer as the sequence is followed. Consequently, we obtained response data X from 426 examinees using the dynamic assessment system. The average correct response rate of all the tasks without hints is 0.28. The parameters of SHMIRT were estimated as the EAP using the MCMC method based on (3) from data X. For cross validation to determine the forgetting parameter L, we partitioned the 426 examinee data into 338 samples as a training dataset, 44 samples as a validation dataset, and 44 samples as a test dataset. Subsequently, the proposed method estimates window size L by incrementing the value by one from the initial value $L = 1$ to maximize the prediction accuracy. The prediction accuracy is calculated using the precision rate between the actual needed hint k_{ij} of the test data and the predicted hints by SHMIRT. The predicted hint $\widehat{k_{ij}}$ is given as

$$\widehat{k_{ij}} = \underset{k \in \{0,1,\cdots,K\}}{\arg\max}\ p(u_{jk} = 1 \mid \widehat{\theta_{it}}), \tag{4}$$

where $\widehat{\theta_{it}}$ represents the estimated θ_{it} as the EAP using MCMC method from past data $\{x_{j-L}, \cdots, x_{j-1}\}$ when $j > L$. When $j \leq L$, $\widehat{\theta_{it}}$ is estimated similarly to methods used for earlier studies [24, 25]. Consequently, we obtained the maximum average prediction accuracy 0.69 when L=9 with $\sigma = 0.10$ for the validation data. Similarly, we calculated the prediction accuracy for predicting $\widehat{k_{ij}}$ using the IRT in Ueno and Miyazawa[24, 25]. Consequently, we obtained the average prediction accuracy 0.60. Results demonstrate that SHMIRT improves the prediction accuracy of earlier method.

Fig. 4 presents a scatter plot illustrating the estimated values of the IRT and SHMIRT for parameter a. The horizontal axis denotes the estimates obtained using IRT, whereas the vertical axis denotes those of SHMIRT. Similarly, Fig. 5 depicts a scatter plot for the estimated values of the IRT and SHMIRT for parameter b. In both scatter plots, a reference line $y = x$ is included. The estimated a parameters of SHMIRT tend to be larger overall than those from the IRT. In contrast, for parameter b, SHMIRT assigns lower difficulty levels to the same items than IRT does. The ability estimates in SHMIRT tend to increase during the learning process. This increase leads to difference of the parameter estimates between those provided by IRT and by SHMIRT.

4.2 Probability-based scaffolding system for discrete mathematics

Using SHMIRT, we developed a scaffolding system to adjust a student's ability change over longer periods. Fig. 3 depicts an example of the proposed system by which adaptive hints are presented. First, the system presents the first task without hints. The tasks are presented according to the determined order in the

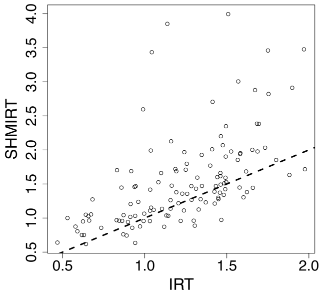


Fig. 4. Scatter plot of parameter a.

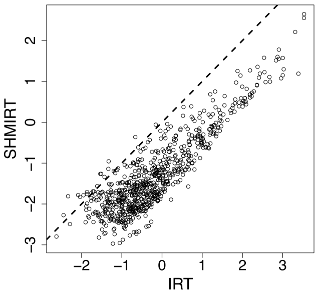


Fig. 5. Scatter plot of parameter b.

course, but those with the predicted correct answer probability greater than 0.5 are skipped so that they are not presented. If a student answers the presented task correctly, then the system estimates the student ability using the student response data. Then the system presents the next task. If the student answers incorrectly, then the system searches the hint database for a hint such that the student predictive correct answer probability is nearest to 0.5.

The predictive correct answer probability of $p(u_{jk} = 1 \mid \widehat{\theta_{it}})$ is estimated using the student's estimated ability $\widehat{\theta_{it}}$ and hint parameters a_j, b_{jk} stored in the database. The ability $\widehat{\theta_{it}}$ is estimated similarly in (4). Then, the system presents the selected hint to the student. If the student answers incorrectly to the task with a hint, then the system provides the correct answer and its explanation. It then presents the next task.

5 Practical Assessment of the System for Longitudinal Learning

5.1 Method

This section presents a description of assessment of the proposed scaffolding system in an actual university course, "Discrete mathematics" (the same course as that described in 3.3), which consists of 15 lectures for 90 min, for one semester in 2023. The participants were first-year technical college students who took a discrete mathematics course in 2023. This group was designated as 'Proposed'. After they took a course lecture once a week, they solved the presented tasks in the scaffolding system for discrete mathematics shown in Fig. 3. The system has 123 tasks. It presents 7–9 tasks with adaptive hints to a student after each lecture. Although each task has 4–12 hints, the system selects and presents only the optimal hints with which the student predictive correct answer probability is the nearest to 0.5, and only if the student answers the task incorrectly without receiving a hint.

We compare the learning performances of the proposed system with those obtained using the following earlier systems.

- Graded hints system: The system presents the graded hints sequentially in the same way as the method explained in section 3.3. System presents the next hint if the participant responds to the task incorrectly. This procedure is repeated until the participant responds correctly. If the participant responds incorrectly to the task when the final hint is presented, then the system presents the correct answer and its explanation. This system was assessed in the same way as the method which was proposed and examined for this study. The participants were first-year technical college students who took a discrete mathematics course in 2021.
- IRT-based hints system: The earlier scaffolding system [24, 25] was assessed in the same way as that for the method proposed for this study. The participants are first-year technical college students who took a discrete mathematics course in 2022.

All participants took pre-tests to assess their prior knowledge before they took the course. Each pre-test consisted of 31 basic mathematics problems. Using analysis of variance (ANOVA), we found no significant difference among the three participant groups. Furthermore, all participants took post-tests after they completed all the scaffolding system tasks. Each post-test consisted of new 25 problems combined with the previously learned knowledge. The maximum score of the post-test is 100.

5.2 Results

Evaluation of probability prediction accuracy As described in this section, we tested and confirmed that the systems presented adaptive hints so that the students' correct answer probabilities to the tasks were approximately 0.5. Fig. 6 depicts the correct answer rates of participants for each week when the selected hints were presented by the proposed system and by the IRT-based hints system[24, 25]. As shown in Fig. 6, although the prediction accuracies of IRT-based hints system become worse as learning proceeds, those of the proposed system maintain their higher prediction accuracies throughout the learning process because the proposed system adjusts a student's ability change over longer periods. By contrast, the abilities estimated using the IRT-based hints system tend to become greater than 0.5 as learning proceeds. The average of the correct answer rates for the proposed method was 0.53 ($\chi^2 = 2.53$, $p < 0.01$ of the χ^2 test). That for IRT-based hints was 0.61 ($\chi^2 = 10.38$, $p < 0.11$), thereby demonstrating that the predictive correct answer probabilities by the proposed system are equivalent to 0.5 with a significance level of 1%, but those by the IRT-based hints system are not. Fig. 7 depicts the moving average numbers of hints (k_{ij}) for a sequence of 10 tasks provided by the three systems to participants. The horizontal axis represents the number of presented tasks. The vertical axis of the left side shows the moving average of k_{ij}. That of the right side presents the difficulty parameters of presented tasks. The average numbers of hints throughout overall period, provided by the graded hints system, the IRT-based hints system, and the proposed system were, respectively, 1.72 (0.68), 1.46 (0.51), and

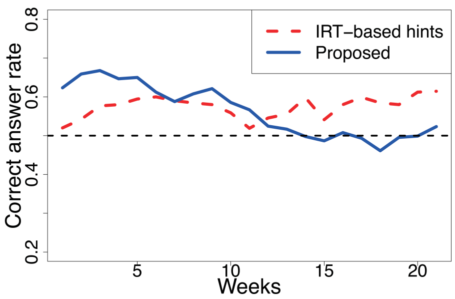


Fig. 6. Correct answer rates transition for each week

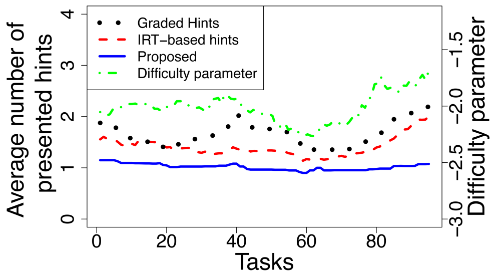


Fig. 7. Average number of presented hints and difficulty parameters of presented tasks.

Table 1. Post-Test Results (maximum score is 100)

Group	Graded hints	IRT-based hints	Proposed
No. examinees	31	30	39
Avg. Pre-test	34.00(12.24)	50.00(19.07)	45.90(49.90)
GPA	2.23(0.48)	2.40(1.14)	2.45(1.26)
Avg. Post-test	57.03**(8.92)	61.18**(12.79)	69.33(11.90)
IPW-Adjusted Avg. Post-test	58.75**(7.21)	60.48**(12.51)	74.99(13.54)
ATE	-7.46**(1.97)	-7.45**(1.98)	10.43(1.83)
Avg. Difference abilities	0.116*(0.465)	0.120*(0.423)	0.373(0.280)
Avg. Attempt	1.814*(0.349)	1.538(0.221)	1.352(0.191)
Avg. Response time (s)	243**(219)	230**(215)	305(377)

Significant difference from the proposed method: *5%, **1%) .

1.02 (0.23) (the values in parentheses represent the standard deviation). The proposed system clearly provides fewer hints than the other systems do. Pea (2004) pointed out that a fading function is a necessary feature for scaffolding system[14]. The graded hints system and the IRT-based hints system increase the average number of hints throughout the learning process. Although it is known that this phenomenon is caused by over-instruction [24, 25], the main reason is that the tasks presented later become more difficult as shown in Fig.7. On the other hand, the proposed system slightly reduces the average number of hints throughout the learning process. These results demonstrate that the system selects and presents only the optimal hints with which the student predictive correct answer probability is the nearest to 0.5.

Evaluation of learning performance This section presents the main evaluation of the proposed system. Post-test results are presented in Table 1, which lists the number of examinees (designated as 'No. examinees') who completed the experiments in each group, the average score from pretests (designated as 'Avg. Pre-test') , the average score of grade point averages (designated as 'GPA'), and average score from post-tests (designated as 'Avg. Post-test').

The values in parentheses in the table are standard deviations. Table 1 also includes the average differences of the SHMIRT-based estimated abilities obtained for the first attempt and the last attempt to each system for the three groups, respectively (designated as 'Avg. Difference abilities'), the average of the attempts to a task in each group (designated as 'Avg. Attempt'), and the average task response time (s) in each group (designated as 'Avg. Response time'). We assessed differences among the groups using one-way ANOVA for results obtained from post-tests. Then we applied the Tukey–Kramer method to evaluate the detected differences. The proposed system performed better than the others for Avg. Post-test, as shown in Table 1, with the relevant significance levels. Nevertheless, this difference cannot be strictly derived because the participants assignments were not randomized for each year experiments. Therefore, we employ propensity score-based covariate adjustment [16], including several variables as covariates, to compare the post-test results. The propensity score is a method proposed by Rosenbaum and Rubin (1983) for estimating causal effects in observational studies where random assignment is not feasible[15]. For this study, the propensity score e_{Ci} for participant i is estimated using the following logistic regression model.

$$e_{Ci} = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_8 x_{8i} + \beta_9 x_{9i})\}} \quad (5)$$

In equation (5), $x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}$, and x_{9i} represent the pre-test score, the university GPA, and grades from seven courses: Calculus I, Calculus II, Linear Algebra I, Linear Algebra II, Mathematics Practice I, Mathematics Practice II, and Mathematical Analysis, which all participants have taken. We determined the generalized propensity scores for each comparison method using logistic regression analysis[8]. To account for potential differences in participants' abilities across different years, we obtained the adjusted average post-test by IPW (designated as 'PW-Adjusted Avg. Post-test') as shown in Table 1. The results demonstrate that the proposed method provides highest learning performance by a significant margin. Additionally, the average treatment effect (ATE) obtained when comparing each method with the other methods was estimated [9, 12, 17]. The results, as presented in Table 1, indicate that the proposed method improves the post-test scores significantly compared to other methods with improvement of 10.43 points. Although the average score of post-test for IRT-based hints system is greater than that of the graded hints system, the ATE results for the two systems are almost identical. The IRT-based hints system can not select the optimal hints to a student accurately for long-term learning. because that hint system does not consider a student's change in ability.

Table 1 also demonstrates that the proposed system performed better than the others for Avg. Difference abilities, with the relevant significance levels. This causes that the proposed system provides fewer hints than the other systems do, as shown in Fig.7.

Another interesting finding is that the results obtained from the proposed system exhibit the least average number of attempts for a task with the longest average time to solve a task. The fact that these learners abilities have improved significantly in Table 1 and the results of the questionnaire in Table 2 which will be described next justify longer engagement times of the learners means active participation by the learners, and not engaging in unproductive behaviors but contributing to learning. Therefore, participants who used the proposed system tended to ponder over a problem for a longer time.

Question analyses We also posed the two questions in Table 2 to participants who used IRT-based hints and the proposed system. Participants answered them by responding using a five-point Likert scale for questions 1) and 2): 1. Strongly disagree, 2. Weakly disagree, 3. I am not sure, 4. Weakly agree, and 5. Strongly agree. Table 2 presents the average scores to the questionnaires for participants who used IRT-based hints and the proposed system. The values in parentheses

Table 2. Average scores and standard deviation from five-point Likert scale questions

	IRT-based hints	Proposed
Question 1: Do you think that you found the correct answers for the tasks using only minimal assistance from the system?	2.67(0.82)	3.50(0.84)
Question 2: Do you think that you tried to ponder the problem	1.67*(0.82)	3.17*(0.41)

(t-test and significant difference: *5%,) in the table represent standard deviations. The results of question 1 demonstrate that the proposed system's average score is higher than that of IRT-based hints, but no significant difference was found. The results of question 2 demonstrate that the proposed system is significantly more effective at facilitating deep thinking than IRT-based hint system is. These results justify that longer engagement times of the learners for the proposed system shown in table 1 were due to their active, autonomous, and deep thought process towards the task. It is superior in this regard because it selects and presents only the optimal hints with which the student predictive correct answer probability is the nearest to 0.5. By contrast, IRT-based hint systems tend to present hints for which the student predictive correct answer probability is higher than that of the proposed system, which leads to over-assistance, as shown in Fig. 6. Consequently, the proposed system facilitates student pondering of problems by themselves, with minimal assistance from the system.

6 Conclusions

We proposed a scaffolding system that provides adaptive hints using a new IRT, Sliding Hidden Markov Item Response Theory (SHMIRT), with adjustment to changes in student ability that have occurred over a longer period of time. The SHMIRT adjusts to a student's changes in ability over longer periods. The proposed system provides hints so that the student's correct response probability approaches 0.5 to each task using SHMIRT. The performance of the proposed system was assessed in an actual university course of "Discrete mathematics" for one semester. Using IPW, we assessed the causal effects of the proposed scaffolding mechanism by comparing those obtained using prior systems for the same course. Results demonstrated the following. 1) SHMIRT improved the accuracy of student performance prediction and enabled the system to provide a hint so that the predictive probability of the student's correct answer approached 0.5 throughout long-term learning. 2) The proposed system selected and presented only minimal hints, by contrast to earlier systems that tended to provide more numerous hints, which led to over-instruction. Consequently, the proposed system improved students' learning performance compared to that supported by the earlier systems.

Our work has the following limitations that should be considered. We used a dataset from only one actual university course. Accordingly, the effectiveness of the proposed system depends on the participants characteristics, the subject, quality of hints, and other factors. Therefore, even though we controlled for potential covariates using IPW, there might still be unobserved variables that influenced the results. In addition, we intend to add adaptive problem selection function[26, 27, 23, 22] to the scaffolding system so that the learner's correct answer probabilities for the presented tasks without a hint are less than 0.5. Our future task is to solve these problems to obtain more general results.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number JP24H00739.

References

1. Adair, A., Pedro, M.S., Gobert, J., Segan, E.: Real-time ai-driven assessment and scaffolding that improves students' mathematical modeling during science investigations. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) AIED 2023. LNCS(LNAI), vol. 13916. pp. 202–216. Springer Nature Switzerland, Cham (2023)
2. Brown, A.L., Ferrara, R.A.: Diagnosing zones of proximal development. Culture, communication, and cognition: Vygotskian perspectives pp. 273–305 (1985)
3. Bruner, J.S.: The role of dialogue in language acquisition. New York, Springer–Verlag (1978)
4. Campione, J.C.: Assisted assessment: A taxonomy of approaches and an outline of strengths and weaknesses. Journal of Learning Disabilities **22**(3), 151–165 (1989). https://doi.org/10.1177/002221948902200303
5. Dylan Molenaar, Daniel Oberski, J.V., Boeck, P.D.: Hidden markov item response theory models for responses and response times. Multivariate Behavioral Research **51**(5), 606–626 (2016). https://doi.org/10.1080/00273171.2016.1192983
6. Gonzalez-Brenes, J., Huang, Y., Brusilovsky, P.: General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In: The 7th International Conference on Educational Data Mining. pp. 84–91 (2014)
7. Gupta, A., Carpenter, D., Min, W., Mott, B., Glazewski, K., Hmelo-Silver, C.E., Lester, J.: Enhancing stealth assessment in collaborative game-based learning with multi-task learning. In: Wang, N., Rebolledo-Mendez, G., Matsuda, N., Santos, O.C., Dimitrova, V. (eds.) AIED 2023. LNCS(LNAI), vol. 13916. pp. 304–315. Springer Nature Switzerland, Cham (2023)
8. Imbens, G.: The role of the propensity score in estimating dose–response functions. Biometrika **87**(3), 706–710 (2000). https://doi.org/10.1093/biomet/87.3.706
9. Joffe, M.M., Have, T.R.T., Feldman, H.I., Kimmel, S.E.: Model Selection, Confounder Control, and Marginal Structural Models. The American Statistician **58**(4), 272–279 (2004). https://doi.org/10.1198/000313004x5824
10. van der Linden, W.J.: Handbook of Item Response Theory: Volume 3: Applications. Chapman and Hall/CRC (2019)
11. Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., Paquette, L.: Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. Journal of Computer Assisted Learning **39**(2), 351–368 (2023). https://doi.org/10.1111/jcal.12761
12. Olmos, A., Govindasamy, P.: A Practical Guide for Using Propensity Score Weighting in R. Practical Assessment, Research and Evaluation **20** (2015)
13. Patz, R.J., Junker, B.W.: Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics **24**(4), 342–366 (1999). https://doi.org/10.3102/10769986024004342
14. Pea, R.D.: The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. Journal of the Learning Sciences **13**(3), 423–451 (2004). https://doi.org/10.1207/s15327809jls1303_6
15. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983). https://doi.org/10.1093/biomet/70.1.41
16. Rubin, D.B.: The use of propensity scores in applied Bayesian inference. Bayesian Statistics 2 pp. 463–472 (1985)

17. Rubin, D.B.: The practical importance of understanding placebo effects and their role when approving drugs and recommending doses for medical practice. Behaviormetrika **47**, 5–18 (2020). https://doi.org/10.1007/s41237-019-00091-7
18. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika **34**(1), 1–97 (1969). https://doi.org/10.1007/BF03372160
19. Schulze Bernd, N., Chounta, I.A.: Beyond the grey area: Exploring the effectiveness of scaffolding as a learning measure. In: Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds.) AIED 2024. LNCS(LNAI), vol. 14829. pp. 365–378. Springer Nature Switzerland, Cham (2024)
20. Tamang, L.J., Banjade, R., Chapagain, J., Rus, V.: Automatic question generation for scaffolding self-explanations for code comprehension. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) AIED 2022. LNCS, vol. 13355. pp. 743–748. Springer International Publishing, Cham (2022)
21. Tao, Y., Papadias, D.: Maintaining sliding window skylines on data streams. IEEE Transactions on Knowledge and Data Engineering **18**(3), 377–391 (2006). https://doi.org/10.1109/TKDE.2006.48
22. Ueno, M., Fuchimoto, K., Kishida, W., Miyazawa, Y.: Computerized adaptive testing to balance exposure bias and measurement accuracy using zero-suppressed binary decision diagrams. IEEE Access **13**, 33883–33903 (2025). https://doi.org/10.1109/ACCESS.2025.3543554
23. Ueno, M., Fuchimoto, K., Tsutsumi, E.: e-Testing from artificial intelligence approach. Behaviormetrika **48**(2), 409–424 (2021)
24. Ueno, M., Miyasawa, Y.: Probability based scaffolding system with fading. In: Conati, C., Heffernan, N., Mitrovic, A., Verdejo, M.F. (eds.) AIED 2015. LNCS(LNAI), vol. 9112. pp. 492–503. Springer International Publishing, Cham (2015)
25. Ueno, M., Miyazawa, Y.: IRT-based adaptive hints to scaffold learning in programming. IEEE Transactions on Learning Technologies **11**, 415–428 (2018). https://doi.org/10.1109/TLT.2017.2741960
26. Ueno, M., Miyazawa, Y.: Uniform adaptive testing using maximum clique algorithm. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) AIED 2019. LNCS(LNAI), vol. 11625. pp. 482–493. Springer International Publishing, Cham (2019)
27. Ueno, M., Miyazawa, Y.: Two-stage uniform adaptive testing to balance measurement accuracy and item exposure. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds.) AIED 2022. LNCS, vol. 13355. pp. 626–632. Springer International Publishing, Cham (2022)
28. Vygotsky, L.S.: Mind in society. Cambridge, MA: MIT Press (1978)
29. Wang, X., Berger, J.O., Burdick, D.S.: Bayesian analysis of dynamic item response models in educational testing. The Annals of Applied Statistics **7**(1), 126–153 (2013). https://doi.org/10.1214/12-aoas608
30. Wilson, K.H., Karklin, Y., Han, B., Ekanadham, C.: Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation. In: 9th International Conference on Educational Data Mining. vol. 1, pp. 539–544 (2016)
31. Wood, D.: Scaffolding contingent tutoring and computer-supported learning. International Journal of Artificial Intelligence in Education **12**, 280–292 (2001)
32. Wood, D.J., Bruner, J.S., Ross, G.: The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry **17**, 89–100 (1976)