International Journal of Distance Education Technologies
Volume 22 • Issue 1

Evaluating the Effectiveness of Bayesian Knowledge Tracing Model-Based Explainable Recommender

Kyosuke Takami, Education Data Science Center, National Institute for Educational Policy Research, Tokyo, Japan*
https://orcid.org/0000-0002-0913-4641

Brendan Flanagan, Center for Innovative Research and Education in Data Science, Kyoto University, Kyoto, Japan
https://orcid.org/0000-0001-7644-997X

Yiling Dai, Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan
https://orcid.org/0000-0001-9900-8763

Hiroaki Ogata, Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

ABSTRACT

Explainable recommendation, which provides an explanation about why a quiz is recommended, helps to improve transparency, persuasiveness, and trustworthiness. However, little research examined the effectiveness of the explainable recommender, especially on academic performance. To survey its effectiveness, the authors evaluate the math academic performance among middle school students (n=115) by giving pre- and post-test questions based evaluation techniques. During the pre- and post-test periods, students were encouraged to use the Bayesian Knowledge Tracing model based explainable recommendation system. To evaluate how well the students were able to do what they could not do, the authors defined growth rate and found recommended quiz clicked counts had a positive effect on the total number of solved quizzes (R=0.343, P=0.005) and growth rate (R=0.297, P=0.017) despite no correlation between the total number of solved quizzes and growth rate. The results suggest that the use of an explainable recommendation system that learns efficiently will enable students to do what they could not do before.

KEYWORDS

Bayesian Knowledge Tracing, Effectiveness, Explainable recommendation, K-12 mathematics, Pre-post tests

INTRODUCTION

Artificial intelligence (AI) in education has enabled the development of e-learning systems that simulate students' knowledge and experience to provide personalized support to students (Nwana, 1990; Self, 1974; Wenger, 2014). AI-supported e-learning refers to the use of AI techniques (e.g., fuzzy logic, decision tree, Bayesian networks, neural networks, genetic algorithms, and hidden Markov models)

DOI: 10.4018/IJDET.337600
*Corresponding Author

in e-learning (i.e., using computer and network technologies for learning or training) (Colchester et al., 2017). A recent meta-review by Kaudi et al. (2021) reported that that the most identified AI-supported e-learning systems were adaptive learning systems and the second most identified kind of AI-enabled learning systems were intelligent tutoring systems, with the recommendation system being the least reported. Some of the recommender systems for personalized learning adopted collaborative filtering (Chen & Cui, 2020; Wind et al., 2018), content-based filtering (Kandakatla & Bandi, n.d.; Lops et al., 2011), and knowledge-based filtering (Haddad & Naser, 2017; Samin & Azim, 2019). These methods are commonly used in recommendation systems, but it is difficult to describe them as AI-supported learning systems since they do not employ Bayesian networks or neural networks. Therefore AI-supported recommendation systems in the educational field have not been well studied.

On the other hand, when we shift our focus to our daily lives, it is clear that recommender systems are everywhere, and AI is being used here as well. For example, Amazon recommends products with collaborative filtering (Smith & Linden, 2017), and Netflix recommends movies using deep learning (Amatriain & Basilico, 2015). In the e-commerce research field of recommendation, explainable recommendations, which provide explanations about why an item is recommended, have received much attention for improving transparency, persuasiveness, and trustworthiness (Zhang & Chen, 2020). Based on these studies, also in education, it is supposed that explanations from a learning system could provide additional benefits for students. Previous research on intelligent tutoring systems has shown that student motivation in system-based self-regulated learning can be improved by prompting and feedback mechanisms, leading to higher achievement (Duffy & Azevedo, 2015). Further, eXplainable AI (XAI) has begun to attract attention in the field of education for emerging concerns about Fairness, Accountability, Transparency, and Ethics (Khosravi et al., 2022). Explanations interpreting the decision-making process of AI are very important for teachers because they must be accountable to students, parents, or governments. Teachers need to know why such feedback was given by the AI, and interpreting why it was given may help teachers improve their teaching skills.

Some explainable recommendation research has been carried out in the field of education: Wikipedia recommendation in learning textbooks (Rahdari et al., 2020), recommendation in programming classes (Barria-Pineda et al., 2021) (both for higher education), and cognitive training for primary or secondary school children (Tsiakas et al., 2020). Various methods of explaining recommendations have been proposed using two different approaches: model-intrinsic and post-hoc approach. In model-intrinsic, rule-based (Conati et al., 2021), keyword-based (Yu et al., 2021), and concept-based (Dai et al., 2022) were proposed to generate explanations. Takami et al. (2021, 2022) proposed methods to generate explanations from the parameters in a learners' knowledge tracing model. Barria-Pineda et al. (2021) adapted a post-hoc approach and combined a concept-based model.

Although various forms of explainable recommenders have been proposed, it is under-explored how effective the explanations of recommended quizzes are, especially on academic performance in a practical school learning environment. Recently, we developed an explanation generator using the parameters from Bayesian knowledge tracing (BKT) models (Takami et al., 2021). In this explanation generator, recommended quizzes were categorized into different feature types according to the values of the model parameters and explanation texts (i.e., "You're not getting the basic skills. Let's go over the basics with this quiz!" or "Watch out for careless mistakes!", etc.) are generated based on these feature types (more details are explored in further sections as well as the Appendix). We reported that comparing the click counts of recommended quizzes with and without explanations of why the quizzes were recommended, the number of clicks was significantly higher for quizzes with explanations in high school mathematics learning (Takami et al., 2022). In the post-experiment student perception survey, the percentage of those convinced by recommended quizzes with explanation was higher than without explanation, and on the question of trust in the system, there were fewer negative answers in the explained recommender group than in the unexplained recommender group. These results indicated the importance of explanation for the recommender system in education. In this study, we used this educational explainable recommendation system to investigate the effects of explanation to academic performance by pre and post-tests to measure academic performance also investigated how the recommendation function can be used to improve academic performance in terms of how much one has improved in solving previously unsolvable problems. Therefore, the research questions were as follows:

1. Does the BKT-based explainable recommender improve knowledge retention?
2. What are the patterns of effective use of the BKT-based explainable recommender for knowledge retention?

BACKGROUND

Knowledge Tracing

Knowledge Tracing (KT) is the task of modeling learner knowledge over time to predict how learners will perform in future interactions (Piech et al., 2015). Its operational process involves collecting data from a learner's performance, such as the correct or wrong responses during practice, or their actions, like the time spent on a question. These data are then utilized to infer the learner's underlying, unobservable characteristics, which may include knowledge, objectives, preferences, and motivational state, among other factors (Gong et al., 2010). KT in e-learning environments has some important advantages; for example, prediction learner's performance (Yang & Cheung, 2018), assisting the learner's needs (Zhang et al., 2020), maintaining learners' motivation during the learning process (Zou et al., 2020), and improving learner's learning efficiency (Shen et al., 2021). The most well-known techniques for KT are Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994) and deep knowledge tracing (Piech et al., 2015). BKT can predict learner performance in e-learning systems (Qiu et al., n.d.). These knowledge tracing methods can help students stay motivated during the learning process as they increase their self-motivation in learning and achieve personalized support by automatically detecting their weak knowledge points (Sou et al., 2020). KT has the potential to accomplish this by estimating a student's underlying hidden qualities, such as knowledge, goals, preference, and motivational state based on observations of the student's performance or activities in the context of AI-supported solutions (Ilić et al., 2023).

Recommender Systems in Education

Recommender systems could be defined and perceived as systems that aim to provide specifically tailored recommendations according to individual users preferences (Zhang et al., 2022). Recommender systems in e-learning environments have been developed using three basic techniques: collaborative filtering (Chen & Cui, 2020; Wind et al., 2018), content-based filtering (Kandakatla & Bandi, n.d.; Lops et al., 2011), and knowledge-based filtering (Haddad & Naser, 2017; Samin & Azim, 2019). In addition to these techniques, methods utilizing machine learning approaches such as association rules with content-based collaborative filtering (Xiao et al., 2018), sequential pattern mining with knowledge-based filtering (Chen et al., 2014), and deep learning (Bhatt et al., 2023) have been proposed, but recommender systems for e-learning are lacking especially using AI technology (Kabudi et al., 2021); with few developed and implemented, there is still insufficient research to thoroughly examine their effectiveness and other aspects.

Explainable AI in Education

Recently eXplainable AI (XAI) has begun to attract attention in the field of education for emerging concerns about Fairness, Accountability, Transparency, and Ethics (FATE) (Khosravi et al., 2022). XAI is one of the emerging methods for increasing trust in AI systems, which promotes the use of methods that "enable human a user to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning, 2017, page 7). Interpreting the decision process of the model and thereby providing explanations can be expected to have a positive impact on students' academic performance by improving their sense of conviction and increasing their confidence in the AI. For the teacher, by presenting an explanation of why the question was recommended to improve academic performance, it shows what explanatory methods are useful for improving academic performance.

Explainable Recommendation

Explainable recommendation has been well studied in the area of recommender systems for improving transparency, persuasiveness, and trustworthiness of users in e-commerce such as Netflix or Amazon (Nunes & Jannach, 2017). This may be due to the fact that it is possible to increase the benefit from a product recommendation system in an explainable way, but explanations in the field of education can be expected not only to improve performance but also to increase motivation to learn since previous research has shown that feedback from a system can increase achievement and motivation (Duffy & Azevedo, 2015).

The Way of Generating Explanation for Recommendation

Recommendation explanations can be generated from different data sources and provided in different display styles (Tintarev & Masthoff, 2015) (e.g., a relevant user or item, a sentence, an image, or a set of reasoning rules). Basically, there are two approaches to generating explanations in recommender systems: model-intrinsic and post-hoc (Zhang & Chen, 2020). In the model-intrinsic approach, the model's mechanism is transparent and the explanation explains exactly how the model generates a recommendation. To this end, the processes of generating recommendations and generating explanations are mutually dependent. In this model-intrinsic approach, the goal of being explainable sometimes can constrain the model from being complex and "deep". For example, deep learning-based knowledge tracing, represented as deep knowledge tracing (DKT) (Piech et al., 2015), to model the knowledge state using recurrent neural network and other side information achieved better prediction accuracy compared to ordinary Bayesian knowledge tracing-based approaches (Su et al., 2018; Wang et al., 2019; Yeung & Yeung, 2018). Deep learning-based approach achieved state-of-the-art accuracy in knowledge state prediction, but it models the relation between the sequential learning activities and the knowledge state implicitly, so it is difficult to interpret the decision process in the model.

In contrast, the post-hoc approach generates the explanation after a recommendation is generated (e.g., providing simple statistical information like "70% of your friends bought this item"), but the explanations by post-hoc does not mean that they are fake; they are just decoupled from the model. As a result, the model is allowed to be a "black box", and the explanation does not necessarily explain why an item is recommended based on the recommender model. In the educational research context, several methods for generating explanations were proposed. In model-intrinsic, rule-based (Conati et al., 2021), keyword-based (Yu et al., 2021), and concept-based (Dai et al., 2022) were proposed to generate explanations. Takami et al. (2021, 2022) proposed methods to generate explanations from the parameters guess and slip in Bayesian knowledge tracing model for mathematics learning systems. Barria-Pineda et al. (2021) adapted a post-hoc approach and combined a concept-based model for explainable recommendations in personalized programming practice systems.

Subject Areas Addressed in AIED Research Articles

In previous Artificial Intelligence in Education (AIED) research areas, there have been a variety of studies on mathematics subjects. Preschools math (Gulz et al., 2020), mathematical instruction (Kelly et al., 1993), metacognitive scaffolding for learning by teachable agent (Matsuda et al., 2020), teachable agent in chat system (Tärning et al., 2019), and personalizing algebra to students' individual interests in an intelligent tutoring system: moderators of impact (Walkington & Bernacki, 2019), modeling and predicting the active video-viewing time in a large-scale e-learning system (Xie et al., 2017). AI supported personalized learning systems could analyze students' data on performance, preferences, and other factors to create customized learning paths and provide targeted support and feedback (Tapalova & Zhiyenbayeva, 2022). There have also been proposals for eXplainable AI for personalized learning, leveraging students' pen strokes, text data, and data on correctness or incorrectness related to self-explanation (Ogata et al., 2023). While there have been many AI-supported learning studies in mathematics, there have been few studies that recommend quizzes by AI and provide further reasons for the recommendation.

In this study, we aim to estimate the learner's knowledge state using the BKT method for quiz recommender. Furthermore, they intend to develop a system that not only provides recommendations for quizzes based on this knowledge state but also incorporates explanations about why these recommendations are made, utilizing the internal parameters of BKT. We validate the effectiveness of this system with actual middle school students.

METHOD

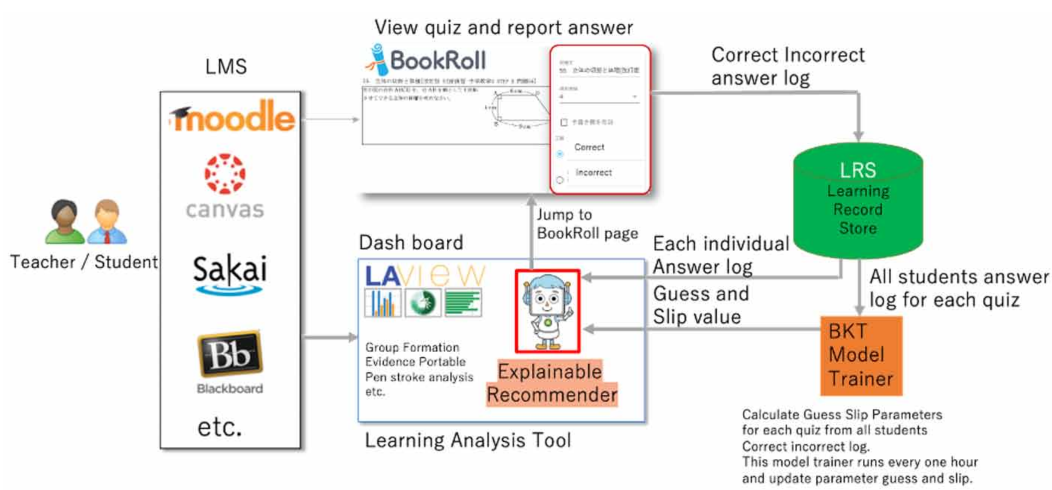
System Overview

The explainable recommender system in this paper is built on a learning system that was developed to support the distribution of learning materials, collection, and automated analysis of learning behavior logs in an open and standards-based approach (Flanagan & Ogata, 2018), as shown in Figure 1. The main components of the framework are: Moodle LMS, which acts as a hub for accessing various courses; the BookRoll reading system for learning material and quiz exercise distribution; an LRS for collecting learning behavior logs from all of the components; and the LAView learning analytics dashboard to provide feedback to students, teachers, and school administrators. This framework enables us to collect and analyze learning behaviors in real time and provide feedback to stakeholders. Quiz books used in the mathematics classes were uploaded to the reading system, and multiple-choice quiz questions were created to enable the collection of answers in the learning log data.

Recommender Using Bayesian Knowledge Tracing Model

The Bayesian knowledge tracing (BKT) model (Corbett & Anderson, 1994) has been used to model student knowledge by calculating the probability that a student knows a skill at a given point in time. This model is used in various educational systems, including tutors for reading skill (Beck & Chang, 2007), computer programming (Corbett & Anderson, 1994), and mathematics (Koedinger, 2002). We

Figure 1. Explainable recommender system architecture



applied a BKT model for their quiz recommender system. First, the guess (giving a correct answer despite not knowing the skill) and slip (knowing a skill but giving a wrong answer) parameters of BKT for each question are calculated from the data of correct and incorrect answers for all questions for all students in the relevant course using the Python Library of Bayesian knowledge tracing models (Badrinath et al., 2021). The parameters of guess and slip are estimated by the BKT-model trainer, as seen in Figure 1, which calculates the parameters using all logs of the course every hour and updates them every hour. Next, using the parameters of guess and slip for each question obtained from the data of all students, each student's individual probability of correct answers for each question is calculated from the following equation, which are individualized Bayesian knowledge tracing models (Yudelson et al., 2013). L is the probability of knowing the knowledge beforehand. T is the probability of acquiring knowledge from an unacquired state. Guess and slip value are estimated as a parameter inside BKT. The conditional probability is used to update the probability of skill mastery according to Equation (3).

$$P(L_{n-1} \mid Correct_n) = \frac{P\left(L_{n-1}\right) * \left(1 - P\left(slip\right)\right)}{P\left(L_{n-1}\right) * \left(1 - P\left(slip\right)\right) + \left(1 - P\left(L_{n-1}\right)\right) * \left(P\left(guess\right)\right)} \tag{1}$$

$$P(L_{n-1} \mid Incorrect_n) = \frac{P\left(L_{n-1}\right) * P\left(slip\right)}{P\left(L_{n-1}\right) * \left(P\left(slip\right)\right) + \left(1 - P\left(L_{n-1}\right)\right) * \left(1 - P\left(guess\right)\right)} \tag{2}$$

$$P(L_n \mid Action_n) = P(L_{n-1} \mid Action_n) + (1 - P(L_{n-1} \mid Action_n)) * P\left(T\right) \tag{3}$$

We set a 50% correct rate as a sweet spot of difficulty – not so hard that students are discouraged but not so easy either; this is called the "Goldilocks zone" (Kidd et al., 2012). Subtracted 0.5 from the correct answer rate P so that those closest to 0 would be recommended, meaning the closer the percentage of correct answers to 0.5 the more likely they are to be recommended, but this method does not take into account the order in which they are learned. Mathematics is a logical subject, and the order of learning is important. For this reason, the following weighting was used so that the smaller the problem number, the more likely it is to be recommended.

$$w_k^2 = \left(QuizNumber\ k \ /\ Total\ number\ of\ quizzes\right)^2 \tag{4}$$

For example, if Q3 (k = 3) and total number of quizzes is 20, $w_3^2 = (3/20)^2$. Therefore, the order weighted quiz correct probability would be as follows:

$$order\ weighted\ \ quiz\ correct\ probability = w_k^2 P_k \quad \left(k = 1, 2, 3, , , n\right) \tag{5}$$

In this study, each quiz generally corresponds to only one quiz, and the correct probability shown here represents the correct probability of a single question.

Five questions are recommended from this order weighted quiz correct probability in descending order of closeness to 0 (note: 0.5 has already been subtracted from the probability). In this way, quiz recommendations are made based on the probability that the student will correctly answer a question as determined by the BKT model, with extremely high or low probability of correct answer quizzes having less weight in the recommendation, and the weighting by w can make it easier to recommend quizzes with a smaller order in the question set.

Explanation Generator Using Bayesian Knowledge Tracing Model Parameters

We developed an explanation generator using BKT parameters guess (giving a correct answer despite not knowing the skill) and slip (knowing a skill but giving a wrong answer) for the quiz BKT recommendation system (Takami et al., 2021). They reported the initial quiz in a component of math learning material tends to have a higher guess value because students have not acquired the skills yet, and the quiz of acquiring a skill and then using that skill tends to have a lower guess value using an acquired skill; therefore, they are not guessing. This means easy new skill quizzes have higher guess values and difficult previous skill required quizzes have lower guess values. From this finding about guess and slip value, recommended quizzes were categorized into different feature types shown by the dotted lines in Figure 2 as a result of reviewing by a teacher with experience in teaching mathematics. According to these parameter-based quiz feature types, explanation texts (e.g., guess high meaning new skills: "You're not getting the basic skills. Let's go over the basics with this quiz!", guess low meaning previous skills required: "Now it's time to challenge applied problems! Make full use of the knowledge you have gained so far!", slip high meaning careless mistakes: "Watch out for careless mistakes!", etc.) are generated (see Figure 2 and Table A1 for more detail). We have implemented this explanation generation algorithm into our recommendation system.

Figure 3 shows a screenshot of the user interface of the implemented recommendation system. Linked to the class schedule information, this week's homework is displayed at the top, and next week's homework is displayed at the bottom. In the middle, the five recommended quizzes are displayed in order of the percentage of correct answers that are close to 50% as calculated by the BKT model and weighted so that the smaller the order, the easier it is to be recommended. The reason for the recommendation will be displayed under the title of the recommended quiz. Students who see these explanations are expected to be convinced of the reason why the quiz was recommended and to be persuaded to solve the quizzes, resulting in improved academic performance by actually solving the quiz. Students can also access the quizzes by clicking on the title of this list of quizzes.

Figure 2. Explanation generation using BKT parameter (Takami et al., 2021)

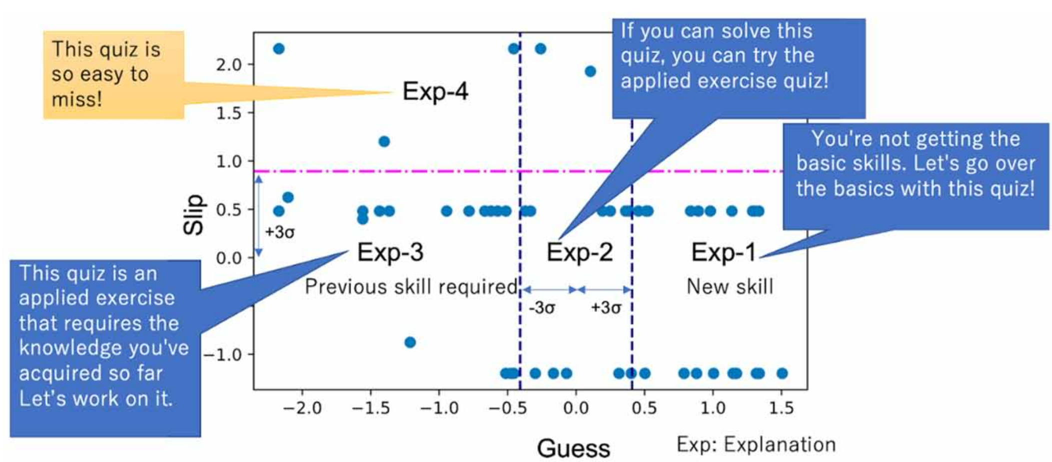
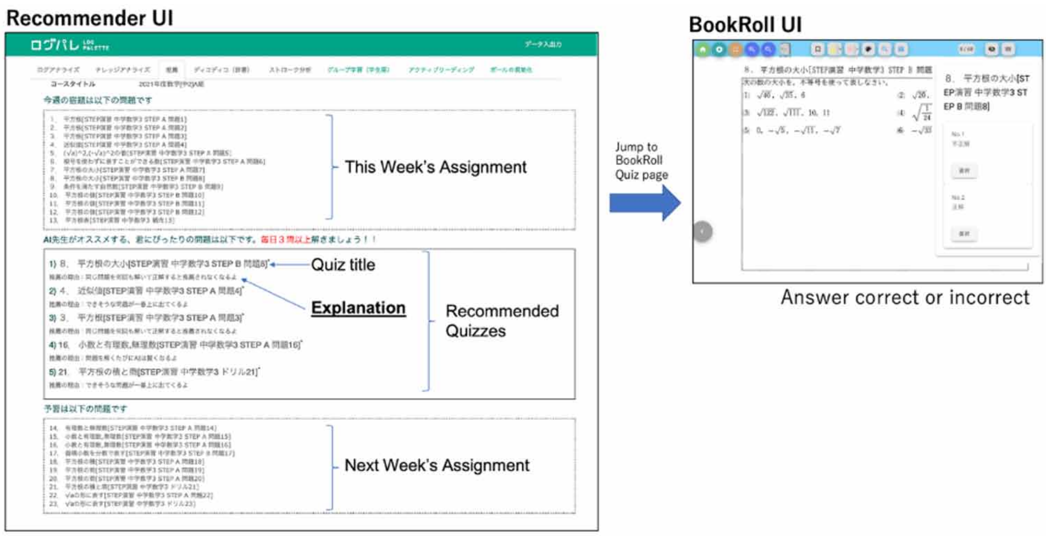


Figure 3. Screenshot of UI



Participants

A total of 115 students from three middle school math classes participated in the study. These classes were the same courses with the same teaching progressions. During the experiment, the students were studying a unit on square roots in mathematics. The teacher asked the students to solve the quizzes by the quiz recommender page. It should be noted that students were strongly encouraged to solve the quizzes and report their answers to the learning system. However, they are not required to do so since they also can choose to solve quizzes in paper-based textbooks.

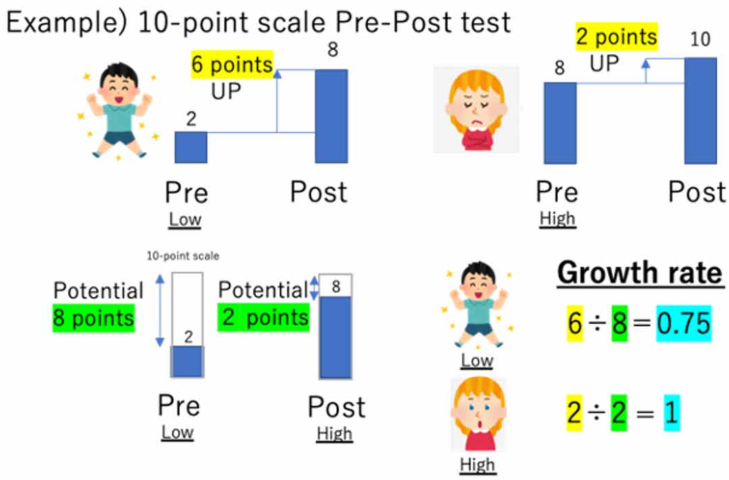
Research Design

Pre- and post-tests are used to measure knowledge gained from participating in junior high school math courses. Pre- and post-test designs are well-suited to investigating the effects of educational intervention and are common in educational research (Dugard & Todman, 1995). The pre-test was conducted on Monday, and the post-test was conducted on Friday of the same week in the classroom using their own tablet PC. The test consisted of 13 square root problems related to the content learned in class last week and graded on a 13-point scale, where each problem is worth one point. The exact same test was given again on Friday to test the student's knowledge retention. The tests are delivered from a tablet PC, and students write their answers with a stylus pen, and the pen strokes are logged. After the students have finished answering the questions, they mark each other next to their seats and report their scores to the system. Reported scores can be checked by the teacher in a handwritten log, so it is almost inconceivable that cheating would occur in a peer evaluation. The test contained the questions, and the participants were encouraged to solve at least three questions recommended by the recommendation system every day during the time between the pre and post-tests.

Growth Rate

Students who score low on the pre-test have more room to grow and are more likely to improve on the post-test, as seen in Figure 4. On the other hand, students with high scores on the pre-test have little room to grow and have difficulty increasing their scores on the post-test. Therefore, in order to evaluate how much the students were able to do what they could not do, we defined the following growth rate, which is normalized by the growth potential.

Figure 4. Growth rate



$$Growth\ potential = \left(Full\ score\ in\ the\ pre\ test\right) - \left(Pretest\ score\right)$$

$$Growth\ rate = \left(Post\ test\ score - Pre\ test\ score\right) / \left(Growth\ potential\right)$$

It should be noted that in this study, students with perfect scores on the pre-test were excluded in order to investigate the extent to which students who initially could not perform well improved. Similar pre and post-test scaling was used in evaluating the e-learning system (Akhuseyinoglu & Brusilovsky, 2022).

Data Collection

We collected the log data from quizzes in middle school mathematics classes during regular class periods from October 11th to the 16th, 2021. During these pre-test and post-study periods, students were encouraged to use the recommendation system to solve at least three questions recommended by the recommendation system every day. This log data included students' recommender page accessed data, recommended quiz clicked data, and answered data right or wrong to quizzes.

RESULTS

Pre and Post-Tests

A total of 115 students participated in this study, and 87 (75.7%) completed both the pre and post-tests. The scores in the post-test were significantly ($P<0.00001$) higher than the pre-test scores, as seen in Table 1. The mean of recommended quiz clicked (rec-clicked) was 3.05, representing the total number of clicks on recommended quizzes in five weekdays divided by the number of participants N. Since students were encouraged to solve the three recommended problems every weekday, the average use of these recommended problems was not very high. In contrast, the average number of quizzes solved per weekday was as high as 13.29, which means that the quizzes were solved not only by recommended problems but also by accessing them directly from the assignment list or book roll, as shown in Figure 3.

Results are expressed as the mean and standard deviation of the total scores obtained in pre- and post-tests. Significance (P value) was found using a paired t-test.

Table 1. Student's pre- and post-test responses

N (full points)	Mean±SD		P value	Mean±SD	
	Pre-test	Post-test		Rec-clicked	Solved quizzes
N=87 (13 points)	9.62±3.31	11.71±2.43	5.11×10-6*	3.05±3.18	13.29±14.75

Growth Rate Calculation

Results are expressed as mean and standard deviation of the total scores obtained in the pre- and post-tests. Significance (P value) was found using a paired t-test.

Table 2 shows that potential growth students who did not receive a perfect score on the pre-test increased their score by an average of 2.95. The students who scored perfectly on the pre-test had a decrease of 0.304 points, but their scores were almost perfect. One of the purposes of the recommendation system is to enable people to use the system to do things they could not do before. Therefore, in this study, we excluded students who had perfect scores on the pre-test and limited the analysis to only those students who had the potential to grow.

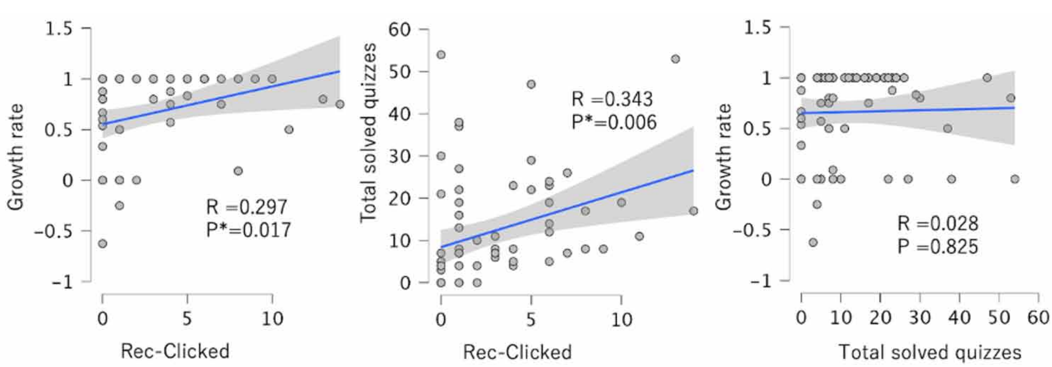
Recommender Effect on Growth Rate and Solved Quizzes

Figure 5 shows the correlation of rec-clicked counts with growth rate and total solved quizzes. Rec-clicked represents how many times each student clicked on a recommended question during the weekday. Total solved quizzes indicates the total number of quizzes solved including both from the recommended quizzes list and the assignments list. We found a significant positive correlation between rec-recommended and growth rate (R=0.297, P=0.017) and between rec-recommended and total solved quizzes (R=0.343, P=0.006). In our previous research, we conducted an A/B test comparing explanations with no explanations, and they observed a significantly higher clicking

Table 2. Potential growth students and full score students in pre-test

	N	Mean±SD			P value	Mean±SD	
		Pre-test	Post-test	Score difference		Rec-clicked	Solved quizzes
Potential growth	64	8.41±3.05	11.36±2.72	+2.95±3.45	6.93×10-8*	3.05±3.18	13.29±14.75
Full score in pre-test	23	13.00±0.00	12.70±0.69	-0.304±0.69	0.0497	3.26±2.33	16.09±13.97
Total students	87	9.62±3.31	11.71±2.43	2.09±3.02	5.11×10-6*	3.05±3.18	13.29±14.75

Figure 5. Correlation of rec-clicked counts with growth rate and total solved quizzes



on recommended quizzes with explanations (Takami et al., 2022). Taking this into consideration, these results could support that clicking on the recommended questions led to improved academic performance and using recommended quizzes led to solving more questions implying explanations why quizzes are recommended increase motivation to solve quizzes for students. Interestingly, we did not find any correlation between solved-quizzes and growth rate, suggesting merely solving a high number of problems does not lead to better growth. A detailed analysis was conducted to examine the relationship between rec-clicked, solved-quizzes and growth rate.

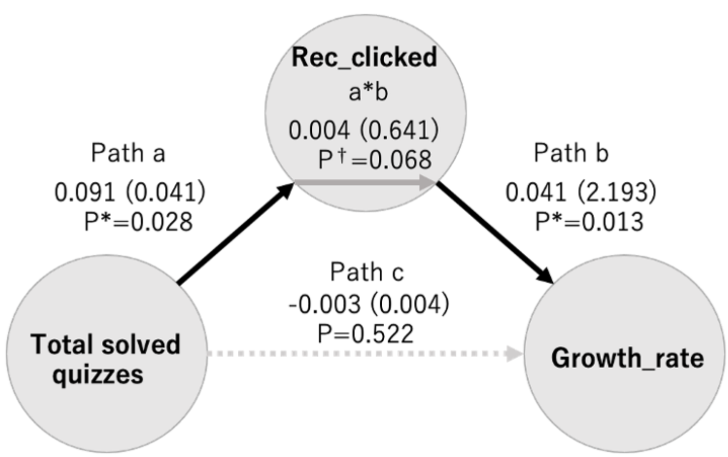
Pre Post-Tests

To quantify and test whether rec-clicked might affect the total solved quizzes on growth rate, we performed a standard mediation analysis (MacKinnon et al., 2007; VanderWeele, 2016). This analysis quantifies the degree to which a relationship between two variables X and Y can be explained by another variable, M. We defined X as total-solved quizzes, Y as growth rate, and M as rec-clicked. Path a and b measure the association between total solved quizzes and the mediator (rec-clicked), and also the association between the mediator and total-solved, and also the association between the mediator and growth rate while controlling for total-solved, respectively. More specifically, path b tests whether rec-clicked predicts variations in growth rate that are conditionally independent of total-solved.

On the other hand, paths c measures the total relationship between total solved quizzes and growth rate controlling for rec-clicked. Finally, product a*b tests the significance of the mediator. We conducted bootstrap tests (5000 interactions) for the statistical significance of the mediators as summarized in Figure 6. From this mediation analysis we found that the total relationship between total solved quizzes and growth rate was not significant (the coefficient for path c = -0.003, z = 0.004, P = 0.522), indicating that the number of solved quizzes had no relationship with growth rate. On the other hand, rec-clicked was associated with growth rate significantly after controlling for total solved (the coefficient for path b=0.041, z=2.193, P*=0.013), indicating that click counts on recommended quizzes involved in growth rate. In addition, solved quizzes was also significantly linked with rec-clicked (coefficient for path a=0.091, z=0.041, P=0.028). The mediation effect of rec-clicked was marginally significant (a*b=0.004, z=0.641, P=0.068).

These results indicate that, with respect to RQ1 (Does the BKT-based explainable recommendation system improve knowledge retention?), solving recommended quizzes according to the level of understanding, rather than just solving many quizzes, leads to improved knowledge retention.

Figure 6. Mediation analysis



Use Case Study

Next, regarding RQ2 (What are the patterns of effective use of the BKT-based explainable recommendation system for knowledge retention?), we investigated some typical usage patterns in order to identify effective usage cases of recommendation functions to improve academic performance. In Figure 7, the horizontal axis shows the quiz number, and the vertical axis shows the number of times. The gray bar indicates the number of times the recommended question has been recommended, and the red color indicates the number of times the recommended question has been clicked. The orange color indicates the number of times the recommended question has been solved. Blue indicates the total number of times the recommended problem was solved, plus the number of times it was solved by accessing it directly from the list or BookRoll. As the red and orange bar graphs above show, a recommender user increased their scores from three to eight by solving a few of the recommended basic problems. In contrast, a non-recommended user solved many different problems on their own and did not improve their performance at all from nine to nine points, despite the system recommended, as the gray bar graph shows. This result suggests that solving many different kinds of problems does not necessarily lead to growth.

Further, we extracted time-series sequential behavior from the learning logs to get a more detailed insight into these two typical examples. Figure 8 shows a sequential behavior pattern. In this figure, if the same event is recorded within a very short period of time (less than one minute), we considered it to be one time. In a recommender user, a basic quiz like Q1 that a student had solved on Day 1 was repeatedly solved even if it was recommended on Day 4, as seen in Figure 8 top. Also the recommended basic quiz Q7 was solved. In a none recommender user, as seen in Figure 8 bottom, basic quizzes like Q2 and Q9 that a student had to solve on Day 1 (actually Q2 was solved on Day 1) were not solved even if they were recommended later on Day 4.

Figure 7. Use case study: Recommender user who solved recommended quizzes and showed growth vs. none recommender user who solved many quizzes but showed no growth

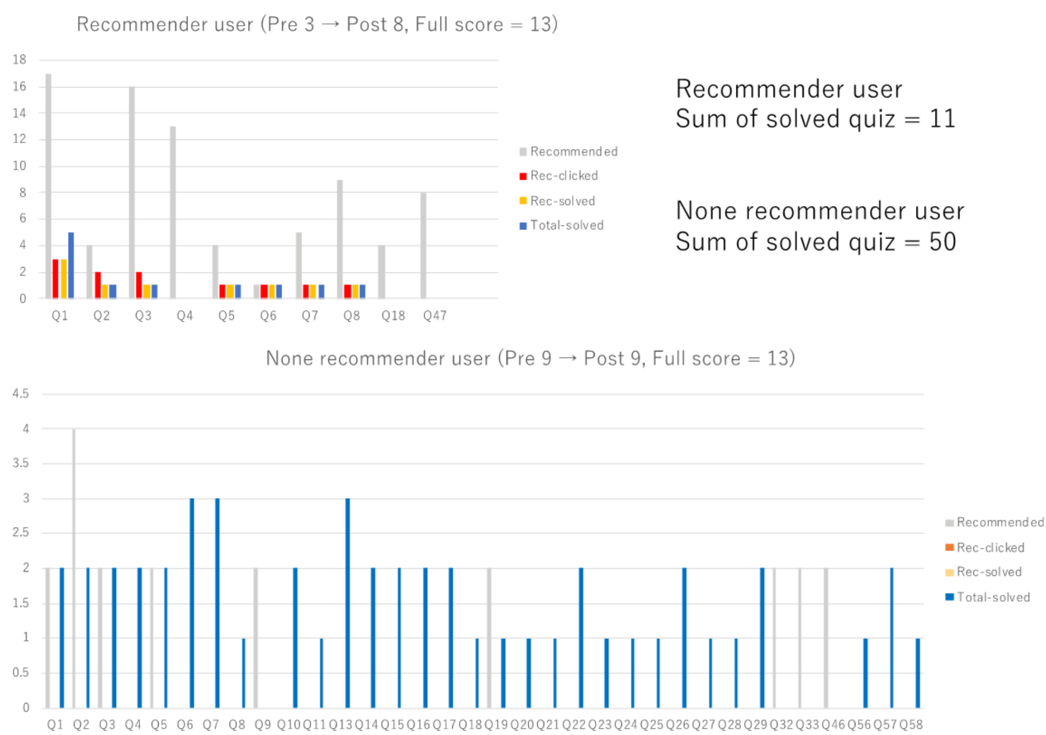
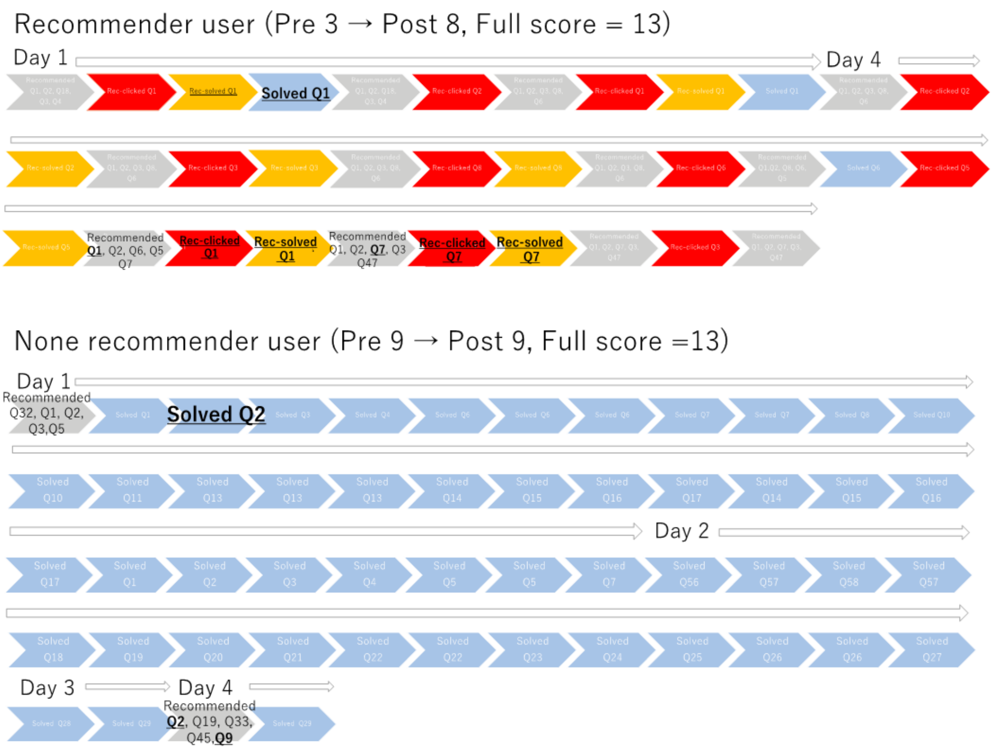


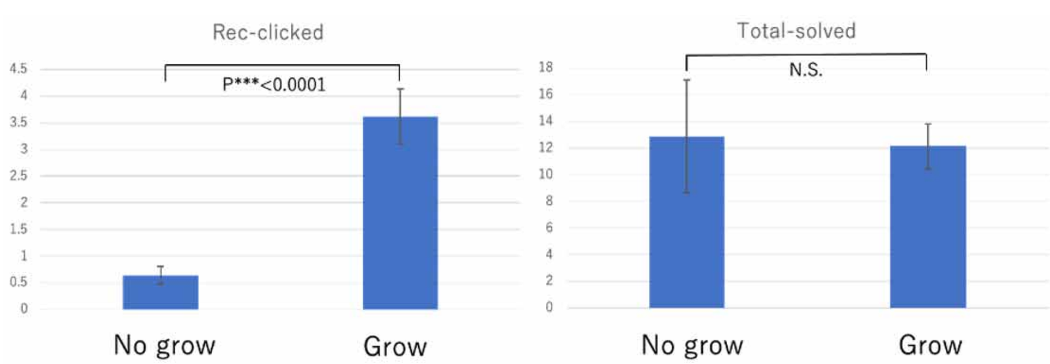
Figure 8. Use case study: Sequential behavior pattern



Quantifying the Case Study for All Potential Growth Students

From the case study analysis, we found that a student who was making good grades solved a few basic recommended problems, while a student who was not making good grades did not solve the recommended problems but solved many different problems of his/her own choosing. We also found that students with higher increasing scores solved a quiz once and solved it again with a recommendation. To quantify these results, we conducted further analysis by separating 50 students with positive growth rate values and 14 students with growth rate values less than 0. Figure 9 shows

Figure 9. Comparison of no grow (Students who scored 0 or less in growth rate, N=14) and grow (students who scored positive values in growth rate, N=50)



a comparison of students with no growth (N=14) who scored a growth rate of 0 or less and students with growth (N=50) who scored positive values in growth rate. The bar graph shows the mean±SEM of rec-clicked (left) and total-solved (right). We found that positive growth rate students had a significantly higher number of clicks on recommended quizzes than no growth students, but there was no significant difference in the total number of quizzes solved. These results are consistent with the mediation analysis shown above indicating using recommended quizzes leads to improvement in understanding retention. Figures 10 and 11 show the mean for each quiz of students who improved their growth rate and had no growth respectively. In positive growth students (Figure 10), as shown in the red (recommended quiz clicked) and orange bar (recommended quiz solved), basic quizzes Q1 to Q8 were recommended and were solved by them. On the other hand, in no growth students, as seen in Figure11, as shown in gray bar (quizzes recommended and viewed them), basic quizzes were recommended but these students rarely solved them as shown in red (recommended quiz were clicked) bar and orange (recommended quiz were solved) bar. This implies that students might have thought they understood the basic problem well enough and did not try to solve it when it was recommended to them.

Finally, we examined whether the explainable recommender led to the same quiz being solved repeatedly. Figure 12 plots the pre- and post-test score in 50 positive growth score students, and orange dots indicate 11 students repeatedly solving the same quiz by recommender. It should be noted that (9,12) and (9,13) are duplicated respectively. Most of the students who solved a problem once and then solved it again by being recommended by the system obtained a high score of 12 to 13 on the post-test on the 13-point scale, but two did not reach a high score on the pre-test even though they solved the problem repeatedly. Two students did not achieve high scores on the post-test even after solving the problems repeatedly. These two students may need more personal support from their teachers.

Figure 10. Mean for each quiz of students who improved their growth rate is indicated by the red (recommended quiz clicked) and orange (recommended quiz solved) bars

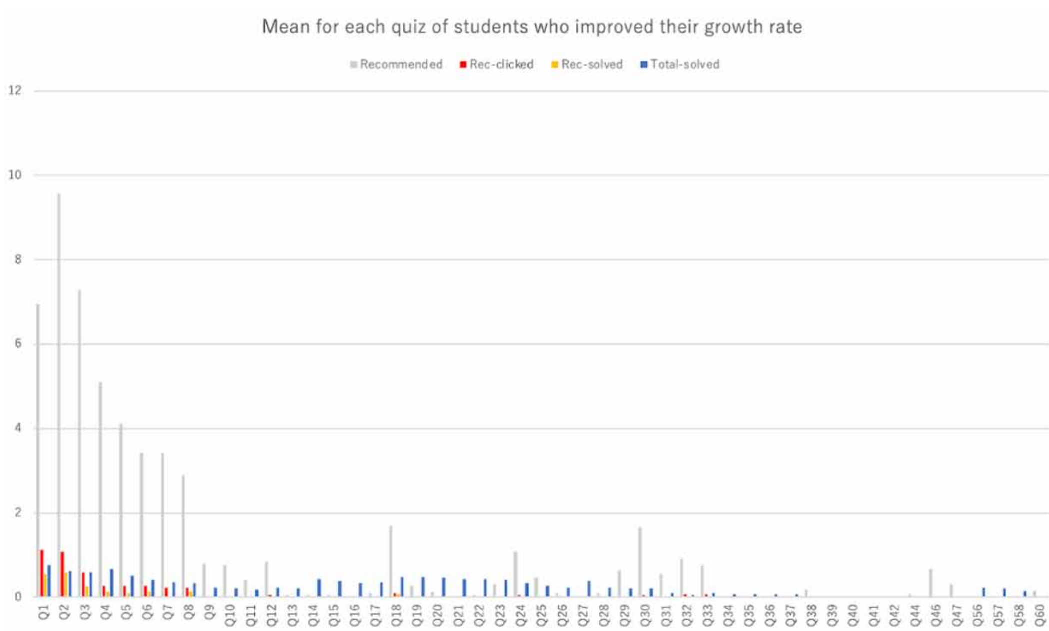


Figure 11. Mean for each quiz of students who had no growth indicating the gray bars (quizzes recommended and viewed them) indicate, that even when basic problems were recommended, they were rarely clicked on and solved

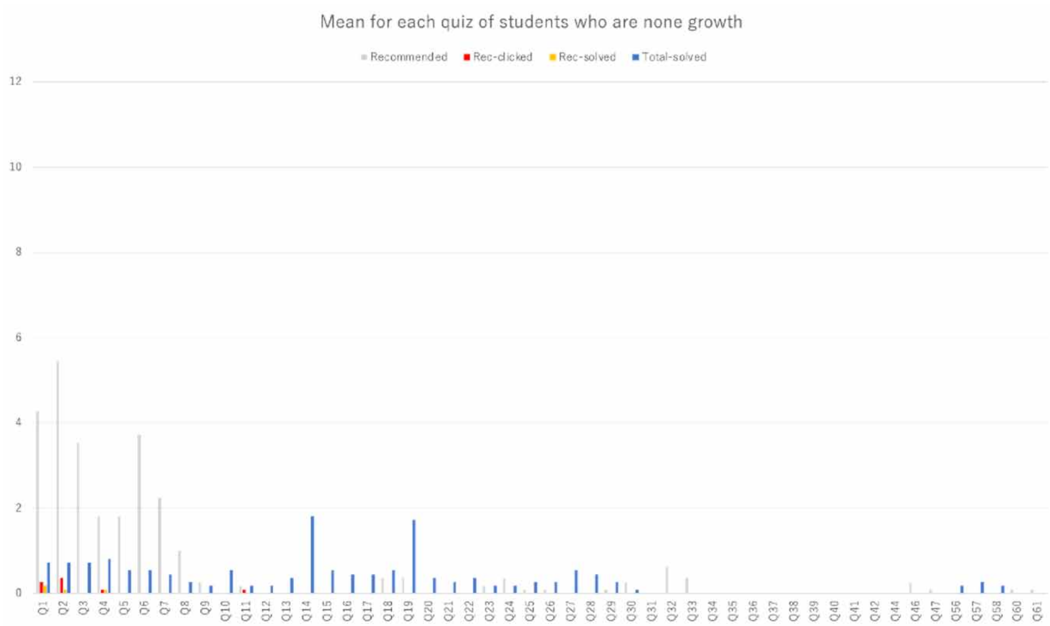
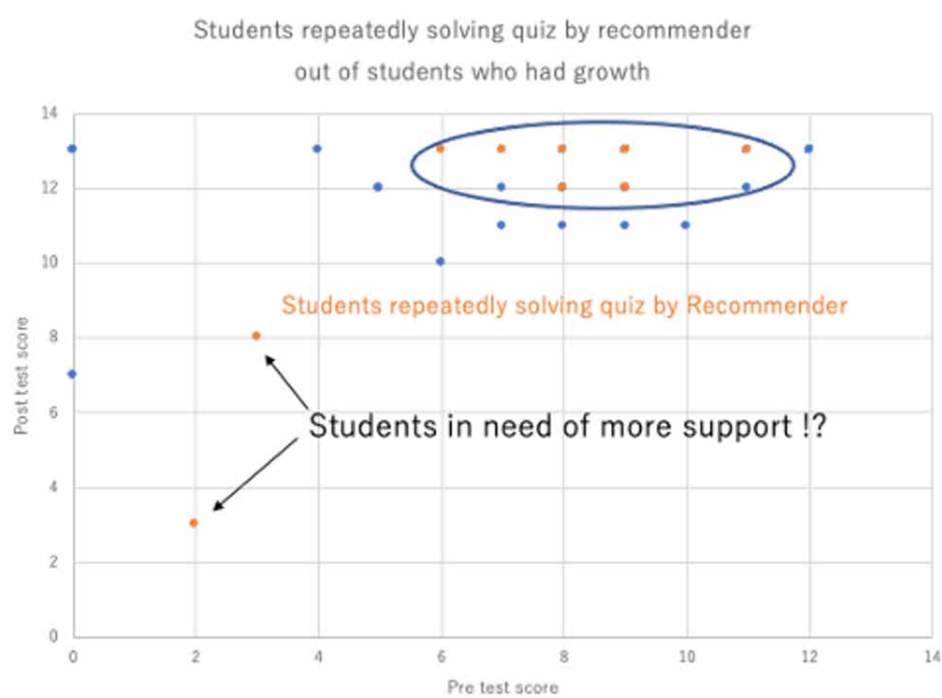


Figure 12. Scatter plot of pre- and post-test scores in students who had growth



DISCUSSIONS

In previous research on recommendation systems in education, methods like collaborative filtering, content-based filtering, and knowledge-based filtering have been employed, with insufficient exploration of those utilizing AI techniques such as fuzzy logic, decision tree, Bayesian networks, neural networks, genetic algorithms, and hidden Markov models.

In our recent study, they developed a Bayesian knowledge tracing (BKT)-based explainable recommender that not only recommends quizzes but also provides an explanation of why the students should solve the recommended quizzes using the parameters from BKT (Takami et al., 2021). They reported explaining the reason why the quiz was recommended significantly affected the positive click rate on recommended quizzes in summer vacation assignments by A/B test (Takami et al., 2022). However, the extent to which explainable recommender affects learning performance was not adequately investigated. Therefore, in this study, we examined the effectiveness of the explainable recommender system by evaluating the knowledge of math lectures among students providing pre and post-tests.

Regarding RQ1 (Does the BKT-based explainable recommender improve knowledge retention?), we showed there was a significant correlation between the number of clicks on the recommended quizzes with explanations and the growth rate (the extent to which the students were able to do what they could not do) in Figure 5. They also found that the number of recommended quizzes correlated with the total number of solved quizzes including not recommended quizzes, implying explanations of why quizzes are recommended increase motivation to solve quizzes for students. Interestingly, there was no significant correlation between the total number of solved quizzes and the growth rate. Mediation analysis indicated the usage of recommended quizzes mediated between the total number of solved quizzes and the growth rate, as seen in Figure 6. This result also suggests the importance of solving recommended quizzes to improve knowledge retention.

Regarding RQ2 (What are the patterns of effective use of the BKT-based explainable recommender for knowledge retention?), we found that easy recommended quizzes were repeatedly solved by positive growth students, as seen in Figures 7, 8, 10, and 12, and not growing students did not use recommended easy quizzes and solved many other different quizzes in their own way, as seen in Figures 7, 8, and 11. These results suggest the explainable recommender can improve students' knowledge retention by recommending easy quizzes according to BKT-estimated students comprehension.

Based on these results, the following implications are considered.

For practical implication, math anxiety is an adverse emotional reaction to math (Hembree, 1990) and was thought to develop in middle school (Berch et al., 2007). The importance of basic numerical and spatial mathematical skills for math anxiety was suggested (Maloney & Beilock, 2012). In our explainable recommender system, by estimating the student's level of understanding with the BKT model, it may be possible to overcome the student's math anxiety by recommending basic problems and generating explanations that convince them. A previous study in our recommender system indicated the explanation of the reason why the quiz recommended had a significantly higher click rate compared to not displaying an explanation (Takami et al., 2022), suggesting the possibility that the explanation will improve acceptance of easy quiz recommendations.

For system development implications, with respect to the difficulty level of the questions to be recommended, Wilson et al. (2019) examined the most efficient learning conditions using machine learning models and models that imitate perceptual learning. The results showed that, depending on the task and assumptions, learning was most effective when the percentage of correct responses was approximately 70 to 85 percent. They suggested that these results are likely to be applicable to humans. However, in our real-world educational recommender system usage study, by recommending and using those with a correct answer rate close to 0.5 and the order weighted, students had good performance. These cases are not the only ones; for instance, a recent study showed there were no significant differences in students' perceptions of the recommended courses between a basic and an advanced model (Yu et al., 2021). These suggest that no matter how high the accuracy of a simulation, it is impossible to know how effective it will be in practice and that it is important to evaluate it by having students use it in a real school, especially in the field of education. This perspective might be also important in considering how AI and education can evolve in harmony.

LIMITATIONS

One limitation of this study is that the experiment was conducted over a short period of usage time of the recommender between pre- and post-tests. However, this short period of time allowed us to check each individual's behavior in detail in the case study, and we were able to gain the knowledge that it may be important to solve the recommended basic problems many times. Another limitation, explanations of the reason for recommendation, were of limited variety, and the importance of the recommended questions may not have been adequately explained. Although there was a small amount of data for both pre- and post-test performance in this study, if a large amount of performance data is collected, it may be possible to predict how much solving the recommended quizzes will improve their performance on the post-test and explain the reason of this recommendation to students so that they can tackle the problems with more conviction. To achieve greater conviction, it is also necessary to consider various methods of explanation. In the age of generative AI, exemplified by ChatGPT (OpenAI, 2022), it could be possible to generate explanations that are more diverse and individually satisfying for learners by designing prompts based on values estimated from the BKT model. If that becomes possible, even if a quiz is recommended that students think they have mastered, a more convincing explanation will help them realize the need for the quiz, engage in learning, and improve their academic performance. To this end, we need to consider more convincing explanations that will engage them in learning.

CONCLUSION

In the context of AI-supported e-learning systems, recommender using AI techniques (e.g., fuzzy logic, decision tree, Bayesian networks, neural networks, genetic algorithms, and hidden Markov models) are lacking, with few developed and implemented. In this study, we examined the effectiveness of the Bayesian knowledge tracing model-based explainable recommender system by evaluating the knowledge of a math lecture among students by giving pre- and post-test questions-based evaluation techniques. Second year junior high school students (n=115) were asked to take the pre-test containing questions about the previous week's math lecture, and the same questions were provided at the end of the week. We obtained test scores from 87 students who responded to both the pre- and post-test. During the pre- and post-test study periods, students were encouraged to use the recommendation system. To evaluate how well the students were able to do what they could not do, we defined growth rate, which was normalized by the growth potential, and found recommended quiz clicked counts had a positive effect on the total number of solved quizzes and growth rate despite no correlation between the total number of solved quizzes and growth rate. our mediation analysis of recommended quiz click counts revealed that there was a marginally significant relationship between recommended quiz click counts indirectly mediated the relationship between total solved quizzes and growth rate. This result suggests that solving recommended quizzes according to the level of understanding, rather than just solving many quizzes, leads to improved knowledge retention. In the case study of a recommender user who had a high growth rate, basic recommended quizzes were solved even if they were quizzes they had already solved once. On the other hand, a non-recommender user who had no growth did not solve recommended basic quizzes. These results suggest that the use of an explainable recommendation system, which can recommend basic quizzes repeatedly based on the estimation of learners' understanding through BKT, will enhance AI-supported mathematical learning by enabling students to do things they could not do before.

ACKNOWLEDGMENT

We would like to thank the middle school students and teachers who participated in this study.

FUNDING

This work was partly supported by JSPS Grant-in-Aid for Early-Career Scientists JP23K17012, JSPS Grant-in-Aid for Scientific Research (B) JP23H01001, JP22H03902, JP20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) JP21K19824, and NEDO JPNP20006.

COMPETING INTERESTS

The author declares no competing interests.

REFERENCES

Akhuseyinoglu, K., & Brusilovsky, P. (2022). Exploring behavioral patterns for data-driven modeling of learners' individual differences. *Frontiers in Artificial Intelligence*, *5*, 807320. doi:10.3389/frai.2022.807320 PMID:35243337

Amatriain, X., & Basilico, J. (2015). Recommender systems in Industry: A Netflix case study. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 385–419). Springer US. doi:10.1007/978-1-4899-7637-6_11

Badrinath, A., Wang, F., & Pardos, Z. (2021). pyBKT: An accessible Python library of Bayesian knowledge tracing models. *Proceedings of The14th International Conferenceon Educational Data Mining (EDM2021)*, (pp. 468–474). IEEE.

Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable recommendations in a personalized programming practice system. *Artificial Intelligence in Education*, (pp. 64–76). IEEE.

Beck, J. E., & Chang, K.-M. (2007). Identifiability: A fundamental problem of student modeling. *In International Conference on User Modeling* (pp. 137-146). Springer Berlin Heidelberg. doi:10.1007/978-3-540-73078-1_17

Berch, D. B., Ashcraft, M. H., Krause, J. A., & Hopko, D. R. (2007). Is math anxiety a mathematical learning disability? *Why Is Math so Hard for Some Children? The Nature and Origins of Mathematical Learning Difficulties and Disabilities*, *457*, 329–348.

Bhatt, S. M., Van Den Noortgate, W., & Verbert, K. (2023). Investigating the use of deep learning and implicit feedback in K12 educational recommender systems. *IEEE Transactions on Learning Technologies, PP*, (99), 1–12.

Chen, F., & Cui, Y. (2020). LogCF: Deep collaborative filtering with process data for enhanced learning outcome modeling. *J Educ Data Min*, *12*(4), 66–99.

Chen, W., Niu, Z., Zhao, X., & Li, Y. (2014). A hybrid recommendation algorithm adapted in e-learning environments. *World Wide Web (Bussum)*, *17*(2), 271–284. doi:10.1007/s11280-012-0187-z

Colchester, K., Hagras, H., Alghazzawi, D., & Aldabbagh, G. (2017). A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms. *Journal of Artificial Intelligence and Soft Computing Research*, *7*(1), 47–64. doi:10.1515/jaiscr-2017-0004

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503. doi:10.1016/j.artint.2021.103503

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. doi:10.1007/BF01099821

Dai, Y., Flanagan, B., Takami, K., & Ogata, H. (2022). Design of a user-interpretable math quiz recommender system for Japanese high school students. *Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior*. CEUR Workshop Proceedings.

Duffy, M. C., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, *52*, 338–348. doi:10.1016/j.chb.2015.05.041

Dugard, P., & Todman, J. (1995). Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology*, *15*(2), 181–198. doi:10.1080/0144341950150207

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 469–484.

Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. *Intelligent Tutoring Systems*, (pp. 35–44). IEEE.

Gulz, A., Londos, L., & Haake, M. (2020). Preschoolers' understanding of a teachable agent-based game in early mathematics as reflected in their gaze behaviors – An experimental study. *International Journal of Artificial Intelligence in Education*, *30*(1), 38–73. doi:10.1007/s40593-020-00193-4

Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense advanced research projects agency (DARPA). *nd Web, 2*(2), 1.

Haddad, I. A. E., & Naser, S. S. A. (2017). Ado-Tutor: Intelligent tutoring system for leaning Ado.net. *European Academic Research*, *4*(10). https://philpapers.org/rec/ELHAIT-2

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*(1), 33–46. doi:10.2307/749455

Ilić, M., Mikić, V., Kopanja, L., & Vesin, B. (2023). Intelligent techniques in e-learning: A literature review. *Artificial Intelligence Review*, *56*(12), 14907–14953. doi:10.1007/s10462-023-10508-1

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. doi:10.1016/j.caeai.2021.100017

Kandakatla, M., & Bandi, K. (n.d.). A content based filtering and negative rating recommender system for e-learning management system. *International Conference on Intelligent Computing and Sustainable System (ICICSS)*, (pp. 167–172). IEEE.

Kelly, A. E., Sleeman, D. H., & Gilhooly, K. J. (1993). Artificial intelligence in education: Using state space search and heuristics in mathematics instruction. *International Journal of Man-Machine Studies*, *38*(4), 725–746. doi:10.1006/imms.1993.1034

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One*, *7*(5), e36399. doi:10.1371/journal.pone.0036399 PMID:22649492

Koedinger, K. R. (2002). *Toward evidence for instructional design principles: Examples from cognitive tutor math 6.* ERIC/CSMEE. https://eric.ed.gov/?id=ED471749

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 73–105). Springer US. doi:10.1007/978-0-387-85820-3_3

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*(1), 593–614. doi:10.1146/annurev.psych.58.110405.085542 PMID:16968208

Maloney, E. A., & Beilock, S. L. (2012). Math anxiety: Who has it, why it develops, and how to guard against it. *Trends in Cognitive Sciences*, *16*(8), 404–406. doi:10.1016/j.tics.2012.06.008 PMID:22784928

Matsuda, N., Weng, W., & Wall, N. (2020). The effect of metacognitive scaffolding for learning by teaching a teachable agent. *International Journal of Artificial Intelligence in Education*, *30*(1), 1–37. doi:10.1007/s40593-019-00190-2

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3), 393–444. doi:10.1007/s11257-017-9195-0

Nwana, H. S. (1990). Intelligent tutoring systems: An overview. *Artificial Intelligence Review*, *4*(4), 251–277. doi:10.1007/BF00168958

Ogata, H., Flanagan, B., Takami, K., Dai, Y., Nakamoto, R., & Takii, K. (2023). EXAIT: Educational eXplainable artificial intelligent tools for personalized learning. *Research and Practice in Technology Enhanced Learning, 19*, 019.

Open A. I. (2022). *Introducing ChatGPT*. Open AI. https://openai.com/blog/chatgpt/

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in Neural Information Processing Systems, 28*. https://proceedings.neurips.cc/paper/5654-deep-knowledge-tracing

Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., & Heffernan, N. T. (2011). Does time matter? Modeling the effect of time in Bayesian knowledge tracing. In EDM (pp. 139-148).

Rahdari, B., Brusilovsky, P., & Thaker, K. (2020). Using knowledge graph for explainable recommendation of external content in electronic textbooks. *ITextbooks*. https://openreview.net/pdf?id=oW5V2-wiGDD

Samin, H., & Azim, T. (2019). Knowledge based recommender system for academia using machine learning: A case study on higher education landscape of Pakistan. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 67081–67093. doi:10.1109/ACCESS.2019.2912012

Self, J. A. (1974). Student models in computer-aided instruction. *International Journal of Man-Machine Studies*, *6*(2), 261–276. doi:10.1016/S0020-7373(74)80005-2

Shen, S., Liu, Q., Chen, E., Huang, Z., Huang, W., Yin, Y., Su, Y., & Wang, S. (2021). Learning process-consistent knowledge tracing. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, (pp. 1452–1460). ACM.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE Internet Computing*, *21*(3), 12–18. doi:10.1109/MIC.2017.72

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-enhanced sequential modeling for student performance prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11864. doi:10.1609/aaai.v32i1.11864

Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022). Educational explainable recommender usage and its effectiveness in high school summer vacation assignment. *LAK22: 12th International Learning Analytics and Knowledge Conference*, (pp. 458–464). ACM. doi:10.1145/3506860.3506882

Takami, K., Flanagan, B., Dai, Y., & Ogata, H. (2021). Toward educational explainable recommender system: Explanation generation based on Bayesian knowledge tracing parameters. *29th International Conference on Computers in Education Conference Proceedings*. https://icce2021.apsce.net/wp-content/uploads/2021/12/ICCE2021-Vol.II-PP.-531-536.pdf

Tapalova, O., & Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIEd for personalised learning pathways. *Electronic Journal of e-Learning*, *20*(5), 639–653. doi:10.34190/ejel.20.5.2597

Tärning, B., Silvervarg, A., Gulz, A., & Haake, M. (2019). Instructing a teachable agent with low or high self-efficacy – Does similarity attract? *International Journal of Artificial Intelligence in Education*, *29*(1), 89–121. doi:10.1007/s40593-018-0167-2

Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 353–382). Springer US. doi:10.1007/978-1-4899-7637-6_10

Tsiakas, K., Barakova, E., Khan, J. V., & Markopoulos, P. (2020). BrainHood: Towards an explainable recommendation system for self-regulated cognitive training in children. *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, (pp. 1–6). ACM. doi:10.1145/3389189.3398004

VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, *37*(1), 17–32. doi:10.1146/annurev-publhealth-032315-021402 PMID:26653405

Walkington, C., & Bernacki, M. L. (2019). Personalizing algebra to students' individual interests in an intelligent tutoring system: Moderators of impact. *International Journal of Artificial Intelligence in Education*, *29*(1), 58–88. doi:10.1007/s40593-018-0168-1

Wang, Z., Feng, X., Tang, J., Huang, G. Y., & Liu, Z. (2019). Deep knowledge tracing with side information. *Artificial Intelligence in Education*, 303–308.

Wenger, E. (2014). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. Morgan Kaufmann.

Wilson, R. C., Shenhav, A., Straccia, M., & Cohen, J. D. (2019). The eighty five percent rule for optimal learning. *Nature Communications*, *10*(1), 4646. doi:10.1038/s41467-019-12552-4 PMID:31690723

Wind, D. K., Jørgensen, R. M., & Hansen, S. L. (2018). Peer feedback with peergrade. *ICEL 2018 13th International Conference on E-Learning*, 184.

Xiao, J., Wang, M., Jiang, B., & Li, J. (2018). A personalized recommendation system with combinational algorithm for online learning. *Journal of Ambient Intelligence and Humanized Computing*, *9*(3), 667–677. doi:10.1007/s12652-017-0466-8

Xie, T., Zheng, Q., Zhang, W., & Qu, H. (2017). Modeling and predicting the active video-viewing time in a large-scale E-learning system. *IEEE Access : Practical Innovations, Open Solutions*, *5*, 11490–11504. doi:10.1109/ACCESS.2017.2717858

Yang, H., & Cheung, L. P. (2018). Implicit heterogeneous features embedding in deep knowledge tracing. *Cognitive Computation*, *10*(1), 3–14. doi:10.1007/s12559-017-9522-0

Yeung, C.-K., & Yeung, D.-Y. (2018). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, (pp. 1–10). ACM. doi:10.1145/3231644.3231647

Yu, R., Pardos, Z., Chau, H., & Brusilovsky, P. (2021). Orienting students to course recommendations using three types of explanation. In *Adjunct proceedings of the 29th ACM conference on user modeling, adaptation and personalization* (pp. 238–245). Association for Computing Machinery. doi:10.1145/3450614.3464483

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. *Artificial Intelligence in Education*, 171–180.

Zhang, N., Du, Y., Deng, K., Li, L., Shen, J., & Sun, G. (2020). Attention-based knowledge tracing with heterogeneous information network embedding. *Knowledge Science, Engineering and Management*, 95–103.

Zhang, Q., Lu, J., & Zhang, G. (2022). Recommender systems in e-learning. *Journal of Smart Environments and Green Computing*. 10.20517/jsegc.2020.06

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval, 14*(1), 1–101.

Zou, Y., Yan, X., & Li, W. (2020). Knowledge tracking model based on learning process. *Journal of Computer and Communications*, *8*(10), 7–17. doi:10.4236/jcc.2020.810002

APPENDIX

Depending on the BKT parameters of the quiz, one of the four texts for each explanation type is randomly selected. In addition to the guess-based explanations, if the slip value is greater than the pink dotted line in Figure 2, an Exp-4 explanation text is added after Exp-1, Exp-2, or Exp-3. For example, if guess= -1.4 and slip= 1.2, then one of the Exp-3 texts and one of the Exp-4 texts will be selected. Hence, an explanation like "Now it's time to challenge applied problems! Make full use of the knowledge you have gained so far! Watch out for careless mistakes!" will be generated.

Table A1. Generated texts for explanation (Takami et al., 2021)

Explanation Type	Texts of Explanations
Exp-1	'You're not getting the basic skills. Let's go over the basics with this quiz!' 'Let's carefully go over some basic skills with this problem!' 'You are not getting the basic skills. Let's work on the basic problems!' 'You don't seem to have the basics down, so with this problem, let's get the basics down!'
Exp-2	'If you can solve this quiz, you can try the applied exercise quiz!' 'This is the skill you need to solve the applied quiz!' 'If you can't solve this problem, you can't solve the applied quizzes!' 'If you can solve this quiz, you can improve your skills in applied quiz!'
Exp-3	'It's an applied quiz that requires some skills. It's great if you can solve it!' 'Now it's time to challenge applied problems! Make full use of the knowledge you have gained so far!' 'This quiz is an applied exercise that requires the knowledge you've acquired so far. Let's work on it.' 'Let's try this quiz! This is a quiz that you can solve by using your learned skills.'
Exp-4	'This quiz is so easy to miss!' 'This quiz is easy for everyone to make mistakes on, so be careful!' 'Watch out for careless mistakes!' 'People often make careless mistakes on this question, so be careful!'

Kyosuke Takami is a senior researcher at Education Data Science Center, National Institute for Educational Policy Research (NIER) in Japan. He has prior expertise in the application of neuroscience in education and received a PhD degree from Osaka University. He has been a high school teacher for more than 9 years. His research interests include: Learning Analytics, Educational Data Mining and Educational Technologies.

Brendan Flanagan received a bachelor's degree from RMIT University and master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. He is currently an Associate Professor at Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, Kyoto University. His research interests include learning analytics, text mining, machine learning, language learning, and blockchain in education.

Yiling Dai is a Program-Specific Researcher at the Academic Center for Computing and Media Studies, Kyoto University. She received a Bachelor's degree from Zhejiang University, a Master's degree from the Graduate School of Business, Rikkyo University, and a PhD degree from the Graduate School of Informatics, Kyoto University. Her research interests include: information retrieval, knowledge discovery, educational data mining and learning analytics.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include Learning Analytics, Evidence-Based Education, Educational data mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.