JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(1), 322–337. https://doi.org/10.18608/jla.2025.8557

Integrating Option Tracing into Knowledge Tracing: Enhancing Learning Analytics for Mathematics Multiple-Choice Questions

Hai Li[1], Wanli Xing[2]*, Chenglu Li[3], Wangda Zhu[4] and Simon Woodhead[5]

Abstract

Knowledge tracing (KT) is a method to evaluate a student's knowledge state (KS) based on their historical problem-solving records by predicting the next answer's binary correctness. Although widely applied to closed-ended questions, it lacks a detailed option tracing (OT) method for assessing multiple-choice questions (MCQs). This paper introduces a general OT method that can be seamlessly integrated into deep knowledge tracing (DKT) methods through data processing techniques and network output modules. Using a million-level assignment record of MCQs from a K–12 math learning platform, which includes two types of knowledge components (KCs), skill and misconception, we converted five different DKT models into deep option tracing (DOT) models. Performance metrics demonstrate that OT enhances KT performance and effectively identifies students' future option selection tendencies. Furthermore, using the best OT model, we extracted students' problem-solving sequence features and learning gains to analyze learning patterns. The results reveal that for beginners in middle school mathematics, consecutive errors in the same skill might lead to greater learning gains. Finally, we applied network analysis to reveal connections between skills based on students' error tendencies. Our work contributes to KT methods and related empirical findings in learning analytics (LA) for knowledge assessment.

Notes for Practice

- Existing knowledge tracing (KT) models typically predict the binary correctness of student answers but lack the ability to track specific options in multiple-choice questions (MCQs). This limitation overlooks valuable insights into student option selection behaviour.
- This study designed an innovative option tracing (OT) method that predicts future selection tendencies based on students' historical assignment records. The knowledge component (KC) simultaneously considers the skill of the correct option and the misconception of the distractors, enhancing the effectiveness of assessing students' knowledge states (KSs).
- Using the OT model, researchers can extract features from students' problem-solving sequences to analyze the relationship between learning patterns and learning gains. The results indicate that for novice mathematics learners, repeatedly making mistakes on the same skill may lead to greater learning gains.
- A skill network analysis method was designed based on students' error tendencies to establish relationships between skills and prerequisite skills. This analysis can provide insights for instructional design and targeted interventions.

Keywords

Knowledge tracing, option tracing, mathematics multiple-choice questions, learning pattern analytics.

Submitted: 07/07/2024 — **Accepted:** 19/12/2024 — **Published:** 24/01/2025

[1] *Email: li.ha@ufl.edu Address: University of Florida, Gainesville, Florida, USA, ORCID iD: https://orcid.org/0009-0004-7299-2042*

[2] *Corresponding author *Email: wanli.xing@coe.ufl.edu Address: University of Florida, Gainesville, Florida, USA, ORCID iD: https://orcid.org/0000-0002-1446-889X*

[3] *Email: chenglu.li@utah.edu Address: University of Utah, Salt Lake City, Utah, USA, ORCID iD: https://orcid.org/0000-0002-1782-0457*

[4] *Email: wangdazhu@ufl.edu Address: University of Florida, Gainesville, Florida, USA, ORCID iD: https://orcid.org/0000-0001-9611-4800*

[5] *Email: simon.woodhead@eedi.co.uk Address: Eedi, London, UK, ORCID iD: https://orcid.org/0000-0002-2192-9797*

1. Introduction

Learner modelling refers to mathematical models that estimate a learner's understanding of a taught subject or their competency in other structures such as metacognition (Pelánek, 2017). It can serve as a foundation for adaptive learning systems such as intelligent tutoring systems. This approach typically relies on data collected from the learner's explicit or implicit interactions with the system to estimate their current knowledge state (KS) and track their learning progress (Pelánek, 2017). Additionally, the contextual information of the scenario, often referred to as the knowledge component (KC), is considered. This includes the knowledge, skills, or concepts contained within the problems (Rochmad et al., 2018).

Knowledge tracing (KT) is one of the prominent applications (Ghosh et al., 2020; Liu, Liu, Chen, Huang, Gao, et al., 2023; Liu, Liu, Chen, Huang, & Luo, 2023) in learner modelling, used to evaluate students' KSs and predict their performance on future tasks, typically represented by binary classifications of whether they will answer the next questions correctly or not. There are generally two forms of representation. The first form uses the average accuracy of all questions in the question bank to evaluate the KS, reflecting the student's overall mastery of the knowledge (Cai et al., 2019; Scruggs et al., 2019). The second form decomposes domain knowledge into fundamental skills and maps the student's performance on problems to the knowledge level of each skill. This means evaluating the KS as the mastery level of different skills (Lu et al., 2023; Y. Huang et al., 2016).

Multiple-choice questions (MCQs) are commonly used in large-scale exams and related online learning platforms due to their objectivity and relatively reliable scoring procedures (Abida et al., 2011). The quicker scoring also allows for more timely feedback on student performance. In the context of the MCQs in the online learning platform discussed in this paper, each question has four options: one correct option and three incorrect options (distractors). The KC in this paper includes two typical attributes in the multiple-choice knowledge space (Stout et al., 2023): skills (correct option) and misconceptions (distractors). The reasons students choose these distractors may vary, but this paper focuses mainly on two primary reasons: (1) a lack of necessary understanding of the KCs being tested, such as skills or concepts, which may lead to misconceptions, and (2) the effects of forgetting (Adler & Clark, 1991; Ebbinghaus, 2013).

Although KT has achieved good results in accuracy assessment, it faces limitations in the context of MCQs. This is because KT only considers binary correctness labels (i.e., correct or incorrect), ignoring the specific options selected by students, making it difficult to diagnose specific student errors. However, error information is crucial for teachers to measure students' mastery and improve instruction, especially since incorrect options in mathematics reflect students' misconceptions and metacognition (Kramarski, 2004; Ndemo & Ndemo, 2018). Errors, including misunderstandings, can play a significant role in mathematics education (Borasi, 1996), and analyzing student errors has the potential to foster inquiry-based learning in mathematics (Hattie, 2008). However, addressing these errors remains a challenging aspect of teaching for educators. To fill this gap, in this paper, we extend the existing KT methods from correctness prediction (binary classification) to predicting students' choices in MCQs (four-class classification, one out of four). This extended method is referred to as option tracing (OT). The OT model can predict students' learning gains, thereby analyzing the relationship between these gains and students' learning patterns, especially their tendencies toward errors. These error tendencies can be used to examine the relationships between different skills.

Specifically, we proposed a general method applicable to deep learning KT based on a high-quality and large-scale dataset of K–12 student responses to mathematical MCQs. This method combines two types of KCs (skill and misconception), including data processing and model output modules, transforming the existing binary classification in KT to multiclass OT. By predicting students' future options, we simultaneously obtained their correctness and error causes and translated them into students' KSs. After comparing the performance of five different deep knowledge tracing (DKT) to deep option tracing (DOT) conversions, we then used the optimal OT model to predict students' learning gains. This allowed us to analyze the relationship between students' answering patterns and their learning gains. Our results suggest that for middle school mathematics beginners, repeated errors on the same skill can lead to greater learning gains. Finally, based on the KS evaluation results, we designed a method to extract network relationships between different skills at the individual student level. The results indicate that the skill percentages might be a critical factor affecting student performance. Specifically, this paper addresses the following research questions (RQs):

RQ1: How can a KT model based on deep learning be expanded and improved into an OT model?
RQ2: Based on the KS evaluation of the model in RQ1, what type of problem-solving patterns lead to higher learning gains for novice mathematics learners in online learning platforms?
RQ3: How can the relationships between different skills within the KS result of students in RQ2 be established?

The contributions of this work are twofold. First, this work originates from KT and designs the OT model for better KS assessment in mathematical MCQ scenarios. Second, the learning analytic methods integrated with the model, including student learning pattern and skill network, provide insights that can support optimization of learning environments and instructional design.

2. Related Works

To assess the state of knowledge, KT refers to estimating students' mastery of skills and concepts based on their historical responses to questions (Ghosh et al., 2021). These estimates are then used to predict their future performance and simulate changes in their KS over time. KT methods play a crucial role in many large-scale online learning platforms (Pu et al., 2018), enabling the automatic assessment of numerous students' knowledge levels. This, in turn, provides scaffolding support, personalized feedback, and recommended learning resources, thereby enhancing learning outcomes (Abdelrahman et al., 2023). Beyond personalized learning, the progress of students' knowledge traced by KT can be combined with KCs to design effective instructional practices for adaptive learning. For instance, teachers can implement knowledge-enhancing data teaching (Abdelrahman & Wang, 2023), dynamically adjusting their instruction based on the class's knowledge levels.

2.1 KT

Before 2010, the early development of KT methods could be categorized into two main approaches. The first approach revolves around Bayesian knowledge tracing (BKT) (Pelánek, 2017; Yudelson et al., 2013), a special case of the hidden Markov model. BKT typically involves multiple binary indicators to study whether a learner has mastered a sub-concept, quantifying user knowledge using probabilistic statistics for calculation and analysis. In the early stages of research on this model, a study (Corbett & Anderson, 1995) used BKT to assess students' current competence levels based on their previous programming exercises. This assessment was then used to guide the tutor's evaluative strategies. The study demonstrated BKT's effectiveness in scaffolding learning and supported the idea that complex knowledge can be broken down into a hierarchical structure of latent KCs and skills. The second approach is item response theory (IRT) models, which use skill models (especially polytomous skill models) as response models, ultimately representing a learner's knowledge level as a vector of high and low values for different concepts (Yeung, 2019). A typical example is the performance factors analysis model (Gong et al., 2010), which uses logistic functions to estimate the probability of KSs. This method often relies on expert-designed features such as the number of student attempts, successes, and failures.

Since 2010, DKT based on deep learning methods has become the mainstream approach for large-scale learner response datasets. Various model methods have gradually developed. Due to its effectiveness on large datasets, DKT has become the new benchmark for KT methods. In this paper's context of the million-scale dataset, we used DKT methods as baselines. The following introduces five models with different characteristics used in this paper, selected to provide a comprehensive overview of KT techniques. Among them are three state-of-the-art deep learning models—AKT, ATDKT, and SparseKT—that represent both the mature application of neural networks in KT (AKT) and the latest advancements in the field (ATDKT and SparseKT). These models were chosen for their ability to handle large-scale datasets and effectively capture complex learning patterns, as evidenced by significant improvements in predictive accuracy (AUC) across multiple datasets, where they have also frequently served as baselines in numerous studies. Notably, AKT is one of the most commonly used baselines, ATDKT enhances the modelling of KC, and SparseKT leverages attention mechanisms to focus on students' primary learning behaviours. Additionally, we included SimpleKT and FoLiBiKT, which incorporate psychometric principles with deep learning techniques to further explore how hybrid models can capitalize on the strengths of both approaches. SimpleKT, for instance, integrates the Rasch model, while FoLiBiKT accounts for students' forgetting behaviours. This diverse selection provides a comprehensive perspective on the current state and evolution of the KT field, enabling a thorough and effective evaluation of our approach.

1. AKT (Ghosh et al., 2020) combines a flexible attention-based neural network model with a series of novel and interpretable components inspired by cognitive and psychometric models. It connects learners' future responses to assessment questions with their past responses to consider the similarity between questions and answers. Compared to the baselines of BKT and DKT, it achieved a 2%–6% improvement in AUC on multiple ASSISTments datasets.

2. ATDKT (Liu, Liu, Chen, Huang, Gao, et al., 2023) incorporates the co-occurrence matrix of skill into the prediction task and uses individualized prior knowledge prediction tasks as auxiliary learning tasks to strengthen the prediction task of KT. This approach explores the intrinsic relationships between questions and KCs. On three math short-answer KT datasets, it achieved about a 2% improvement in AUC and accuracy compared to the AKT model (and other DKT baselines).

3. Inspired by the Rasch model in psychometrics, SimpleKT (Liu, Liu, Chen, Huang, & Luo, 2023) explicitly models question-specific variations to capture individual differences between questions covering the same set of KCs. These KCs generalize the concepts or skills needed for learners to complete tasks or steps in problems. An attention function extracts time-aware information embedded in student learning interactions. The model is highly interpretable and has a simple structure. On multiple public KT datasets, it achieved a slight improvement in AUC compared to the AKT model and other DKT baselines.

4. SparseKT (S. Huang et al., 2023) improves the robustness and generalizability of attention methods through heuristic sparsification, selecting only the samples with the highest attention scores. This approach reduces the model's focus on irrelevant learning behaviours of students' KSs. It achieved about a 1% improvement in AUC compared to the AKT model on several open-source datasets.

5. FoLiBiKT (Im et al., 2023) leverages advanced capabilities to capture long-term dependencies in context, reflecting the forgetting behaviour in learning as a linear bias added to several baseline models' attention modules. On multiple public datasets, it achieved a 1%–3% improvement in AUC compared to the AKT model.

2.2 OT

Few methods exist for OT, and most belong to the realm of deep learning. Ghosh and colleagues (2021) designed five DOT models based on a K–12 math multiple-choice dataset, including models like recurrent neural networks (RNN) and long short-term memory (LSTM) structures, as well as the AKT model. They compared these with previous works based on IRT, which used collaborative filtering methods to design OT models as a baseline. In the DOT models presented here, the KCs only considered the skills required for the questions. Among these six models, AKT performed the best with an OT AUC of 0.736. An and colleagues (2022) developed a novel deep learning multitask model, incorporating LSTM structures to achieve both KT and OT, and used student performance prediction as a downstream task. This model achieved the best OT AUC of 0.742 in English MCQs.

Previous works have achieved OT with some success, but they employed specially designed network structures and methods that lack scalability to broader scenarios. The method designed in this paper can be extended to most DKT methods with binary classification. Additionally, past works only considered KCs for math skills, where different question options usually cannot provide effective conceptual or classification parameters (all four options' KC parameters are skills). The lack of prior knowledge makes it difficult for the model to establish connections between the different knowledge granules of options. However, our study used a high-quality dataset where each question's options included one skill and three misconceptions. This use of mathematical concepts enhances the potential support of OT for KT and aids in analyzing the characteristics of students' learning gain patterns.

3. Dataset

Our dataset is sourced from a large online mathematics learning platform, Eedi, which comprises K–12 MCQs. The platform's users are students aged 10 to 13 in the UK and the US. Each question includes one correct option (representing one skill) and three distractors (representing three misconceptions). The dataset comprises 1,597 questions, which cover 14 primary skills and correspond to 63 secondary skills. The distractors encompass 801 types of misconceptions. For a small portion of questions with missing distractors, we treated the question and its options as a textual input and used SentenceBERT to supplement the missing misconceptions by identifying the semantically closest matches among all available options (a commonly used semantic retrieval method; Reimers & Gurevych, 2019). While this method effectively addresses the issue of missing misconceptions, it introduces some accuracy challenges, which could be mitigated through further optimization in future work. The secondary skills offer a finer granularity of knowledge for the correct options, and, therefore, they are used as KC parameters, while the primary skills would be used in the visualization analytics. Figure 1 shows an example of a question.

The dataset of student assignment records contains a total of 1,292,204 entries. Each entry includes the student's choices, whether they were correct, and the submission time, involving 25,340 students. The sequence of tasks attempted by the students is determined by teacher-assigned skills and randomly selected from a question bank. The dataset is randomly split into 80% for the training set and 20% for the validation set. No initial KS is required for the training set. However, for the validation set, an initial KS of 10 questions is set for 5,294 students, meaning only students with more than 11 question records are included in the validation set. This method ensures higher reliability in predicting student accuracy based on a uniform initial state, and only the assignment records following this initial state are validated. After screening, we obtained 233,200 assignment records from 4,537 students. For RQ2, the initial KS is also set to 10 questions to analyze learning patterns, with the final KS set to 30 questions to calculate learning gains. A total of 14,135 students with at least 30 task records are included. For RQ3, the skill network analytics is based on the error tendencies identified in RQ2, using the same dataset.

4. RQ1: OT

4.1 Methods

As previously mentioned, the characteristics of the five KT models we selected from related work are summarized in Table 1. Considering the characteristics of the models, the KC parameter is only effective in three of them. For KT, the KC parameter of a problem is represented as skill, while for OT, the KC parameters of the options include both skill and misconception. The goal of this research is not to compare all KT methods but to design a framework that generalizes the tracking of students'

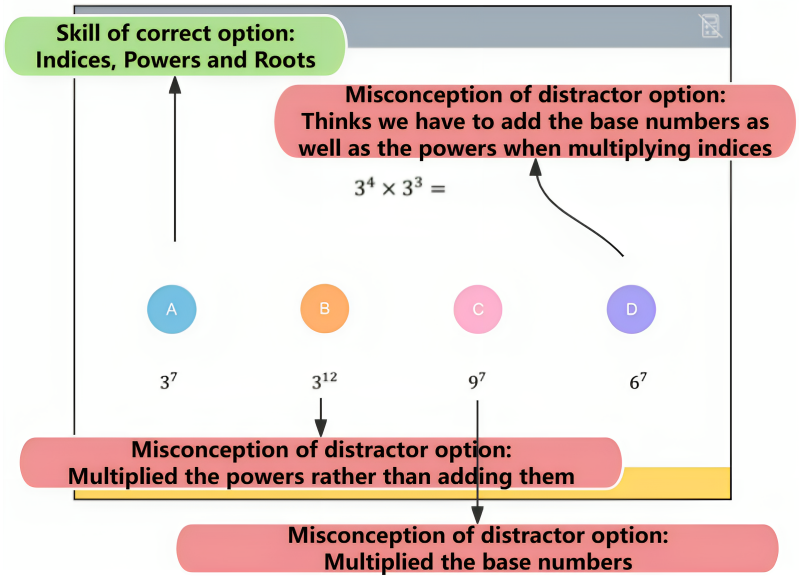


Figure 1. An example of an MCQ.

Table 1. Five KT models.

Name	Embedded with KC	Description
AKT	Yes	Considers psychometric models, linking learners' future responses to assessment questions with their past responses to account for the similarity between questions and answers
ATDKT	Yes	Incorporates the KC co-occurrence matrix into the prediction task and uses personalized prior knowledge prediction as an auxiliary task to enhance KT predictions
SimpleKT	No	Inspired by the Rasch model in psychometrics, explicitly models item-specific variations to capture individual differences between questions covering the same set of KCs, using an attention mechanism to extract temporally aware information from student learning interactions
SparseKT	No	Uses a heuristic sparsification method, selecting only samples with the highest model attention, thereby reducing the model's focus on irrelevant student learning behaviour samples
FoLibiKT	Yes	Leverages advanced capabilities for capturing long-term contextual dependencies, reflecting forgetting behaviours in learning as linear biases, and integrating this attention mechanism into several baseline models

mathematical option choices. As long as the KT model conforms to the binary classification sequence prediction task, it can be modified into an OT model using this method.

We proposed a general method of converting KT into OT, that is, transforming binary classification into four-class classification. OT can predict the likelihood of students choosing options for future exercises based on their learning history, and it can also be viewed as the current state of knowledge. The specific process is as follows:

1. **Data processing:** Split each MCQ into four options as a group. Set the time for each group of samples to the same value (equivalent to the time of the choice). Set the concept of each sample in the group (embedding with KCs) as one skill and three misconceptions, respectively. Ensure that the model's batch size is a multiple of four (e.g., 64). During training and prediction, treat one group as a unit and change the label from whether it is correct to whether it is selected.

2. **Output module:** Add a Softmax activation layer at the last layer of the model's binary classification output head, ensuring that the sum of the output probabilities for the four options in each group is one.

Using the aforementioned method, both model training and prediction will treat the four options as a group and output the probabilities of the four options. To generalize our designed method, it needs to satisfy two assumptions: (1) When a student handles an MCQ, there is no time difference in learning gain among the options. In other words, the student has considered all four options when making a choice. (2) Each option has one and only one KC representation, which can be a skill or a

misconception. This method can also be extended to different numbers of classifications for MCQs by modifying the parameters involved in the method. The mathematical expression of the OT model we designed is as follows.

Defining the OT scenario for four options, each corresponding to a KC, involves using the student's historical discrete time steps of questions and corresponding choices to estimate the student's current KS in order to predict the student's choice for the next question. The sequence of problems in the student's historical interactions is represented as questions 1 to n: $Q_1, Q_2, \ldots, Q_n$. The sequence of historical choices is $q(1,1), q(1,2), q(1,3), q(1,4), \ldots, q(n,1), q(n,2), q(n,3), q(n,4)$, and the KC sequence corresponding to these choices is $c(1,1), c(1,2), c(1,3), c(1,4), \ldots, c(n,1), c(n,2), c(n,3), c(n,4)$. The time sequence corresponding to these choices is $t(1,1), t(1,2), t(1,3), t(1,4), \ldots, t(n,1), t(n,2), t(n,3), t(n,4)$, where for any question index i, $q(i,1) + q(i,2) + q(i,3) + q(i,4) = 1$ and $t(i,1) = t(i,2) = t(i,3) = t(i,4)$. The OT needs to predict the probability vector of the student's next question choices, $q(n+1,1), q(n+1,2), q(n+1,3), q(n+1,4)$.

For a question bank of m questions, where the index sequence of correct options is $y_1, y_2, \ldots, y_m$, the model training obtains the student's implicit KT (for DKT, it is represented as a knowledge-embedding vector) to predict the probabilities of the student's choices for all questions. The average value of these probabilities represents the student's explicit KS, KS = Mean$(Qy_1, Qy_2, ..., Qy_m)$, and it can also be statistically divided by skill to obtain the mastery level of different skills.

4.2 Results

To fairly compare the performance of KT and the improved OT method, we converted the predicted option probabilities into the corresponding student's KT correct answer rates (binary classification). We also provided metrics for option selection (four classifications). The evaluation metrics include the commonly used AUC, F1, and accuracy for KT assessments. The results are shown in Table 2.

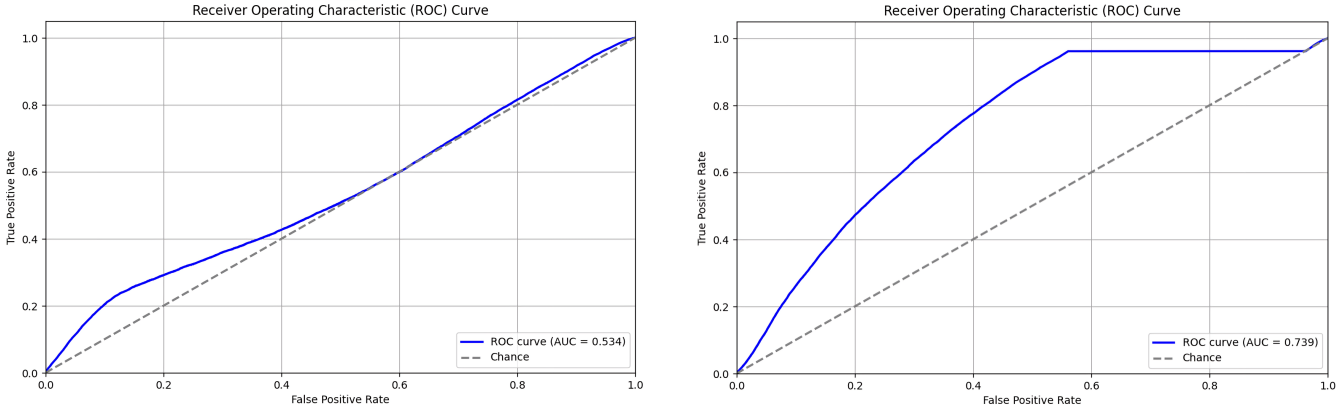
Table 2. Performance metrics for different models across KT and OT tasks.

	KT (Binary)				KT (Binary)			OT (Four-Class)		
Model	**AUC**	**F1**	**Accuracy**	**Model**	**AUC**	**F1**	**Accuracy**	**AUC**	**F1**	**Accuracy**
ATK	**0.561**	0.740	0.650	**AOT**	0.718	0.747	0.700	0.785	0.577	0.848
ATDKT	0.534	0.686	0.527	**ATDOT**	**0.739**	**0.777**	**0.712**	**0.810**	**0.604**	**0.849**
SimpleKT	0.560	0.745	0.657	**SimpleOT**	0.624	0.678	0.660	0.682	0.485	0.834
SparseKT	0.511	0.696	0.587	**SparseOT**	0.667	0.721	0.683	0.721	0.534	0.837
FoLiBiKT	**0.561**	**0.754**	**0.669**	**FoLiBiOT**	0.662	0.702	0.670	0.719	0.537	0.842

Comparing the KT (binary) metrics between OT models and KT models, such as comparing ATK and AOT or ATDKT and ATDOT, we observe that aside from a decrease in F1 scores for FoLiBiKT and SimpleKT, all other models and metrics show an increase. This indicates the effectiveness of our approach, as OT methods generally enhance the performance of KT (binary). The lower F1 scores for FoLiBiKT and SimpleKT might be explained by the sensitivity of F1 to class imbalances, particularly the issue of having 3 zeros and 1 one. Similar patterns can be observed by looking at the OT (four-class) metrics. For AUC and accuracy, OT's four-class predictions generally outperform KT's binary predictions, while F1 shows the opposite trend. This can also be due to the fixed ratio of 3 zeros to 1 one, which causes imbalances that affect classification performance.

The assessment of KS requires the predicted probabilities from KT because KS is calculated based on probability values of students' correct rates. Similarly, AUC is calculated based on the order of probability values and can be considered the most important KT metric. Meanwhile, F1 and accuracy serve as supplementary classification metrics. Among the KT (binary) metrics, SimpleOT, which does not include KC embedding, performs the worst. The best-performing ATDOT and the second-best ATK have a common feature: they consider the knowledge similarity between questions, including the embedding encodings of the items and KC. ATDOT achieved the highest AUC of 0.7301 and was selected as the model for subsequent RQ2 analysis.

We present the AUC metric curves for ATDKT and ATDOT models in Figure 2. Compared to the ATDKT model, the ROC curve of the ATDOT model rises more sharply near the midpoint. This indicates that the model achieves a higher true positive rate (TPR), which reflects the model's ability to correctly identify positive samples. It can be inferred that, in addition to the correct skill options, the model, including three distractor misconceptions as auxiliary information, can more accurately predict whether students can correctly answer the question. However, the false positive rate (FPR) has not significantly improved (as TPR increases rapidly within a lower FPR range), indicating that the model's ability to identify students choosing incorrect options still needs enhancement.



(a) ATDKT. **(b)** ATDOT.

Figure 2. ROC curve of two models in KT (binary).

5. RQ2: Learning Pattern Analytics

5.1 Methods

The purpose of this section is to use the best OT model in RQ1 to assess knowledge gains and analyze the problem-solving patterns of students with high learning gains. These findings should aid in question-recommendation systems or instructional design. First, we set a generalizable scenario: within a student's sequence of problem solving, the learning platform can assign a set of problems for each learning cycle. The number of problems is determined by the teacher, typically set at 10, 20, or 50 problems per week based on experience; this study uses 20 problems. The process of calculating learning gains is as follows: a student's initial KS is determined by 10 problems. After the student works on 20 problems, the final KS is predicted based on the 30 problems, and the difference between the two KS measurements represents the learning gain. We selected the initial KS and learning cycle to favour beginners in mathematics since it is based on the first 30 problems practised on the learning platform. However, the model can adjust parameters to generalize this method to more scenarios. In the subsequent linear regression analysis, we use the final KS as the dependent variable.

In addition to learning gains, we also extracted the problem-solving patterns of students on 20 problems. We started by statistically analyzing the characteristics of the problems students encounter, including the number of different skills and misconceptions. Then, we added the student's initial KS and the accuracy of the problems solved. After incorporating these four basic features, we added four features of the behavioural sequence for the 20 problems. For two responses to adjacent problems involving the same skill, based on the assumption, a student may have four outcomes: Right-Right (high-level behaviour), Wrong-Right (progressive behaviour), Right-Wrong (forgetting behaviour or insufficient knowledge mastery), and Wrong-Wrong (insufficient knowledge mastery). Among these, progressive behaviour is significant for measuring mastery level, such as determining if a student understands a specific knowledge concept to adjust the learning pace (Shen et al., 2022). On the other hand, a student's knowledge proficiency might decline due to the forgetting effect (Adler & Clark, 1991; Ebbinghaus, 2013). Therefore, we counted the occurrences of the four outcomes for problems with the same skill and converted them into percentages. In this way, we used eight features to describe students' learning patterns.

5.2 Results

In the results of learning gains, the average level of 4,869 students increased (mean $= 0.072, \text{SD} = 0.080$), with a mean initial KS of 0.728 (SD $= 0.092$). Conversely, the average level of 9,266 students decreased (mean $= -0.070, \text{SD} = 0.062$), with a mean initial KS of 0.811 (SD $= 0.044$). It appears that students with lower initial levels have more room for improvement. Moreover, the larger number of students with a decline in average levels suggests that the performance on the subsequent 20 questions was worse for most students compared to their initial 10 questions. The results of the linear regression, with Final KS as the dependent variable and eight learning pattern characteristics as independent variables, are presented in Table 3.

It can be observed in Table 3 that the initial KS and correct rate are positively correlated with the final KS, and the positive coefficients are statistically significant. The coefficients for the number of skills and the number of misconceptions are not significant, indicating that a diverse range of practice questions does not have a statistically significant association with learning gains. Next, we analyze the four types of question-answering behaviour patterns. All coefficients are significant and positive. On one hand, this indicates that diverse learning patterns are beneficial. On the other hand, given that the number of 20 practice questions is fixed, it means we should measure the contribution of different behaviours to learning gains based on the magnitude of the coefficients. The coefficients for Right-Wrong and Wrong-Wrong are higher than those for Right-Right and

Table 3. Linear regression results for learning pattern analytics.

Feature	Coef.	Std.Error	2.5% CI	97.5% CI
Initial KS	0.035***	0.007	0.022	0.049
Number of skills	0.002	0.002	−0.002	0.006
Number of misconceptions	<0.001	<0.001	−0.000	0.000
Correct rate	0.103***	0.019	0.066	0.140
Right-Right	0.111***	0.011	0.089	0.132
Wrong-Right	0.109***	0.008	0.094	0.124
Right-Wrong	0.152***	0.009	0.134	0.169
Wrong-Wrong	0.162***	0.008	0.145	0.178

Note. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Wrong-Right, by nearly half. This suggests that errors play a crucial role in learning gains. Previous research indicates that errors are considered a natural and important element of the learning process (Türkdoğan et al., 2013). Students can learn from their mistakes and facilitate knowledge acquisition through a favourable error environment. Additionally, for Right-Wrong, although we cannot be certain whether it is due to forgetting or not mastering the same skill type, the higher coefficient value implies that even repeated mistakes can still lead to greater subsequent learning gains.

In summary, given the same initial KS, a higher correct rate during the learning phase and more Right-Wrong and Wrong-Wrong response records can result in a higher final KS (i.e., greater learning gains).

5.3 Visual Analytics of Students' KSs

Our model can assess students' KSs to monitor learning progress. Student progress is significant for both learning and teaching, as it not only indicates the current KSs of the students but also shows whether their learning speed meets the targets. To demonstrate that the OT designed in this paper can capture meaningful student progress, we will select one student for a visualization analysis of learning progress. Following similar methods as in previous works (Shen et al., 2022; Abdelrahman & Wang, 2019), we calculate the change in KS for each question practised by the student, resulting in a heatmap composed of 20 squares for 20 questions. For KS, as previously mentioned, mastery can be expressed based on the accuracy of questions split into different skills. We select the top seven skills with the highest frequency (mean frequency as the threshold) out of 14 skills for radar chart visualization, and we separately plot radar charts for the initial and final KS.

First, according to the heatmap in the upper part of Figure 3, we can identify some progress patterns from the correct and incorrect interactions. Initially, the student completed eight consecutive basic arithmetic questions, with four correct and four incorrect answers. The KS fluctuated, showing similar changes whether the answers were initially correct followed by incorrect, as seen in questions 1 and 2, or vice versa, as seen in questions 4 and 5 and questions 6 and 7. The KS decreased after a correct response to the first question and increased after an incorrect response to the second question, suggesting that the model's assessment of KS is not solely dependent on correctness but also on the statistical impact of prior knowledge. One possible explanation is that overlapping knowledge learning could lead to diminishing returns and slowing progress. Learning from mistakes on challenging questions might yield greater knowledge benefits than repeated correct practice of the same skill. Next, the student tackled three questions on indices, powers, and roots (blocks 9 to 11), all of which were incorrect, leading to a decline in KS. Finally, the student completed nine basic arithmetic questions. Despite errors causing changes of 0.06 and 0.09 (blocks 12 and 13), KS still increased, indicating that students can gain knowledge from mistakes and mitigate knowledge decay. Despite multiple incorrect answers, the student answered questions 16 and 20 correctly, resulting in a KS increase, aligning with the learning curve where students often arrive at the correct answer after several failures, learning from their mistakes.

In the lower part of Figure 3, the student's initial KS is 0.832 (rounded to two decimal places in the figure but described here with three decimal places), corresponding to the radar chart on the left, and the final KS is 0.917, shown on the right radar chart, indicating a KS improvement of 0.085 through practice. All of the student's skills improved, with the most significant gain in Factors, Multiples, and Primes, from 0.825 to 0.937. Although the question sequence only included two skills, different skills can influence and promote each other, highlighting the interaction between skills.

6. RQ3: Skill Network Analytics

6.1 Methods

Based on KS results in RQ2, we further investigated whether our model can uncover intrinsic relationships among skills and capture meaningful information. In practice, for certain skills A and B, if learners struggle to comprehend skill A, they often find skill B challenging to grasp. In such cases, these two skills are typically closely related, potentially having prerequisites or

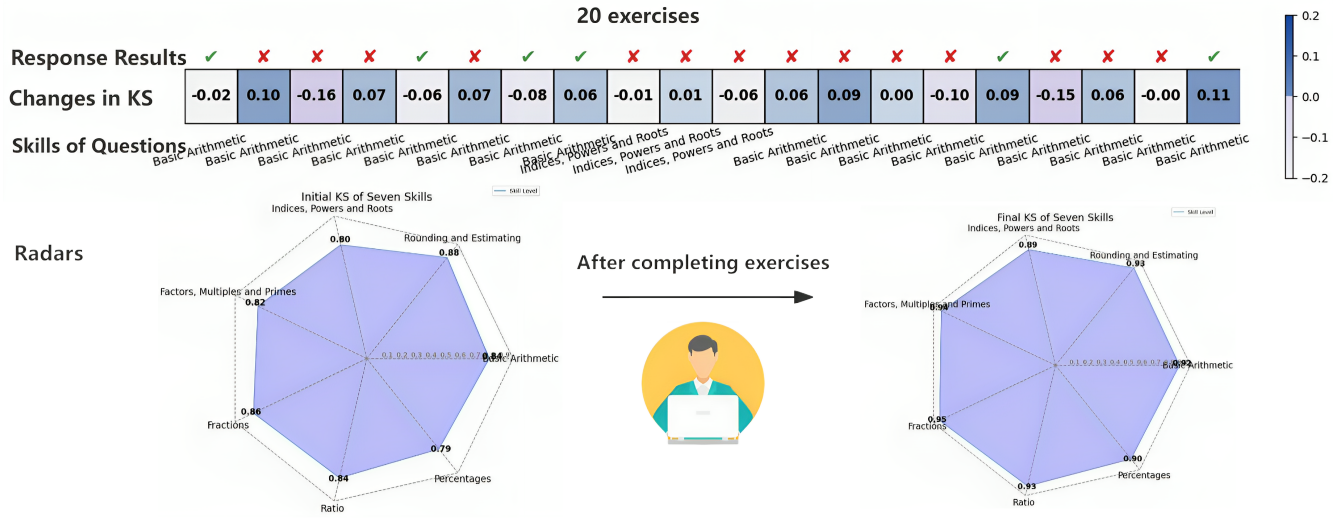
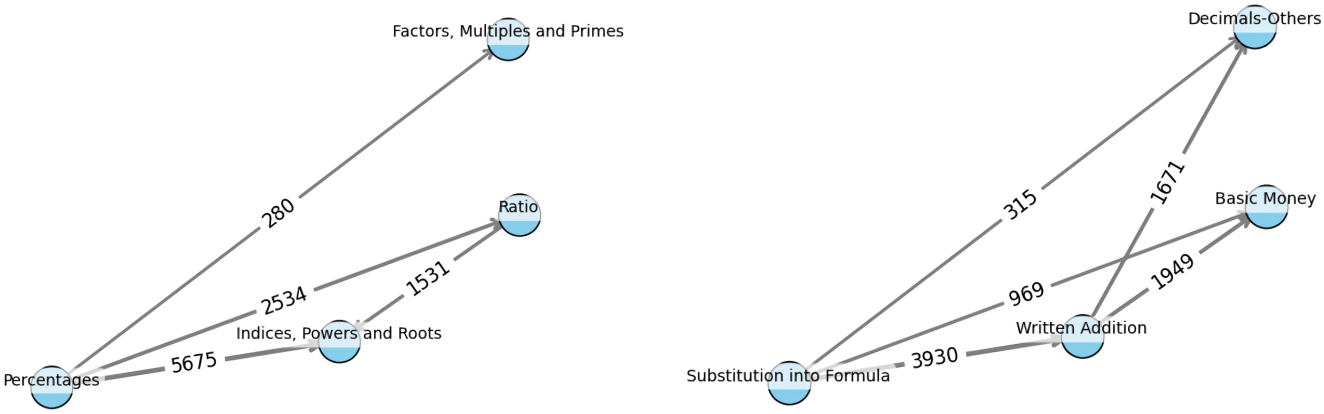


Figure 3. The heatmap of the student's KS changes and the radar charts of skills.

inclusion relationships, with skill A considered a prerequisite for skill B (Lu et al., 2023). We devised a novel approach to identify such relationships between skills by leveraging students' KSs. Specifically, the two skills with the poorest student performance may indicate a prerequisite relationship. Assuming each skill has at least one prerequisite, the less poor skill might serve as a prerequisite for the poorer skill. Thus, each student's final KS can be assigned a potential prerequisite relationship for one skill. Through this, each student contributes a vote count for associated skill relationships, with the most votes determining the final result for each skill. In a directed graph, such skill relationships can also be termed *adjacent skills*. Skills and their adjacent skills act as nodes, with directed links forming edges. Figure 4 depicts the graphical representation generated using the network analysis tool NetworkX based on primary and secondary skills, where the edge threshold is set to 150 (i.e., 10% sample proportion).



(a) Primary skills. **(b)** Secondary skills.

Figure 4. Skill network visualization. The edge is directed, and its weight represents the number of votes; the greater the quantity, the higher the degree of relevance.

6.2 Results

In Figure 4, we observe meaningful connections between skills. It is evident that Percentages are crucial prerequisites for three primary skills. Within secondary skills, Substitution into Formula is a significant prerequisite for three skills, including Written Addition, and it also shows a progressive relationship where Written Addition serves as a prerequisite not only for Basic Money but also for Decimals-Others. Additionally, Ratio occupies a central position among two skills (subplot (a)), while Written Addition holds a central position among three skills (subplot (b)), suggesting its bridging role in learning these skills. Considering that our method selects only the most relevant and poorest-performing skills, this stringent approach may limit the identification of additional skill relationships. Our network graph analytics reveals interconnections among skills, which could

aid in recommending question assignments for adaptive systems and designing learning pathways in education.

7. Discussion and Future Works

This paper designs a general method that can modify DKT models into DOT models, including data processing and network output modules. The results show that in the context of K–12 mathematics, MCQs, based on the high-quality large dataset provided in this paper, which includes two types of KCs (mathematical skills and misconceptions), the performance of DOT surpasses that of DKT. Moreover, DOT is more advantageous from the perspective of LA—its option prediction is particularly useful for analyzing answering patterns, especially when it involves error analysis. On the one hand, this work contributes to artificial intelligence in education (AIED) by promoting the application of KT in multiple-choice scenarios through technological innovation, potentially benefiting educational resource recommendation systems and intelligent tutoring systems. On the other hand, this work contributes to LA by allowing DOT to perform more detailed option analyses based on response patterns. The results indicate that errors involving the same skill may have greater potential for learning gains. Additionally, the designed network analytics method establishes relationships between different skills based on students' KSs, revealing that Percentages may be a key skill affecting student performance. These technical methods and analyses provide new perspectives for learning pattern analytics and instructional design.

KS assessment plays a crucial role in MCQ scenarios. From the perspective of cognitive psychology, designing well-structured sequences of MCQs can reflect students' decision-making processes and cognitive strategies when facing complex problems (Coderre et al., 2004). When students select among multiple options, they must integrate varied knowledge and skills to filter and judge the information. KT technology not only displays students' mastery of specific knowledge points but also reveals their thinking processes and strategy applications in problem solving. It can infer their misunderstandings or knowledge gaps and design targeted interventions. Moreover, traditional assessment methods work well but can only offer static, overall performance evaluations, whereas KT can reflect students' learning progress and knowledge changes in real time. This dynamic assessment not only helps teachers adjust teaching strategies in a timely manner and provide personalized instruction based on the specific needs of different students (Du et al., 2022) but also assists students in self-monitoring and adjusting their learning pace, thus improving learning efficiency and outcomes.

This study's analysis of learning patterns reveals that multiple errors in the same skill can lead to greater subsequent learning gains, with misconceptions being a typical form of mathematical error. According to the theoretical framework of conceptual change (Vosniadou, 1994; Stafylidou & Vosniadou, 2004), learning, which requires the restructuring of existing knowledge, is inherently challenging and may result in misconceptions. However, these misconceptions can, in fact, facilitate the assimilation of new information into a student's existing knowledge base. Notably, this framework has already been applied and validated in the field of mathematics, where conceptual learning plays a significant role because many misconceptions can be observed in this subject (Mishra, 2020). Furthermore, misconceptions are one of the main sources of systematic errors in mathematics (Rakes & Ronau, 2019). When learners' understanding of a concept fundamentally differs from its scientific meaning, misconceptions are likely to arise. Additionally, due to the spiral nature of the mathematics curriculum—characterized by the overlapping of concepts—it is almost impossible to define any concept in mathematics without using other concepts (Crooks & Alibali, 2014). As a result, students' misconceptions about previous mathematical topics may introduce new misconceptions into newly learned topics (Biber et al., 2013), a manifestation of constructivism. Conversely, a high level of conceptual understanding can help students solve various forms of mathematical problems, including new problems they have never encountered before (DiNapoli & Miller, 2022). Therefore, learning mathematical concepts is crucial for students, and fully understanding misconceptions may provide a more efficient method to support students' understanding of the subject. For example, one study (Isotani et al., 2010) developed an intelligent tutoring system that linked elementary students' mathematical errors to corresponding misconceptions. This system adjusted feedback, including recommending subsequent questions, to help students better understand decimal concepts and address specific misconceptions arising from internalized incorrect behaviours and thought processes. The study demonstrated promising initial results for the intelligent tutor. Similarly, another study (Resnick et al., 1989) encourages educators to use mathematical misconceptions as diagnostic tools to assess the nature of students' understanding of mathematical topics by analyzing the conceptual origins of their errors. Yet another study (McLaren et al., 2012) explored the impact of interactive erroneous examples on middle school students' learning of decimals. This study provided students with deliberately incorrect examples of decimal problems, which they were tasked to explain and correct, thereby engaging in deep metacognitive processes. The findings revealed that students who interacted with these erroneous examples significantly outperformed those who engaged in traditional problem solving in terms of long-term retention and conceptual understanding. This approach not only reinforces the conceptual change framework by using misconceptions as learning tools but also suggests a practical method for applying this theory in educational settings, particularly in subjects like mathematics where misconceptions are prevalent and often deeply rooted. Consequently, teachers need to plan instructional and intervention activities by considering students' KSs and misconceptions. In this context, data-driven approaches like OT can help educators understand the thoughts behind students' errors, turning learning obstacles into integral parts of mathematics

learning and teaching.

Some educators argue that having students review and discuss their errors can be valuable for learning (McLaren et al., 2012). In particular, mathematics education may benefit from students engaging with their mistakes, as this practice encourages critical thinking about mathematical concepts and stimulates reflection and inquiry. On one hand, incorrect examples can provide a deeper learning experience by helping students build upon their initial understanding of decimals, gradually leading to a more profound comprehension over time. Moreover, such examples may foster generative processing, which involves deeper cognitive engagement with instructional materials by organizing them and linking them to prior knowledge. One explanation for generative processing is based on the "desirable difficulty" effect (Schmidt & Bjork, 1992), whereby errors generated from challenging problems may enhance long-term memory. Specifically, engaging in exercises that require explaining and correcting mistakes can help students process mathematical problems at a deeper level, aiding them in overcoming misconceptions and establishing a lasting understanding of mathematical concepts (Adams et al., 2014). On the other hand, directly addressing students' misconceptions may force them to recognize the inaccuracies in their prior beliefs (Cunningham et al., 2004), necessitating the adoption of new concepts or improved knowledge calibration. Furthermore, in teaching, by directing students' attention to common errors through incorrect examples, they are more likely to engage deeply with correct concepts (VanLehn, 1999) and develop new mental representations of the material. In support of this, some studies have already demonstrated the importance of mathematical errors in promoting learning (McLaren et al., 2015). For instance, one study, using Bayesian tracing to assess mathematical skills (McLaren et al., 2012), found that when sixth- and seventh-grade students encountered errors while solving math test problems, their likelihood of making common mistakes subsequently decreased. Additionally, another qualitative experiment (Durkin & Rittle-Johnson, 2012) found that contrasting incorrect examples with correct ones helped students learn accurate concepts and procedures. As a result, incorrect examples can lead to greater procedural knowledge, higher conceptual understanding, and fewer misconceptions following the post-test.

One possible reason for errors, besides misunderstanding, is the forgetting effect. The process of students' knowledge construction is not static but evolves over time (Chen et al., 2017). This evolution is due to the fact that students learn and forget over time, making it more challenging to track their knowledge. These processes of learning and forgetting have been validated by two classic theories in educational psychology: the learning curve theory (Yang et al., 2022) and the Ebbinghaus forgetting curve theory (Ebbinghaus, 2013). The former uses learning curves to represent how performance improvements result from more extensive trials or practice. The latter posits that students' KSs or knowledge levels change over time. Research has already shown that the frequency with which students learn a target skill affects forgetting behaviour: the more they practise, the less likely they are to forget (Farr, 1987). Future work can consider incorporating forgetting-related behaviours to achieve accurate knowledge modelling and further analyze students' error patterns. For instance, simulating students' forgetting behaviour during intervals between exercises and considering the lag time of different interactions with the same skill on error rates could be investigated.

This study also employs automatic visualization techniques of prerequisite skills to analyze the relationships between different skills, aiming to provide references for teachers' instructional design. According to the educational theory of knowledge transfer (Thorndike & Woodworth, 1901; Britton, 2002), when learners acquire a skill, it can not only affect the proficiency of the current skill but also induce changes in related skills. Mathematical skills and concepts are particularly important when students solve similar problems, because knowledge can transfer between skill concepts, and the effects of learning can propagate through multiple relationships within the knowledge structure. Therefore, in using knowledge structures for KT, it is possible to consider the propagation of these effects to achieve more accurate KS assessments. Future work can dynamically modify a student's KS in a prerequisite or related skill based on their interactions with other relevant skills, thereby allowing the model to benefit from the student's interactions with different skills throughout their entire learning sequence. Additionally, considering that metacognition involves students' ability to self-regulate problem-solving strategies (Dixon & Brown, 2012), it is necessary to abstract common themes from problem-solving experiences. This approach emphasizes the need to consider common patterns of students' misconceptions, specific subject errors, and difficulties in teaching (Rakes & Ronau, 2019; Hayati & Setyaningrum, 2019). Understanding how mathematical misconceptions relate to other topics within the overall conceptual framework can help identify broader patterns in students' long-term mathematical development.

Future work could incorporate the OT method designed in this paper to enhance mathematics teaching strategies, particularly by focusing on learning from mistakes. Some studies already support this perspective. For example, some researchers have begun developing curricula that encourage teachers to integrate incorrect examples into their mathematics lessons (Curtis et al., 2009). Another study found that presenting students with incorrect examples in an interactive intelligent tutoring system could improve their metacognitive skills, problem-solving abilities, and conceptual understanding (Tsovaltzi et al., 2010). However, the related teaching strategies remain challenging. For instance, teaching decimals is complex because teachers are not always aware of common misconceptions and may incorrectly attribute wrong answers to incorrect underlying misunderstandings (Stacey et al., 2001). Research suggests that simple correction of students' errors, without addressing underlying misconceptions, only demonstrates a basic level of content knowledge depth. The highest level of content knowledge depth is achieved only

when misconceptions are conceptually identified and corrected (Webb, 2002). Teachers need to help learners correctly learn the concepts by correcting the misconceptions found in their cognitive structures (Chi, 2005). Misconceptions should be marked as such in the learners' memory (Van Den Broek & Kendeou, 2008), and the likelihood of selecting incorrect options in the future should be reduced (Siegler, 2002). Therefore, one potential direction for improving teaching strategies is to present students with incorrect examples. Incorrect examples may help students better calibrate their knowledge by emphasizing the critical features of both incorrect and correct examples, thus helping them to recognize correct concepts as correct and incorrect concepts as incorrect (Van Den Broek & Kendeou, 2008). In this process, reducing mathematical errors involves guiding students to think freely and apply the skills acquired from learning (Rochmad et al., 2018). It is also important to note that some studies have found that students with low prior knowledge may not have sufficient mathematical understanding to comprehend and learn from incorrect examples. Their ability to learn from errors is lower than that of students with higher mathematical proficiency (Große & Renkl, 2007). The potential importance of prior knowledge highlights that simply exposing students to incorrect examples may not be sufficient to improve learning outcomes, as students may not understand why an example is incorrect (Stark et al., 2011). Future research could analyze learning outcomes based on students' individual levels of mathematical literacy. More detailed investigations into students' participation patterns, such as engagement time recorded in learning logs (Lyu et al., 2024) or an analysis of students' response behaviours combined with qualitative teacher feedback (Li, Xing, Li, Zhu, & Heffernan, 2024) could provide valuable insights into improving engagement. Moreover, education technologists could develop automated feedback technologies to deliver detailed explanations of students' errors (Li, Li, et al., 2024) or employ automated quality assessment methods (Li, Guo, et al., 2024) to design more effective distractors. For instance, multimodal assessments of item difficulty and cognitive load (integrating both visual and textual elements) could facilitate the creation of tasks with higher learning potential (Li, Xing, Li, Zhu, & Oh, 2024).

Although the scenario conditions of our designed method are limited to a learning environment where only one correct option is present in MCQs and where the options must include KCs representing both skills and misconceptions, which restricts its practicality in many learning contexts, our method has substantial potential to be adapted to other OT scenarios. Future work could explore two other commonly used scenarios. The first common scenario involves situations where some options do not necessarily contain KCs. In such cases, future work could involve predicting common error types and skill types to serve as the KCs for the options to be tracked. This aligns with a typical educational text classification task. The second common scenario addresses the application of the method to multiple-selection MCQs, where the number of correct answers and the number of student selections can vary (e.g., 1, 2, 3, or even 4). In this case, we need to adapt the model to a multi-label framework and design a new output module to predict the number of options a student may choose. A simple approach would be to duplicate the existing output module, modify the number of neurons in the output to one, and introduce new labels during training that correspond to the number of options selected by the student. This would enable tracking more complex multiple-selection MCQs. If the model predicts that a student selects N options, the output should then choose the N options with the highest probabilities.

This study has several limitations. To begin with, from the perspective of the model, both KT and OT assess students' KSs in the form of time series. Specifically, breaking a single question into four options expands the sequence length by four times, which not only affects the speed and space of algorithm prediction but also reduces the accuracy of predictions due to the excessively long sequences. In other words, overly long time series lead to information overload, and the inclusion of too many KCs increases the task's difficulty, resulting in decreased accuracy in KS assessments. Therefore, in this study, we controlled the sequence length for beginner math students to within 30*4. However, as the sequence length increases, evaluating more proficient learners requires an even longer time series, posing a challenge for OT. Consequently, future work could explore how to optimize learning process modelling to accommodate OT's longer time series. For example, one possible approach is to truncate sequences and retain only the most recent 30 responses while retraining the model to handle cases where students have completed a much larger number of questions (e.g., 100). This method could involve incorporating new KC parameters, such as the number of questions truncated and the corresponding accuracy rate. In doing so, the model could maintain high effectiveness even with extended sequences. In practice, the initial KS and the total number of questions should align with multiples of the set of exercises provided to students by the platform. This allows for a comprehensive evaluation of the student's KS within a truncated sequence. Moreover, expanding the model's applicability requires consideration of scenarios where students' problem-solving sequences extend significantly. Given the model's capacity constraints, our approach suggests focusing on the most recent sequence while integrating earlier sequences as KCs to ensure a more comprehensive data consideration. From the perspective of the analytical method, it is essential to recognize that students' learning gains are influenced not only by the patterns of problem-solving sequences but also by students' demographic attributes, such as grade level and gender. Therefore, including these attributes in the analysis will enhance the value and applicability of the research findings. Moving forward, future work can further improve its effectiveness in personalized learning by using dynamic real-time intelligent tutoring systems based on KT. These systems can adjust teaching strategies according to learners' demographic attributes, KSs, and real-time responses to problems, thereby enhancing engagement and learning outcomes.

8. Conclusion

KT and OT are both pivotal technologies in innovative educational technology. An OT-based KS assessment system can provide targeted feedback and suggestions based on students' choices, while learning pattern analytics can identify students' learning obstacles, aiding teachers in optimizing teaching effectiveness. This paper proposes a general method for converting KT to OT, including data processing and network output modules. The results demonstrate that OT not only identifies students' option choices but also achieves better performance in KS assessment. Based on our OT model, learning pattern analytics suggests that for beginners in middle school mathematics, repeated mistakes in the same skill have the potential to yield greater learning gains. Skill network analytics uncovers the relationships between skills based on students' error tendencies in learning patterns; for example, Percentages might be a key skill affecting student performance. The proposed technical framework and analysis results can contribute to research on KS assessment in AIED, as well as learning pattern analytics in LA.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work is supported by the Learning Engineering Virtual Institute under grant number G-23-2137070 and the Jaffe Foundation under grant number AGR00026932. Any opinions, findings, and conclusions or recommendations expressed in this paper, however, are those of the authors and do not necessarily reflect the views of the funding agency.

References

Abdelrahman, G., & Wang, Q. (2019). Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2019), 21–25 July 2019, Paris, France (pp. 175–184). ACM. https://doi.org/10.1145/3331184.3331195

Abdelrahman, G., & Wang, Q. (2023). Learning data teaching strategies via knowledge tracing. *Knowledge-Based Systems*, *269*, 110511. https://doi.org/10.1016/j.knosys.2023.110511

Abdelrahman, G., Wang, Q., & Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, *55*(11), article number 93. https://doi.org/10.1145/3578362

Abida, K., Azeem, M., & Gondal, M. B. (2011). Assessing students' math proficiency using multiple-choice and short constructed response item formats. *International Journal of Technology, Knowledge and Society*, *7*(3), 135. https://doi.org/10.18848/1832-3669/CGP/v07i03/56206

Adams, D. M., McLaren, B. M., Durkin, K., Mayer, R. E., Rittle-Johnson, B., Isotani, S., & Van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior*, *36*, 401–411. https://doi.org/10.1016/j.chb.2014.03.053

Adler, P. S., & Clark, K. B. (1991). Behind the learning curve: A sketch of the learning process. *Management Science*, *37*(3), 267–281. https://doi.org/10.1287/mnsc.37.3.267

An, S., Kim, J., Kim, M., & Park, J. (2022). No task left behind: Multi-task learning of knowledge tracing and option tracing for better student assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(4), 4424–4431. https://doi.org/10.1609/aaai.v36i4.20364

Biber, C., Tuna, A., & Korkmaz, S. (2013). The mistakes and the misconceptions of the eighth grade students on the subject of angles. *European Journal of Science and Mathematics Education*, *1*(2), 50–59. https://doi.org/10.30935/scimath/9387

Borasi, R. (1996). *Reconceiving mathematics instruction: A focus on errors. Issues in curriculum theory, policy, and research series*. Ablex Publishing.

Britton, S. (2002). Are students able to transfer mathematical knowledge? *Second International Conference on the Teaching of Mathematics*, 1–6 July 2002, Hersonissos, Crete, Greece. http://users.math.uoc.gr/~ictm2/Proceedings/pap267.pdf

Cai, D., Zhang, Y., & Dai, B. (2019). Learning path recommendation based on knowledge tracing model and reinforcement learning. In *Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications* (ICCC 2019), 6–9 December 2019, Chengdu, China (pp. 1881–1885). IEEE. https://doi.org/10.1109/ICCC47050.2019.9064104

Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y., & Hu, G. (2017). Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (CIKM 2017), 6–10 November 2017, Singapore (pp. 989–998). ACM. https://doi.org/10.1145/3132847.3132929

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*(2), 161–199. https://doi.org/10.1207/s15327809jls1402_1

Coderre, S. P., Harasym, P., Mandin, H., & Fick, G. (2004). The impact of two multiple-choice question formats on the problem-solving strategies used by novices and experts. *BMC Medical Education*, *4*, article number 23. https://doi.org/10.1186/1472-6920-4-23

Corbett, A. T., & Anderson, J. R. (1995). Knowledge decomposition and subgoal reification in the ACT programming tutor. In J. Greer (Ed.), *Artificial Intelligence and Education, 1995: The Proceedings of World Conference on Artificial Intelligence in Education* (AI-ED 1995), 16–19 August 1995, Washington, DC, USA. AACE.

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, *34*(4), 344–377. https://doi.org/10.1016/j.dr.2014.10.001

Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Stanovich, P. J. (2004). Disciplinary knowledge of K–3 teachers and their knowledge calibration in the domain of early literacy. *Annals of Dyslexia*, *54*, 139–167. https://doi.org/10.1007/s11881-004-0007-y

Curtis, D. A., Heller, J. I., Clarke, C., Rabe-Hesketh, S., & Ramirez, A. (2009). The impact of *Math Pathways and Pitfalls* on students' mathematics achievement. In *Proceedings of the Annual Meeting of the American Educational Research Association*, 13–17 April 2009, San Diego, CA. AERA.

DiNapoli, J., & Miller, E. K. (2022). Recognizing, supporting, and improving student perseverance in mathematical problem-solving: The role of conceptual thinking scaffolds. *The Journal of Mathematical Behavior*, *66*, 100965. https://doi.org/10.1016/j.jmathb.2022.100965

Dixon, R. A., & Brown, R. A. (2012). Transfer of learning: Connecting concepts during problem solving. *Journal of Technology Education*, *24*(1), 2–17. https://doi.org/10.21061/jte.v24i1.a.1

Du, H., Xing, W., & Zhang, Y. (2022). Misconception of abstraction: When to use an example and when to use a variable? In J. Vahrenhold, K. F. ad Matthias Hauswirth, & D. Franklin (Eds.), *Proceedings of the 2022 ACM Conference on International Computing Education Research* (ICER 2022), 7–11 August 2022, Lugano, Switzerland, and online (pp. 28–29, Vol. 2). ACM. https://doi.org/10.1145/3501709.3544276

Durkin, K., & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, *22*(3), 206–214. https://doi.org/10.1016/j.learninstruc.2011.11.001

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences*, *20*(4), 155. https://europepmc.org/article/MED/25206041

Farr, M. J. (1987). *The long-term retention of knowledge and skills: A cognitive and instructional perspective*. Springer-Verlag.

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD 2020), 6–10 July 2020, online (pp. 2330–2339). ACM. https://doi.org/10.1145/3394486.3403282

Ghosh, A., Raspat, J., & Lan, A. (2021). Option tracing: Beyond correctness analysis in knowledge tracing. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *International conference on artificial intelligence in education. AIED 2021. Lecture notes in computer science* (pp. 137–149, Vol. 12748). Springer. https://doi.org/10.1007/978-3-030-78292-4_12

Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems. ITS 2010. Lecture notes in computer science* (pp. 35–44, Vol. 6094). Springer. https://doi.org/10.1007/978-3-642-13388-6_8

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction*, *17*(6), 612–634. https://doi.org/10.1016/j.learninstruc.2007.09.008

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hayati, R., & Setyaningrum, W. (2019). Identification of misconceptions in middle school mathematics utilizing certainty of response index. *Journal of Physics: Conference Series*, *1320*(1), 012041. https://doi.org/10.1088/1742-6596/1320/1/012041

Huang, S., Liu, Z., Zhao, X., Luo, W., & Weng, J. (2023). Towards robust knowledge tracing models via k-sparse attention. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2023), 23–27 July 2023, Taipei, Taiwan (pp. 2441–2445). ACM. https://doi.org/10.1145/3539618.3592073

Huang, Y., Hollstein, J. D. G., & Brusilovsky, P. (2016). Modeling skill combination patterns for deeper knowledge tracing. In *Late-Breaking Results, Posters, Demos, Doctoral Consortium and Workshops Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation* (UMAP 2016), 13–16 July 2016, Halifax, Nova Scotia, Canada. https://ceur-ws.org/Vol-1618/PALE4.pdf

Im, Y., Choi, E., Kook, H., & Lee, J. (2023). Forgetting-aware linear bias for attentive knowledge tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (CIKM 2023), 21–25 October 2023, Birmingham, UK (pp. 3958–3962). ACM. https://doi.org/10.1145/3583780.3615191

Isotani, S., McLaren, B. M., & Altman, M. (2010). Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems. ITS 2010. Lecture notes in computer science* (pp. 346–348, Vol. 6095). Springer. https://doi.org/10.1007/978-3-642-13437-1_66

Kramarski, B. (2004). Making sense of graphs: Does metacognitive instruction make a difference on students' mathematical conceptions and alternative conceptions? *Learning and Instruction*, *14*(6), 593–619. https://doi.org/10.1016/j.learninstruc.2004.09.003

Li, H., Guo, R., Li, C., & Xing, W. (2024). Automated quality assessment of multimodal mathematical stories generated by generative artificial intelligence. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale* (L@S 2024), 18–20 July 2024, Atlanta, Georgia, USA (pp. 110–121). ACM. https://doi.org/10.1145/3657604.3662029

Li, H., Li, C., Xing, W., Baral, S., & Heffernan, N. (2024). Automated feedback for student math responses based on multi-modality and fine-tuning. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (LAK 2014), 24–28 March 2014, Indianapolis, Indiana, USA (pp. 763–770). ACM. https://doi.org/10.1145/3636555.3636860

Li, H., Xing, W., Li, C., Zhu, W., & Heffernan, N. (2024). Positive affective feedback mechanisms in an online mathematics learning platform. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale* (L@S 2024), 18–20 July 2024, Atlanta, Georgia, USA (pp. 371–375). ACM. https://doi.org/10.1145/3657604.3664666

Li, H., Xing, W., Li, C., Zhu, W., & Oh, H. (2024). Are simpler math stories better? Automatic readability assessment of GAI-generated multimodal mathematical stories validated by engagement. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13554

Liu, Z., Liu, Q., Chen, J., Huang, S., Gao, B., Luo, W., & Weng, J. (2023). Enhancing deep knowledge tracing with auxiliary tasks. In Y. Ding, J. Tang, J. Sequeda, L. Aroyo, C. Castillo, & G.-J. Houben (Eds.), *Proceedings of the ACM Web Conference 2023* (WWW 2023), 30 April–4 May 2023, Austin, Texas, USA (pp. 4178–4187). ACM. https://doi.org/10.1145/3543507.3583866

Liu, Z., Liu, Q., Chen, J., Huang, S., & Luo, W. (2023). SimpleKT: A simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*. https://doi.org/10.48550/arXiv.2302.06881

Lu, Y., Wang, D., Chen, P., Meng, Q., & Yu, S. (2023). Interpreting deep learning models for knowledge tracing. *International Journal of Artificial Intelligence in Education*, *33*(3), 519–542. https://doi.org/10.1007/s40593-022-00297-z

Lyu, B., Li, C., Li, H., & Xing, W. (2024). Roles of joining time, technology use, and social interaction in sustaining student participation in an online mathematics discussion board. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale* (L@S 2024), 18–20 July 2024, Atlanta, Georgia, USA (pp. 398–402). ACM. https://doi.org/10.1145/3657604.3664672

McLaren, B. M., Adams, D., Durkin, K., Goguadze, G., Mayer, R. E., Rittle-Johnson, B., Sosnovsky, S., Isotani, S., & Van Velsen, M. (2012). To err is human, to explain and correct is divine: A study of interactive erroneous examples with middle school math students. In A. Ravenscroft, S. Lindstaedt, C. Delgado Kloos, & D. Hernández-Leo (Eds.), *21st century learning for 21st century skills. EC-TEL 2012. Lecture notes in computer science* (pp. 222–235, Vol. 7563). Springer. https://doi.org/10.1007/978-3-642-33263-0_18

McLaren, B. M., Adams, D. M., & Mayer, R. E. (2015). Delayed learning effects with erroneous examples: A study of learning decimals with a web-based tutor. *International Journal of Artificial Intelligence in Education*, *25*(4), 520–542. https://doi.org/10.1007/s40593-015-0064-x

Mishra, L. (2020). Conception and misconception in teaching arithmetic at primary level. *Journal of Critical Reviews*, *7*(5), 936–939. https://www.researchgate.net/publication/344066409_CONCEPTION_AND_MISCONCEPTION_IN_TEACHING_ARITHMETIC_AT_PRIMARY_LEVEL

Ndemo, Z., & Ndemo, O. (2018). Secondary school students' errors and misconceptions in learning algebra. *Journal of Education and Learning*, *12*(4), 690–701. https://doi.org/10.11591/edulearn.v12i4.9556

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*, 313–350. https://doi.org/10.1007/s11257-017-9193-2

Pu, Y., Wu, W., Han, Y., & Chen, D. (2018). Parallelizing Bayesian knowledge tracing tool for large-scale online learning analytics. In *Proceedings of the 2018 IEEE International Conference on Big Data* (Big Data), 10–13 December 2018, Seattle, Washington, USA (pp. 3245–3254). IEEE. https://doi.org/10.1109/BigData.2018.8622355

Rakes, C. R., & Ronau, R. N. (2019). Rethinking mathematics misconceptions: Using knowledge structures to explain systematic errors within and across content domains. *International Journal of Research in Education and Science*, *5*(1), 1–21. https://www.ijres.net/index.php/ijres/article/view/482

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084

Resnick, L. B., Nesher, P., Leonard, F., Magone, M., Omanson, S., & Peled, I. (1989). Conceptual bases of arithmetic errors: The case of decimal fractions. *Journal for Research in Mathematics Education*, *20*(1), 8–27. https://doi.org/10.2307/749095

Rochmad, R., Kharis, M., Agoestanto, A., Zahid, M. Z., & Mashuri, M. (2018). Misconception as a critical and creative thinking inhibitor for mathematics education students. *Unnes Journal of Mathematics Education*, *7*(1), 57–62. https://doi.org/10.15294/ujme.v7i1.18078

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Scruggs, R., Baker, R. S., & McLaren, B. M. (2019). Extending deep knowledge tracing: Inferring interpretable knowledge and predicting post-system performance. *arXiv preprint arXiv:1910.12597*. https://doi.org/10.48550/arXiv.1910.12597

Shen, S., Chen, E., Liu, Q., Huang, Z., Huang, W., Yin, Y., Su, Y., & Wang, S. (2022). Monitoring student progress for learning process-consistent knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 8213–8227. https://doi.org/10.1109/TKDE.2022.3221985

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). Cambridge University Press.

Stacey, K., Helme, S., Steinle, V., Baturo, A., Irwin, K., & Bana, J. (2001). Preservice teachers' knowledge of difficulties in decimal numeration. *Journal of Mathematics Teacher Education*, *4*, 205–225. https://doi.org/10.1023/A:1011463205491

Stafylidou, S., & Vosniadou, S. (2004). The development of students' understanding of the numerical value of fractions. *Learning and Instruction*, *14*(5), 503–518. https://doi.org/10.1016/j.learninstruc.2004.06.015

Stark, R., Kopp, V., & Fischer, M. R. (2011). Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. *Learning and Instruction*, *21*(1), 22–33. https://doi.org/10.1016/j.learninstruc.2009.10.001

Stout, W., Henson, R., & DiBello, L. (2023). Optimal classification methods for diagnosing latent skills and misconceptions for option-scored multiple-choice item quizzes. *Behaviormetrika*, *50*(1), 177–215. https://doi.org/10.1007/s41237-022-00172-0

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes. *Psychological Review*, *8*(4), 553–564. https://doi.org/10.1037/h0071363

Tsovaltzi, D., Melis, E., McLaren, B. M., Meyer, A.-K., Dietrich, M., & Goguadze, G. (2010). Learning from erroneous examples: When and how do students benefit from them? In M. Wolpers, P. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Sustaining TEL: From innovation to learning and practice. EC-TEL 2010. Lecture notes in computer science* (pp. 357–373, Vol. 6383). Springer. https://doi.org/10.1007/978-3-642-16020-2_24

Türkdoğan, A., Baki, A., & Çepni, S. (2013). The anatomy of mistakes: Categorizing students' mistakes in mathematics within learning theories. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *1*(1), 13–26. https://dergipark.org.tr/en/pub/turkbilmat/issue/21560/231415

Van Den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Applied Cognitive Psychology*, *22*(3), 335–351. https://doi.org/10.1002/acp.1418

VanLehn, K. (1999). Rule-learning events in the acquisition of a complex skill: An evaluation of Cascade. *The Journal of the Learning Sciences*, *8*(1), 71–125. https://doi.org/10.1207/s15327809jls0801_3

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*(March), 1–9. http://ossucurr.pbworks.com/w/file/fetch/49691156/Norm%20web%20dok%20by%20subject%20area.pdf

Yang, S., Liu, X., Su, H., Zhu, M., & Lu, X. (2022). Deep knowledge tracing with learning curves. In K. S. Candan, T. N. Dinh, M. T. Thai, & T. Washio (Eds.), *Proceedings of the 2022 IEEE International Conference on Data Mining Workshops* (ICDMW 2022), 28 November–1 December 2022, Orlando, Florida (pp. 282–291). IEEE. https://doi.org/10.1109/ICDMW58026.2022.00046

Yeung, C.-K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*. https://doi.org/10.48550/arXiv.1904.11738

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In H. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education. AIED 2013. Lecture notes in computer science* (pp. 171–180, Vol. 7926). Springer. https://doi.org/10.1007/978-3-642-39112-5_18