Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Tracing intermittent knowledge states based on knowledge-context-aware augmentation of Q-matrices

Lankun Guo

lankun.guo@gmail.com

Hunan Normal University

Jianghui Yi

Hunan Normal University

Jinhai Li

Kunming University of Science and Technology

Zhiqiang Yuan

Hunan Normal University

Research Article

Keywords: Knowledge tracing, Q-matrix, Item response theory, Knowledge concept

Posted Date: July 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9760933/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Tracing intermittent knowledge states based on knowledge-context-aware augmentation of Q-matrices

Lankun Guo[1*], Jianghui Yi[1], Jinhai Li[2], Zhiqiang Yuan[1]

[1] Key Laboratory of Computing and Stochastic Mathematics (Ministry of Education), School of Mathematics and Statistics, Hunan Normal University, Changsha, 410081, Hunan Province, China .

[2] Data Science Research Center and Faculty of Science, Kunming University of Science and Technology, Kunming, 650500, Yunnan Province, China.

*Corresponding author(s). E-mail(s): lankun.guo@gmail.com;
Contributing authors: jianghuiyi@hunnu.edu.cn; jhlixjtu@163.com;
zhqyuan@hunnu.edu.cn;

Abstract

In most existing knowledge tracing models, students' knowledge state is assumed to constantly change with the variation of individual exercises and even allowed for significant fluctuation. However, students usually work on exercises in batches within short time windows, separated by relatively longer intervals during which forgetting and latent learning may occur. Such special application scenarios require optimized knowledge tracing models to support the tracing of students' knowledge state with characteristics of stability in short time range and variability across different time stages. In this paper, a knowledge tracing model is developed based on knowledge-context-aware augmentation of Q-matrices for the tracing of students' intermittent knowledge states. Particularly, an approach of topologically sorting knowledge concepts is proposed by exploiting the complete paths contained in the knowledge concept graph. Then, a method of augmenting original Q-matrices is developed by requiring the updated Q-matrices to have best compatibility with the topological sorting on knowledge concepts. Finally, based on the augmentation of Q-matrices, a knowledge tracing model is developed by exploring Bayesian formulas and item response theory, which is specifically designed to trace students' intermittent mastery on knowledge concepts. Data

experiments are conducted on three public datasets and two newly collected datasets in order to verify the efficiency of the proposed model.

Keywords: Knowledge tracing, Q-matrix, Item response theory, Knowledge concept

1 Introduction

With the support of machine learning and artificial intelligence technologies, intelligent education achieves more automated evaluation of students' learning process, which provides teachers with scientific and objective evaluation criteria, thereby improving the quality of teaching and learning (Ouyang and Jiao 2021; Tapalova and Zhiyenbayeva 2022). Many intelligent education systems such as Massive Online Open Course, Knewton.com and KhanAcedemy.org have been developed, in which learners can obtain more personalized learning support and their learning behaviors possess stronger adaptive characteristics (Kerimbayev et al. 2025). Specifically, measuring and evaluating students' dynamic knowledge state is one of the core tasks of intelligent education systems.

In the field of intelligent education, knowledge tracing predicts learners' dynamic knowledge mastery status by analyzing their learning behavioral data, and thereby provides support for targeted teaching and personalized learning (Liu et al. 2021; Abdelrahman et al. 2023). The initial research on knowledge tracing can be traced back to Corbett and Anderson (1994) who used Bayesian networks to model students' learning process. The subsequent research can be summarized into several categories including: Bayesian models based on probability theory (De La Torre 2009; Pelánek 2017; Käser et al. 2017), logistic models based on logistic functions (Cen et al. 2006; Pavlik Jr et al. 2009), and deep learning models based on neural networks (Liu et al. 2025). Typically, learners, exercises or tests, and knowledge concepts are essential entities in constructing knowledge tracing models. As the main training data, learners' behavioral performance on learning resources, such as students' scoring rate on exercises or tests, can be collected in the actual teaching process. Moreover, the cognitive attributes of learners, such as forgetting rate and learning rate, are often considered as necessary model parameters. Furthermore, the connection between exercises (tests) and knowledge concepts, as well as the correlation between knowledge concepts, are important for analyzing learners' mastery on knowledge concepts. The difficulty and discrimination of exercises (tests) are also important parameters in knowledge tracing models.

The performance data of learners is generally presented in the form of time series. In processing such data, most knowledge tracing models assume that students' mastery on knowledge concepts is constantly changing, and even allow for significant fluctuation on knowledge states. However, learning is a gradual process in which students' mastery on knowledge concepts generally does not fluctuate at random. Moreover, students' mastery on knowledge concepts usually has the characteristic of intermittent fluctuation. For example, in an independent exam, it can be considered that students' mastery on knowledge concepts is basically stable and unchanged. Meanwhile, for

multiple exams with relatively larger time intervals, students' mastery on knowledge concepts is allowed to fluctuate slightly. These cognitive features are rarely considered in existing deep learning based knowledge tracing models.

We give a toy example to explain the phenomenon of intermittent changes on learners' knowledge state. As shown in Fig. 1, it is assumed that a student completed a learning process by participating several tests $\{\tau_1, \tau_2, \tau_3\}$ in sequence. In each test τ_i, several exercises $\{e_j^{(i)}\}$ are contained respectively and the student's actual scoring on each exercise is recorded. About the connection between exercises and knowledge concepts, we assume that an exercise may be related to several knowledge concepts. Moreover, a graph which demonstrates the interconnection among knowledge concepts is provided. Based on the historical performance data and information data on exercises or knowledge concepts, the dynamic knowledge state, i.e., the change of the student's mastery on knowledge concepts, is expected to be captured.

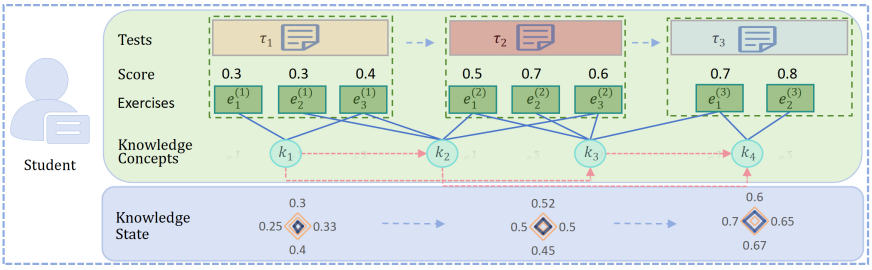


Fig. 1 A toy example of tracing intermittent knowledge states

Besides the tracing of intermittent knowledge states, more challenges exist on the accuracy and interpretability of knowledge tracing tasks. On the one hand, the quality of Q-matrix greatly affects the efficiency of cognitive diagnosis and knowledge tracing models. Most probability based knowledge tracing models regard a Q-matrix as an invariant supportive information for the purpose of knowledge tracing. Particularly, in many deep learning based knowledge tracing models, a Q-matrix is often regarded as a graph or hyper-graph based on which information fusion is conducted. However, the implicit connection between exercises and knowledge concepts are usually lost. Such incompleteness of Q-matrices inevitably reduces the accuracy of knowledge tracing. On the other hand, deep learning based knowledge tracing models can leverage the flexibility and powerful computing power of neural networks. The cooperation of convolutional neural networks or attention mechanisms can enhance the fusion ability of deep learning in knowledge tracing tasks, thereby improving the prediction accuracy. However, the inherent "black box" attribute of deep learning leads to the lack of sufficient interpretability in capturing knowledge states, thereby affecting the credibility and transparency of the models.

This paper proposes a knowledge tracing model for students' intermittent knowledge state based on knowledge-context-aware augmentation of Q-matrices. Firstly, by exploiting the complete paths contained in the knowledge concept graph, an approach of mining topological index on knowledge concepts is developed. Secondly,

a novel approach of augmenting the original Q-matrices is developed by requiring the updated Q-matrices have best fitness with the topological index on knowledge concepts. Thirdly, as the main module of the proposed model, an intermittently knowledge tracing model based on the knowledge-context-aware augmentation of Q-matrices is developed. The distinctive feature of the proposed model is admitting knowledge states' characteristics of stability in short time range and variability across different stages. Besides students' mastery on knowledge concepts, the proposed model also takes into account students' cognitive characteristics and objective attributes of exercises, which enhances the model's interpretability and rationality.

2 Related Work

2.1 Implicit Knowledge Mining

Q-matrices play an important role in constructing cognitive diagnosis or knowledge tracing models. In particular, a Q-matrix depicts the connection between learning resources (such as tests or exercises) and knowledge concepts. As a necessary component of cognitive diagnosis models, a Q-matrix often serves as a bridge between learners' mastery on knowledge concepts and their proficiency on learning resources (Liu et al. 2018). In the deep learning based knowledge tracing models, a Q-matrix is often regarded as a graph or hyper-graph based on which information fusion is conducted (Nakagawa et al. 2019; Ye and Shan 2023; Xu et al. 2025). In addition, some works have been conducted on revising Q-matrices in the form of a component embedded into a complex network architecture (Yang et al. 2022; Cui et al. 2024; Huo et al. 2025) or probability-based models (Chen et al. 2017).

Because Q-matrices are often provided by educational experts, only those knowledge concepts explicitly related to exercises are identified. However, there usually exist implicit knowledge concepts related to exercises due to the complex interrelation between knowledge concepts. From the prediction point of view, learners' performance on exercises depends on not only their mastery on the explicit knowledge concepts, but also that on those implicit ones. Moreover, since there exists difference on the degree of correlation between learning resources and knowledge concepts, a Q-matrix should be allowed to be decimal, i.e., such correlation is better to be diversified and hierarchically organized. Even though techniques such as graph neural networks can benefit the capturing of the potential connection between exercises and implicit knowledge concepts, the module of revising Q-matrices therein lose more or less interpretability. Therefore, methods of mining implicit knowledge concepts are required to improve the efficiency of cognitive diagnosis or knowledge tracing models.

2.2 Cognitive Diagnosis

Cognitive diagnosis aims to analyze learners' knowledge state based on their observed performance in the learning process (Anderson 1986; Leighton and Gierl 2007; Guo et al. 2024). In this respect, item response theory (IRT) provides an approach of diagnosing learners' latent characteristics including the mastery on knowledge concepts and entities' attributes such as differentiation and difficulty (Harris 1989; Cai et al.

2016; Khan and Ghosh 2021). It also provides a straightforward approach of modeling the relationships between latent variables and explanatory variables (Zwinderman 1991; Christensen 2006). The well-known IRT models include Lord's two-parameter normal ogive model (Lord 2012), the Rasch model (Rasch 1980) and Birnbaum's three-parameter logistic model (Birnbaum 1968). Recently, many extended IRT-based models have been developed by incorporating more cognitive variables or designing new strategy of simulating learners' performance (Reckase 1997; Bradshaw and Templin 2014; Zhan et al. 2018). IRT-based mechanism or ideas have also been embedded into neural network architectures, which improves the accuracy of predicting learners' performance (Yeung 2019; Tsutsumi et al. 2023; Guo et al. 2025).

One of the advantages of cognitive diagnostic models is the good interpretability of estimated parameters. The mastery on knowledge concepts of students, the discrimination and difficulty of exercises, as well as the guessing or slip rates of students on exercises, are common parameters admitted in cognitive diagnostic models. Particularly, the accuracy and completeness of Q-matrices greatly affect the efficiency of cognitive diagnostic models. However, the construction of a Q-matrix usually requires the support of a large amount of high-quality data. If the completeness and accuracy of the Q-matrix are not enough, the models' performance usually significantly decrease. Moreover, forgetting and learning factors are rarely considered in classical cognitive diagnostic models, which reduces their capability of capturing rational knowledge state changes. Furthermore, most cognitive diagnostic models are based on strong assumptions, such as the Markov features of knowledge state changes and linear relationship between cognitive parameters, which leads to poor applicability of the models in dealing with complex problems or nonlinear relationship scenarios. In addition, cognitive diagnostic models mostly rely on the binary characteristics on observed performance of learners. When the actual data does not conform to the assumptions, the classification error of the model may significantly increase, especially in multidimensional capability assessment tasks. Therefore, it is necessary to construct cognitive diagnostic models which accommodate revised Q-matrices and more cognitive parameters.

2.3 Knowledge Tracing

The mainstream knowledge tracing models include Bayesian network based models and deep learning based models. On the one hand, knowledge tracing models based on Bayesian networks typically utilize prior probabilities and consider interpretable parameters related to the attributes of learners and exercises (Delen et al. 2020; Bulut et al. 2023). The advantage of such type of models lies in the strong interpretability, but they also have limitations such as ignoring forgetting effect and differences in students' learning rates. On the other hand, knowledge tracing models based on LSTM or GRU can capture complex temporal features by modeling dynamic changes in knowledge states through recurrent neural networks (Liu et al. 2023; Wu and Ling 2023; Ning et al. 2025). Specifically, by introducing attention mechanisms or leveraging the advantages of graph neural networks, the knowledge concepts association modeling ability of deep learning based knowledge tracing models can often be significantly improved (Pandey and Srivastava 2020; Liu et al. 2019; Abdelrahman and Wang 2022; Wang et al. 2023; Park et al. 2024). In addition, in order to realize a stable assessment

of students' knowledge state, enhancement of knowledge tracing models was studied from the perspective of causal reference (Li et al. 2023; Zhu et al. 2023; Huang et al. 2024; Lu et al. 2025).

Although deep learning based knowledge tracing models can capture the connection among complex knowledge concepts, both forgetting and learning factors are often embedded into a specific neural network structure. The most typical methods are the forgetting gates mechanism and output activation in LSTM. On the one hand, sufficient training data is usually demanded for deep learning and the interpretability of model parameters is weak. Although changes in knowledge states, as intermediate representation states, are generally attempted to be captured in existing models, it is usually difficult to simulate students' mastery on specific knowledge concepts. On the other hand, in processing time series data, such models generally assume that students' mastery on knowledge concepts is constantly changing, and even allow for significant fluctuation on knowledge states. However, from the practical perspective, students' mastery on knowledge concepts generally has the characteristic of intermittent fluctuation. For example, in an independent exam, it can be assumed that students' mastery on knowledge concepts is basically stable and unchanged. Meanwhile, for multiple exams with relatively larger time intervals, students' mastery on knowledge concepts is allowed to fluctuate slightly. These cognitive features are rarely considered in existing deep learning based knowledge tracing models.

3 Problem Formulation and Model Overview

3.1 Problem Formulation

Suppose that there are a set S of students, a set $\boldsymbol{\tau} = \{\tau_i\}_{i=1,2,\cdots,N}$ of integrative competencies tests and a set K of knowledge concepts. Particularly, for each test τ_i, a set E_i of exercises is contained. It is assumed that an exercise does not appear repeatedly across all the tests. From practical viewpoint, the time series of tests $\{\tau_i\}$ facilitate students' gradual mastery of knowledge concepts and also provide an evaluation basis for students' learning efficiency at different stages. From the perspective of knowledge tracing, the performance of students on the test at the end of each stage provides data basis for measuring students' dynamic knowledge states.

Assume that the tests are conducted successively on students in S at specific points in time. The real scoring rates of each student on each exercise are collected. For convenience, the notation $R_{se}^{(i)}$ is used to represent the real scoring rate of s on an exercise $e \in E_i$. During the time window between two adjacent tests, students have autonomy on learning strategies and their learning possesses more personalized features, which makes the corresponding data relatively incomplete. For instance, the data about $R_{se}^{(i)}$ may be inaccurate or has not even been collected. For convenience, the notation $\Delta t^{(i)}$ is used to denote the interval length of the time window from τ_i to τ_{i+1}. The notation $[i, i+1]$ is used to denote such time window for convenience. In order to deal with the associated knowledge tracing problems, especially on the mastery of knowledge concepts, methods or models are required to admit characteristics of stability in short time range and variability across different stages.

The notion of Q-matrix provides a bridge between students' mastery on knowledge concepts and their performance on learning resources. In practice, Q-matrices are usually provided by experts having years of teaching experience. Particularly, for each test τ_i, the connection between exercises and knowledge concepts is represented by a matrix $\mathbf{Q}^{(i)} = \{q_{ek}^{(i)} \mid e \in E_i \ \& \ k \in K\}$. In particular, $q_{ek}^{(i)} = 1$ means that knowledge concept k is explicitly related to exercise $e \in E_i$, and vice versa. For convenience, each Q-matrix $\mathbf{Q}^{(i)}$ can be also represented by $\{\mathbf{Q}_e^{(i)}\}_{e \in E_i}$ where the $|K|$-dimensional vector $\mathbf{Q}_e^{(i)}$ just serves as the one-hot representation vector of exercise e describing which knowledge concepts are explictly related to exercise $e \in E_i$.

According to the theory of cognitive structure transfer, all meaningful learning is always based on the original cognitive structure. In other words, the learning on new-coming knowledge concepts is based on the already-learned ones. This is in line with an old China's teaching philosophy of "reviewing the past and learning new things". Therefore, we consider the "is prior to" relationship between knowledge concepts. Moreover, we allow such relationship to be multi-valued in order to describe the different strength of such relationship. Formally, such preparatory relationship can be represented by a mapping $\mathbb{T} : K \times K \longrightarrow [0, 1]$. For knowledge concepts k and k', the value $\mathbb{T}(k, k')$ describes the transfer degree of k being a preparatory knowledge concept of k'. In addition, a knowledge concept graph $\mathbf{G}$ can be obtained if an arrow from k to k' is drawn if $\mathbb{T}(k, k') \neq 0$.

Based on the notions mentioned above, the problem can be formulated as follows: given students' performance $\mathbf{R} = \{R_{se}^{(i)}\}$ on individual exercises, Q-matrices $\mathbf{Q}^{(i)}$ for each test τ_i and a preparatory relationship $\mathbb{T}$ on knowledge concepts in K, the goal of the paper is to develop a knowledge tracing model for the purpose of capturing students' dynamic knowledge states, especially the mastery on knowledge concepts.

3.2 Model Overview and Main Contributions

The main contribution of the paper is to develop a knowledge tracing model based on the mining of topological index on knowledge concepts and augmentation of Q-matrices. As shown in Fig. 2, the model has a three-module framework.

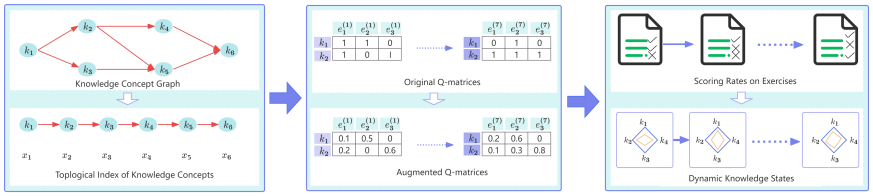


Fig. 2 Three modules of the proposed knowledge tracing model

(1) In the first module, the notion of topological index is introduced in order to evaluate the relative position degree of knowledge concept in the desired linear order on knowledge concepts. An approach of mining such topological index is developed by exploiting complete paths contained in the knowledge concept graph.

(2) In the second module, a method of evaluating the correlation between exercises and all related knowledge concepts is proposed. It can be seen as a novel approach of augmenting original Q-matrices by requiring the updated Q-matrices have best fitness with the topological index on knowledge concepts.

(3) In the module of knowledge tracing, a method of tracing students' knowledge states is proposed by admitting characteristics of stability in short time range and variability across different stages. Besides students' mastery on knowledge concepts, the proposed model also takes into account students' cognitive characteristics (such as forgetting rate and learning rate) and objective attributes of exercises (such as discrimination degree and difficulty).

4 Q-matrix Augmentation based on Topological Index of Knowledge Concepts

High quality feature expression of exercises can effectively improve the effectiveness of knowledge tracing. For the task of knowledge tracing, the related knowledge concepts are often used to characterize exercises. Particularly, the correlation between exercises and knowledge concepts is generally expressed by the notion of Q-matrix. However, Q-matrices are generally sparse and incomplete because exercises are often labeled by explicit knowledge concepts without those implicit ones. In order to improve the feature representation of exercises and the efficiency of knowledge tracing models, it is valuable to develop methods for identifying exercises' implicit knowledge concepts and quantifying the correlation between exercises and associated knowledge concepts.

The interconnection among knowledge concepts provides supplementary information for capturing more relevant knowledge concepts of exercises besides the explicit ones. Moreover, mining implicit knowledge concepts of exercises is also an inevitable requirement for knowledge tracking tasks. We use a simple example to explain this point. Suppose that k_1 is a preparatory knowledge concept of k_2. During the learning process of a student s on k_2, it will be inevitable to relearn or recall k_1. In the other words, the subsequent learning on knowledge concept k_2 activates the consolidation learning on the preparatory knowledge concept k_1. By answering specific exercises, students' mastery on relevant knowledge concepts can be enhanced. The performance of students on exercises, such as scoring rates, can also be used to evaluate their mastery on relevant knowledge concepts. Suppose that student s completes an exercise e_1 for the learning of k_1, and in a period of time another exercise e_2 is completed for the learning of k_2. On the one hand, the performance of s on e_1 (respectively, e_2) can be used to evaluate the mastery of k_1 (respectively, k_2). On the other hand, the performance of s on e_2 also provides useful information for evaluating the mastery on k_1. In order to better capture the dynamic changes of the mastery of s on k_1, it is necessary to capture the implicit correlation between e_2 and k_1. From the illustrative example, it can be seen that in order to capture the dynamic changes of students on individual knowledge concepts, it is necessary to capture the implicit connection between the exercises and knowledge concepts.

To improve the readability of symbols and formulas, the notations and the corresponding explanations are summarized in Table 1.

Table 1 Notations with descriptions for Section 4

Notation	Description
$\mathbf{G}$	a knowledge concept graph
$\mathbb{P}$	the set of all complete knowledge concept paths in $\mathbf{G}$
$\mathrm{p} \in \mathbb{P}$	a complete knowledge concept path
$k, k' \in \mathrm{p}$	two nodes of path p
$k \preceq k'$	k appears earlier than k' in p or $k = k'$
l_{p}	the length of path p
$\mathrm{p}^{(j)}$	the j-th node of path p
$x_{\mathrm{p}^{(j)}}$	the topological index of knowledge concept $\mathrm{p}^{(j)}$
$y_{\mathrm{p}^{(j)}}$	the location degree of knowledge concept $\mathrm{p}^{(j)}$ in path p
$\psi(\mathrm{p})$	positive correlation function value of path p
$\phi(\mathrm{p})$	penalty function value of path p
$\mathrm{I}(k \rtimes k')$	indicator function: $\mathrm{I}(k \rtimes k') = 1$ if k is a prerequisite of k' but $x_k > x_{k'}$
$\overleftarrow{K}_j$	the set of knowledge concepts that are earlier than $\mathrm{p}^{(j)}$ in p with larger index
$\overrightarrow{K}_j$	the set of knowledge concepts that are later than $\mathrm{p}^{(j)}$ in p with smaller index

4.1 Topological Index Mining of Knowledge Concepts

Specific sequences of knowledge concepts are often needed in the real teaching and learning. For the representation of knowledge concepts through neural networks, a linear order on knowledge concepts may be required (Lu et al. 2022). On the one hand, knowledge concepts with the connection between them provide an environment in which different learning strategies are formed. A basic principle of arranging rational knowledge concept sequences is that the order on knowledge concepts should coordinate, as much as possible, with the connection between knowledge concepts. On the other hand, graph-style structures derived from connection between knowledge concepts provide a reference to evaluate the rationality of knowledge concept sequences. In particular, the order-related information contained in such graphs can be exploited to measure the adaptability or fitness of knowledge concept sequences.

In this subsection, an approach of mining topological index of knowledge concepts based on the preparatory relationship among knowledge concepts is proposed. In this end, the notation $\mathbb{P}$ is used to represent the set of all complete knowledge concept paths in $\mathbf{G}$. Here, a path p is said to be complete if: (i) there exists no preparatory knowledge concept for the first node of p; (ii) the last node of p is not preparatory to any others; (iii) any node cannot be inserted between any adjacent nodes in p. For any path $\mathrm{p} \in \mathbb{P}$, we use $k \in \mathrm{p}$ to mean that k is a node of p. For any $k, k' \in \mathrm{p}$, we use $k \preceq k'$ to describe that k appears earlier than k' in p or $k = k'$. In the following, the appearance of $k \preceq k'$ always implies that k and k' are contained in a path p. The notation l_{p} is used to describe the length of p, i.e., the number of knowledge concepts contained in p. In addition, the notation $\mathrm{p}^{(j)}$ represents the jth node in the path p.

In order to evaluate the relative position of knowledge concepts in a desired linear order, a variable x_k is assigned to each $k \in K$. For any path $\mathrm{p} \in \mathbb{P}$ and $j \in \{1, 2, \cdots, l_{\mathrm{p}}\}$, $x_{\mathrm{p}^{(j)}}$, intuitively, serves as a topological index which evaluates the relative position degree of knowledge concept $\mathrm{p}^{(j)}$ in the desired linear order on knowledge concepts. For better fitness between the linear order on knowledge concepts and preparatory relationship between knowledge concepts, the value of $x_{\mathrm{p}^{(j')}}$ is expected to be bigger than $x_{\mathrm{p}^{(j)}}$ if $j'>j$ in that the position $\mathrm{p}^{(j')}$ is more rearward than $\mathrm{p}^{(j)}$ in the path p. Generally, it is expected that the bigger $x_{\mathrm{p}^{(j)}}$, the further rearward the position of $\mathrm{p}^{(j)}$ in the path p.

The paths derived from the knowledge concept graph contain order-related information with respect to the preparatory relationship. For any path $\mathrm{p} \in \mathbb{P}$ and $j \in \{1, 2, \cdots, l_{\mathrm{p}} - 1\}$, the notation $y_{\mathrm{p}^{(j)}}$ is used to represent the accumulated product from current transfer degree to the end one with respect to the path p. Formally,

$$y_{\mathrm{p}^{(j)}} = \prod_{i=j}^{l_{\mathrm{p}}-1} \mathbb{T}(\mathrm{p}^{(i)}, \mathrm{p}^{(i+1)}). \tag{1}$$

In the special case when $j = l_{\mathrm{p}}$, we set $y_{\mathrm{p}^{(j)}} = 1$. For any $j, j' \in \{1, 2, \cdots, l_{\mathrm{p}}\}$, if $j < j'$, i.e., the knowledge concept $\mathrm{p}^{(j')}$ is more backward than $\mathrm{p}^{(j)}$ in the path p, then the value $y_{\mathrm{p}^{(j')}}$ is bigger than or equal to $y_{\mathrm{p}^{(j)}}$, and vise versa. Intuitively, the value of $y_{\mathrm{p}^{(j)}}$ describes the accumulative preparatory degree of $\mathrm{p}^{(j)}$ to the terminal node with respect to the path p. The value of $y_{\mathrm{p}^{(j)}}$ just describes the location degree of the knowledge concept $\mathrm{p}^{(j)}$ in the path p.

The fixed arrangement of order on knowledge concepts in a path p allows introduce vectors $\mathbf{X}_{\mathrm{p}} = \{x_{\mathrm{p}^{(1)}}, x_{\mathrm{p}^{(2)}}, \cdots, x_{\mathrm{p}^{(l_{\mathrm{p}})}}\}$ and $\mathbf{Y}_{\mathrm{p}} = (y_{\mathrm{p}^{(1)}}, y_{\mathrm{p}^{(2)}}, \cdots, y_{\mathrm{p}^{(l_{\mathrm{p}})}})$. Taking an advantage of cosine similarity between $\mathbf{X}_{\mathrm{p}}$ and $\mathbf{Y}_{\mathrm{p}}$, a function

$$\psi(\mathrm{p}) = \cos(\mathbf{X}_{\mathrm{p}}, \mathbf{Y}_{\mathrm{p}}) \tag{2}$$

is defined to quantify the positive correlation between the accumulative preparatory degree and the position indicators of knowledge concepts.

Moreover, a penalty mechanism is exploited in order to reduce the contradiction between the values of the topological index and the preparatory relationship among knowledge concepts. Particularly, for any path p, the following notion is introduced:

$$\phi(\mathrm{p}) = \frac{1}{1 + e^{-\beta \cdot \chi(\mathrm{p})}} \tag{3}$$

where β is a hyper-parameter and $\chi(\mathrm{p})$ is $\sum_{k,k' \in \mathrm{p}} \mathrm{I}(k \rtimes k')(x_{k'} - x_k)$ for which $\mathrm{I}(k \rtimes k')$ is an indicator function defined as follows: $\mathrm{I}(k \rtimes k') = 1$ if k is a preparatory to k' but $x_k > x_{k'}$; otherwise, $\mathrm{I}(k \rtimes k') = 0$.

By the Bayesian formulation, it is easy to verify that

$$P(\mathbf{X}|\mathbf{G}) \propto P(\mathbf{G}|\mathbf{X})P(\mathbf{X}). \tag{4}$$

Adopting that the distribution of $\mathbf{G}|\mathbf{X}$ is independent of each other across different complete paths, we define the unnormalized probability $P(\mathbf{G}|\mathbf{X})$ as follows:

$$P(\mathbf{G}|\mathbf{X}) = \prod_{\mathrm{p}\in\mathbb{P}} [(1-\omega)\cdot\psi(\mathrm{p}) + \omega\cdot\phi(\mathrm{p})] \tag{5}$$

where $\omega \in [0,1]$ is a hyper-parameter used to adjust the proportion of $\psi(\mathrm{p})$ and $\phi(\mathrm{p})$. Intuitively, the value of $P(\mathbf{G}|\mathbf{X})$ describes the fitness between $\mathbf{X}$ and the preparatory relationship among knowledge concepts. In addition, as a prior assumption, the values of components of $\mathbf{X}$ follow a normal distribution, i.e.,

$$P(\mathbf{X}) = \prod_{k\in K} \mathcal{N}(x_k \mid \mu_x, \sigma_x). \tag{6}$$

Then, the task of assigning topological index to knowledge concepts for a linear order is transformed as solving the optimization problem of the following function:

$$F(\mathbf{X}) = \sum_{\mathrm{p}\in\mathbb{P}} \log\left[(1-\omega)\cdot\psi(\mathrm{p}) + \omega\cdot\phi(\mathrm{p})\right] - \sum_{k\in K} \frac{(x_k-\mu_x)^2}{2\sigma_x^2}. \tag{7}$$

Specifically, the goal is reset as searching for appropriate values of $\{x_k\}_{k\in K}$ which make the function F reach its maximum.

To solve the corresponding optimization problem, gradient descent method is applied. In this end, for any variable x_k, the partial derivative is calculated:

$$\frac{\partial F(\mathbf{X})}{\partial x_k} = \sum_{\mathrm{p}\in\mathbb{P}} \frac{(1-\omega)\cdot\partial\psi(\mathrm{p})/\partial x_k + \omega\cdot\partial\phi(\mathrm{p})/\partial x_k}{(1-\omega)\cdot\psi(\mathrm{p}) + \omega\cdot\phi(\mathrm{p})} - \frac{x_k-\mu_x}{\sigma_x^2}. \tag{8}$$

Particularly, if a knowledge concept k coincides with $\mathrm{p}^{(j)}$ in p, then

$$\begin{cases} \dfrac{\partial\psi(\mathrm{p})}{\partial x_k} = \dfrac{y_{\mathrm{p}^{(j)}}|\mathbf{X}_\mathrm{p}|^2 - x_{\mathrm{p}^{(j)}}(\mathbf{X}_\mathrm{p}\cdot\mathbf{Y}_\mathrm{p})}{|\mathbf{X}_\mathrm{p}|^3\times|\mathbf{Y}_\mathrm{p}|} \\ \dfrac{\partial\phi(\mathrm{p})}{\partial x_k} = \beta\cdot\phi(\mathrm{p})(\phi(\mathrm{p})-1)(|\overleftarrow{K}_j| - |\overrightarrow{K}_j|)) \end{cases} \tag{9}$$

where $\overleftarrow{K}_j$ is the set of knowledge concepts of which the position in p is more forward than $\mathrm{p}^{(j)}$ and the index is larger than $x_{\mathrm{p}^{(j)}}$. That is, $\overleftarrow{K}_j = \{k \in \mathrm{p} \mid k \preceq \mathrm{p}^{(j)} \ \&\ x_k > x_{\mathrm{p}^{(j)}}\}$. $\overrightarrow{K}_j$ is the set of knowledge concepts of which the position in p is more backward than $\mathrm{p}^{(j)}$ and the index is smaller than $x_{\mathrm{p}^{(j)}}$. That is, $\overrightarrow{K}_j = \{k \in \mathrm{p} \mid \mathrm{p}^{(j)} \preceq k \ \&\ x_k < x_{\mathrm{p}^{(j)}}\}$.

4.2 Knowledge-context-aware Q-matrix Augmentation

In this subsection, we propose an approach of augmenting original Q-matrices by exploiting the knowledge concept graph. On the one hand, the preparatory relationship

between knowledge concepts is took advantage of as a context for the mining of implicit knowledge concepts. On the other hand, the topological index of knowledge concepts is exploited to quantify the correlation between exercises and related knowledge concepts.

Without loss of generality, suppose that E is a set of exercises. For any exercise $e \in E$, the notation K_e is used to denote the set of all explicit knowledge concepts of e. The notation $\widetilde{K}_e$ is used to denote the set of knowledge concepts which appear in the paths with explicit knowledge concepts of e being afterward nodes, i.e.,

$$\widetilde{K}_e = \{k \in K \mid (\exists \mathrm{p} \in \mathbb{P}, k' \in K_e)(k \preceq k')\}.$$

Obviously, $\widetilde{K}_e$ can be obtained based on the knowledge concept graph. In order to quantify the correlation between exercises and all related knowledge concepts (explicit or implicit), a matrix $\widetilde{\mathbf{Q}} \in R^{|E|\times|K|}$ is introduced where variable $\tilde{q}_{ek}$ is assigned for each pair of $e \in E$ and $k \in \widetilde{K}_e$. We hope to obtain the most likely distribution of correlation between exercises and knowledge concepts with the preparatory relationship on knowledge concepts as the precondition.

From a probabilistic perspective, the values of variables contained in $\widetilde{\mathbf{Q}}$ maximizing the probability $P(\widetilde{\mathbf{Q}}|\mathbf{X})$ become the goal of the optimization problem. Following the Bayesian formulation, it holds that

$$P(\widetilde{\mathbf{Q}}|\mathbf{X}) \propto P(\mathbf{X}|\widetilde{\mathbf{Q}}) \times P(\widetilde{\mathbf{Q}}). \tag{10}$$

Particularly, for the conditional event $\mathbf{X}|\widetilde{\mathbf{Q}}$, assuming that the probability is unnormalized and the distribution is independent of each other across different exercises, a calculation method of $P(\mathbf{X}|\widetilde{\mathbf{Q}})$ is defined as follows:

$$P(\mathbf{X}|\widetilde{\mathbf{Q}}) = \prod_{e\in E} \sigma(\cos(\mathbf{X}_e, \widetilde{\mathbf{Q}}_e)) \tag{11}$$

where $\sigma(\cdot)$ is the well-known Sigmoid function. Intuitively, the value of $\cos(\mathbf{X}_e, \widetilde{\mathbf{Q}}_e)$ quantifies the coordination level between the augmented Q-matrix values of e and the knowledge concept graph. In addition, as a prior assumption, the values of components of $\widetilde{\mathbf{Q}}$ follow a normal distribution, i.e.,

$$P(\widetilde{\mathbf{Q}}) = \prod_{e\in E} \prod_{k\in\widetilde{K}_e} \mathcal{N}(\tilde{q}_{ek} \mid \mu_q, \sigma_q). \tag{12}$$

For convenience of calculation, a logarithmic processing is made on $P(\widetilde{\mathbf{Q}}|\mathbf{X})$. Thus, the problem of augmenting Q-matrices is transformed into an optimization problem of finding the maximum of the following function:

$$\Phi(\widetilde{\mathbf{Q}}) = \sum_{e\in E} \log(\sigma(\cos(\mathbf{X}_e, \widetilde{\mathbf{Q}}_e))) + \sum_{e\in E} \sum_{k\in\widetilde{K}_e} \log(\mathcal{N}(\tilde{q}_{ek} \mid \mu_q, \sigma_q)). \tag{13}$$

Gradient descent method is used to solve the optimization problem. Particularly, only the partial derivatives for the pair (e, k) satisfying $k \in \widetilde{K}_e$ are needed:

$$\frac{\partial \Phi}{\partial \tilde{q}_{ek}} = \left[1 - \sigma(\cos(\mathbf{X}_e, \widetilde{\mathbf{Q}}_e))\right] \frac{\partial \cos(\mathbf{X}_e, \widetilde{\mathbf{Q}}_e)}{\partial \tilde{q}_{ek}} - \frac{\tilde{q}_{ek} - \mu_q}{\sigma_q^2}. \tag{14}$$

Algorithm 1 is developed towards the augmentation of a Q-matrix. The first half part of the algorithm is for the topological index of knowledge concepts and the second half is for the augmented Q-matrix.

Example 1 An illustrative knowledge graph is presented by the left diagram in Fig. 3. It contains ten knowledge concepts $\{k_i \mid i = 1, 2, \cdots, 10\}$. The preparatory relationship degree is shown in the form of directed edges with weights.

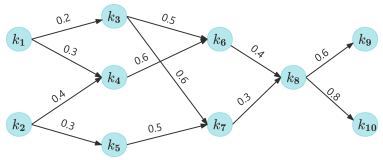


Fig. 3 An illustrative knowledge graph

The topological index values of knowledge concepts are obtained through Algorithm 1. The initial values are randomly assigned. Fig. 4 shows the trend of the topological index values of each knowledge concepts with the increase of iteration numbers, from which a good convergence can be observed. According to the final topological index values, the knowledge concepts are linearly ordered as sequence: $k_1 \to k_3 \to k_2 \to k_4 \to k_5 \to k_7 \to k_6 \to k_8 \to k_{10} \to k_9$.

Example 2 This example is a continuation of Example 1. An illustrative original Q-matrix Q is presented in the left sub-table in Table 2. It contains six exercises $\{e_i \mid i = 1, 2, \cdots, 6\}$. If $q_{e_i k_j} = 1$, i.e., knowledge concept k_j is associated with exercise e_i, then the label at row e_i and column k_j is 1. The updated Q-matrix $\widetilde{Q}$ is presented in the right sub-table in Table 2. Obviously, the correction between exercises and knowledge concepts is augmented and quantified through preparatory relationship between knowledge concepts. For instance, in the original Q-matrix, exercise e_3 is only related to knowledge concept k_7. However, in the updated Q-matrix, e_3 is related to more knowledge concepts k_1, k_2, k_3 and k_5 with redistributed correlation degrees.

5 Tracing Intermittent Knowledge State of Students

Taking problem-solving-based learning as a specific application scenario, knowledge tracing aims to analyze students' mastery on knowledge concepts based on their performance on designed exercises or questions. In order to model students' performance

Algorithm 1 Q-matrix augmentation based on topological index of knowledge concepts

Require: $\mathbf{G}$: a knowledge concept graph; $\mathbf{Q}$: an original Q-matrix; M, M': the most iteration times; h, h': step widths for gradient descent method; ϵ, ϵ': convergence hedges

Ensure: $\widetilde{\mathbf{Q}}$: an augmented Q-matrix

Obtaining the complete paths $\mathbb{P}$ in $\mathbf{G}$ and calculating y_{p} for each $\mathrm{p} \in \mathbb{P}$;
Obtaining $\widetilde{K}_e$ for each $e \in E$ through traversal searching on $\mathbb{P}$;
Initializing $\{x_k\}_{k \in K}$ which store the topological indexes of knowledge concepts;
Initializing (int) $m = 0$;
while $m < M$ and $\epsilon < d$ **do**
    Backing up temporary topological indexes $\{\hat{x}_k\}_{k \in K} \longleftarrow \{x_k\}_{k \in K}$;
    for $k \in K$ **do**
        Initializing (double) Sum=0;
        for $\mathrm{p} \in \mathbb{P}$ **do**
            if $k \in \mathrm{p}$ **then**
                Sum $\longleftarrow$ Sum $+\partial\psi/\partial x_k + \partial\phi/\partial x_k$ according to Equation (9);
            end if
        end for
        Calculating $\bigtriangledown_k \longleftarrow$ Sum$/|\mathbb{P}|$ and updating $x_k \longleftarrow \hat{x}_k + h * \bigtriangledown_k$;
    end for
    Calculating iteration distance $d = ||\hat{x} - x||$;
    $m++$;
end while
Initializing $\widetilde{\mathbf{Q}} = \{\tilde{q}_{ek} \mid k \in \widetilde{K}_e\}$ which stores the augmented Q-matrix;
Initializing (int) $m' = 0$;
while $m' < M'$ and $\epsilon' < d'$ **do**
    Backing up temporary revised Q-matrix $\widehat{\mathbf{Q}} = \{\hat{q}_{ek}\}_{e \in E, k \in K} \longleftarrow \{\tilde{q}_{ek}\}_{e \in E, k \in K}$;
    for $e \in E$ **do**
        for $k \in K$ **do**
            Calculating $\bigtriangledown_{\tilde{q}_{ek}}$ according to Equation (14) and updating $\tilde{q}_{ek} \longleftarrow \hat{q}_{ek} + h' * \bigtriangledown_{\tilde{q}_{ek}}$;
        end for
    end for
    Calculating distance d' between $\Phi(\widehat{\mathbf{Q}})$ and $\Phi(\widetilde{\mathbf{Q}})$;
    $m'++$;
end while
return $\widetilde{\mathbf{Q}}$

on exercises, besides their mastery on knowledge concepts, learners' cognitive characteristics (containing forgetting rate and learning rate) and objective attributes of learning resources (such as discrimination and difficulty) are often took into account.

To facilitate students' gradual mastery of knowledge concepts, a learning progress is often phased and thus can be decomposed into a time series of several stages. At the end of each stage, a specially arranged test or exam is often organized in order

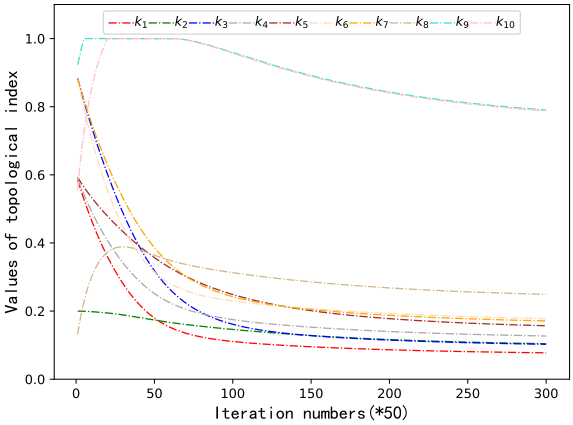


Fig. 4 The trend of the topological index values of knowledge concepts

Table 2 An illustrative original Q-matrix (left) and its updated Q-matrix (right)

	Original Q-matrix										Updated Q-matrix									
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}
e_1	0	0	0	1	0	0	0	0	0	0	0.55	0.74	0	1	0	0	0	0	0	0
e_2	0	0	0	0	0	1	0	0	0	0	0.49	0.57	0.54	0.64	0	1	0	0	0	0
e_3	0	0	0	0	0	0	1	0	0	0	0.42	0.58	0.57	0	0.71	0	1	0	0	0
e_4	0	0	0	0	0	0	0	1	0	0	0.34	0.43	0.46	0.51	0.52	0.63	0.67	1	0	0
e_5	0	0	0	0	0	0	0	0	1	0	0.22	0.24	0.25	0.27	0.34	0.30	0.31	0.40	1	0
e_6	0	0	0	0	0	0	0	0	0	1	0.24	0.29	0.31	0.26	0.30	0.37	0.36	0.46	0	1

to evaluate students' mastery on knowledge concepts. In a short period of time of an independent test or exam, it can be assumed that a student's cognitive state remains unchanged. During the time window between two adjacent tests, knowledge states' changes often have great diversities. In order to deal with such **K**nowledge **T**racing problems with **IN**termittent characteristic, especially on the mastery of knowledge concepts, we propose a knowledge tracing model **INKT** which admits characteristics of stability in short time range and variability across different stages.

To improve the readability of symbols and formulas, the notations and the corresponding explanations are summarized in Table 3. Fig. 5 is presented in order to make the proposed model more clearer.

Before a learning process in which tests $\{\tau_1, \tau_2, \cdots, \tau_N\}$ are specially organized, an initial mastery $I_{sk}^{(0)}$ is assigned for each pair of student s and knowledge concept k. From the perspective of educational psychology, students' learning ability and forgetting level are important factors that affect their knowledge state changes. Considering the personalized differences in students' learning of knowledge concepts, we use $c_{sk}^{(i)}$ to represent the learning rate of student s on knowledge concept k at the time window

Table 3 Notations with descriptions for Section 5

Notation	Description
τ_i	a test
E_i	exercises contained in τ_i
$R_s^{(i)}$	real scoring rate of student s on τ_i
$I_{sk}^{(0)}$	initial mastery of s on knowledge concept k
$f_{sk}^{(i)}$	forgetting rate of s on k in time window $[i, i+1]$
$c_{sk}^{(i)}$	learning rate of s on k at the time window $[i, i+1]$
ρ_{sk}	indirect parameter for the forgetting rate of s on k
$\gamma_e^{(i)}$	proportion of the preset score of e to the total score of τ_i
$I_{sk}^{(i)}$	mastery on k of s at the stage represented by τ_i
$\theta_{se}^{(i)}$	proficiency degree of s on exercise $e \in E_i$
$\eta_{se}^{(i)}$	simulated scoring rate of s on $e \in E_i$
$\eta_s^{(i)}$	simulated scoring rate of s on τ_i
$\lambda_e^{(i)}$	differentiation of $e \in E_i$
$d_e^{(i)}$	difficulty of $e \in E_i$

$[i, i+1]$. For forgetting factors, we assume that the forgetting rate $f_{sk}^{(i)}$ of a student s on knowledge k at the time window $[i, i+1]$ is determined by

$$f_{sk}^{(i)} = \exp(-\rho_{sk} * \Delta t^{(i)}) \tag{15}$$

where the parameter ρ_s is referred from the Ebbinghaus Forgetting Curve. Regarding the knowledge states of students at different learning stages, it is assumed that the mastery on a knowledge concept k of a student s at the stage i, denoted by $I_{sk}^{(i)}$, is determined by the mastery at the previous stage with $c_{sk}^{(i-1)}$ and $f_{sk}^{(i-1)}$ as coefficient weights. Formally, a recursive formula is defined as follows:

$$I_{sk}^{(i)} = (c_{sk}^{(i-1)} + f_{sk}^{(i-1)}) I_{sk}^{(i-1)}. \tag{16}$$

It is presumed that the latent capacity of students on exercises is determined by their mastery on related knowledge concepts. Hence, the proficiency degree of a student s on an exercise e appearing in the test τ_i is defined as follows:

$$\theta_{se}^{(i)} = \sum_{k \in K} I_{sk}^{(i)} \tilde{q}_{ek}^{(i)}. \tag{17}$$

In IRT, the difficulty and discrimination are regarded as inherent characteristics of items, which do not vary with the sample of students. In this sense, the simulated scoring rate of s on the exercise e appearing in the test τ_i is defined as follows:

$$\eta_{se}^{(i)} = \frac{1}{1 + \exp[D\lambda_e^{(i)}(\theta_{se}^{(i)} - d_e^{(i)})]} \tag{18}$$

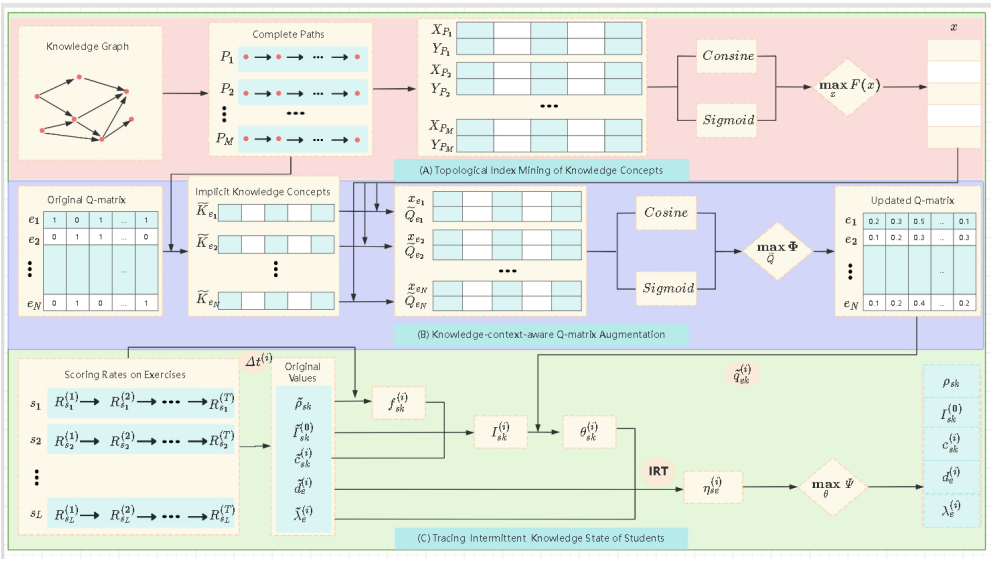


Fig. 5 The overall framework of the proposed knowledge tracing model

where $\lambda_e^{(i)}$ and $d_e^{(i)}$ are differentiation and difficulty of the exercise $e \in E_i$, respectively.

Given the data of $\mathbf{G}$, $\{\mathbf{Q}^{(i)}\}$ and $\mathbf{R}$ as input, our goal is to develop a method of learning the parameters $\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}$ and $\boldsymbol{\rho}$. First of all, each Q-matrix $\mathbf{Q}^{(i)}$ can be updated as $\widetilde{\mathbf{Q}}^{(i)}$. Then, assuming that the parameters to learn are independent of each other, the following equation is obtained by the Bayesian formula (see the proof in Appendix A.1):

$$P(\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho} | \widetilde{\mathbf{Q}}, \mathbf{R}) \propto P(\mathbf{R} | \mathbf{I}^{(\mathbf{0})}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}, \widetilde{\mathbf{Q}}) P(\mathbf{I}^{(\mathbf{0})}) P(\boldsymbol{\lambda}) P(\mathbf{d}) P(\mathbf{c}) P(\boldsymbol{\rho}). \quad (19)$$

Then, by formulating the log posterior distribution, the loss function is defined as follows:

$$\Psi = \log(P(\mathbf{R} | \mathbf{I}^{(\mathbf{0})}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}, \widetilde{\mathbf{Q}}) P(\mathbf{I}^{(\mathbf{0})}) P(\boldsymbol{\lambda}) P(\mathbf{d}) P(\mathbf{c}) P(\boldsymbol{\rho})). \quad (20)$$

Finally, the problem of knowledge tracing is transformed equivalently to an optimization problem of minimizing the function Ψ.

For student s and knowledge concept k, we use the notation $\overline{I}_{sk}^{(0)}$ to represent a preliminary assessment of students' mastery on knowledge concepts based on his/her historical data. Notations $\overline{\lambda}_e^{(i)}$, $\overline{d}_e^{(i)}$, $\overline{c}_{sk}^{(i)}$ and $\overline{\rho}_s$ are given similarly. Moreover, we assume that the parameters to learn follows special distribution as follows:

$$\eta_{se}^{(i)} \sim \mathcal{N}(\mu_{\eta_{se}^{(i)}} = R_{se}^{(i)}, \sigma^2_{\eta_{se}^{(i)}}), \quad I_{sk}^{(0)} \sim \mathrm{Be}(\alpha_{I_{sk}^{(0)}}, \beta_{I_{sk}^{(0)}}), \quad \lambda_e^{(i)} \sim \mathrm{Be}(\alpha_{\lambda_e^{(i)}}, \beta_{\lambda_e^{(i)}}),$$

$$d_e^{(i)} \sim \mathrm{Be}(\alpha_{d_e^{(i)}}, \beta_{d_e^{(i)}}), \quad c_{sk}^{(i)} \sim \mathrm{Be}(\alpha_{c_{sk}^{(i)}}, \beta_{c_{sk}^{(i)}}), \quad \rho_{sk} \sim \mathrm{Be}(\alpha_{\rho_{sk}}, \beta_{\rho_{sk}})$$

for which $\overline{I}_{sk}^{(0)}$ (respectively, $\overline{d}_e^{(i)}$, $\overline{c}_{sk}^{(i)}$, $\overline{\rho}_s$) serves as the maximum point for the distribution of $I_{sk}^{(0)}$ (respectively, $d_e^{(i)}$, $c_{sk}^{(i)}$, ρ_s); $R_s^{(i)}$ (respectively, $\overline{\lambda}_e^{(i)}$) serves as the mean for the distribution of $\eta_s^{(i)}$ (respectively, $\lambda_e^{(i)}$).

Based on these distributions, the loss function is transformed as follows:

$$\begin{aligned}\Psi = &\sum_{i=1}^{N}\sum_{s\in S}\sum_{e\in E_i}\left[-\frac{(\eta_{se}^{(i)}-R_{se}^{(i)})^2}{2\sigma^2_{\eta_{se}^{(i)}}}\right]+\sum_{s\in S}\sum_{k\in K}\left[(\alpha_{I_{sk}^{(0)}}-1)\ln(I_{sk}^{(0)})+(\beta_{I_{sk}^{(0)}}-1)\ln(1-I_{sk}^{(0)})\right]\\
&+\sum_{i=1}^{N}\sum_{e\in E_i}\left[(\alpha_{\lambda_e^{(i)}}-1)\ln(\lambda_e^{(i)})+(\beta_{\lambda_e^{(i)}}-1)\ln(1-\lambda_e^{(i)})\right]\\
&+\sum_{i=1}^{N}\sum_{e\in E_i}\left[(\alpha_{d_e^{(i)}}-1)\ln(d_e^{(i)})+(\beta_{d_e^{(i)}}-1)\ln(1-d_e^{(i)})\right]\\
&+\sum_{i=1}^{N}\sum_{s\in S}\sum_{k\in K}\left[(\alpha_{c_{sk}^{(i)}}-1)\ln(c_{sk}^{(i)})+(\beta_{c_{sk}^{(i)}}-1)\ln(1-c_{sk}^{(i)})\right]\\
&+\sum_{s\in S}\sum_{k\in K}\left[(\alpha_{\rho_{sk}}-1)\ln(\rho_{sk})+(\beta_{\rho_{sk}}-1)\ln(1-\rho_{sk})\right].\end{aligned} \tag{21}$$

The hyper-parameters of the distributions play two roles. One is to determine the degree of freedom of parameter iteration around the reference value. The other is to control the proportion of losses corresponding to each parameter in the loss function. If the training data is small-scale, the parameters $I_{sk}^{(0)}$, $\lambda_e^{(i)}$, $d_e^{(i)}$, $c_{sk}^{(i)}$ and ρ_s are updated directly by using gradient descent method. In the case of large-scale data, Stochastic or mini-batch gradient descent method may be used. Specifically, the derivative of each parameter is computed. For demonstration purpose, only the derivative for $I_{sk}^{(0)}$ is presented here and all derivatives can be seen from Appendix A.2:

$$\nabla_{I_{sk}^{(0)}}=\frac{\alpha_{I_{sk}^{(0)}}-1}{I_{sk}^{(0)}}-\frac{\beta_{I_{sk}^{(0)}}-1}{1-I_{sk}^{(0)}}+\sum_{i=1}^{N}\sum_{e\in E_i}\left[-\frac{\eta_{se}^{(i)}-R_{se}^{(i)}}{\sigma^2_{\eta_{se}^{(i)}}}(\frac{\partial\eta_{se}^{(i)}}{\partial I_{sk}^{(0)}})\right] \tag{22}$$

where

$$\frac{\partial\eta_{se}^{(i)}}{\partial I_{sk}^{(0)}}=D\lambda_e^{(i)}\tilde{q}_{ek}^{(i)}\eta_{se}^{(i)}(\eta_{se}^{(i)}-1)\prod_{j=0}^{i-1}(c_{sk}^{(j)}+f_{sk}^{(j)}). \tag{23}$$

Algorithm 2 is developed for the learning of parameters.

Because the gradient of each parameter can be computed independently based on the results of the previous iteration, the gradient components' computation is distributed into multiple threads. Moreover, an incremental update strategy is applied in cumulative multiplication and accumulation. For instance, the notation $\mathcal{E}_{sk}^{(i)}$, for convenience, is used to denote the cumulative multiplication $\prod_{j=0}^{i-1}(c_{sk}^{(j)}+f_s^{(j)})$. It can

Algorithm 2 Parameters learning of **INKT** model

Require: $\mathbf{G}$: a knowledge concept graph; $\mathbf{Q}^{(i)}$: original Q-matrices of tests τ_i; $\mathbf{R}$: scoring rates of all students on all tests; $\overline{I}_{sk}^{(0)}$, $\overline{\lambda}_e^{(i)}$, $\overline{d}_e^{(i)}$, $\overline{c}_{sk}^{(i)}$, $\overline{\rho}_s$: prior parameter values; $M^{''}$: a most iteration time; $\epsilon^{''}$: a convergence hedge

Ensure: $\mathbf{\Theta} = \{\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}\}$

Obtaining the augmented Q-matrices $\widetilde{\mathbf{Q}}^{(i)}$ based on $\mathbf{G}$ and $\mathbf{Q}^{(i)}$ by Algorithm 1;
Initializing the parameters $\widetilde{\mathbf{\Theta}}$ and the gradient $\overline{\nabla}_\Psi$;
Initializing (int) $m = 0$;
while $m < M^{''}$ and $\epsilon^{''} < d$ **do**
    Computing $\mathcal{E}_{sk}^{(i)}$ for each triplet (s, k, i) and $\eta_{se}^{(i)}$ for each triplet (s, e, i);
    for each student $s \in S$ **do**
        Computing ∇_{ρ_s} after computing $\frac{\partial \eta_{se}^{(i)}}{\partial \rho_s}$ for each pair (e, i);
        for each knowledge concept $k \in K$ **do**
            Computing $\nabla_{I_{sk}^{(0)}}$ after computing $\frac{\partial \eta_{se}^{(i)}}{\partial I_{sk}^{(0)}}$ for each pair (e, i);
            Computing $\frac{\partial \eta_{se}^{(j)}}{\partial c_{sk}^{(i)}}$ for each triplet (e, i, j) with $i < j$;
            for each test τ_i **do**
                Computing $\nabla_{c_{sk}^{(i)}}$ based on $\{\frac{\partial \eta_{se}^{(j)}}{\partial c_{sk}^{(i)}} \mid i < j, e \in E_i\}$;
            end for
        end for
    end for
    for each test τ_i **do**
        for each exercise $e \in E_i$ **do**
            Computing $\nabla_{\lambda_e^{(i)}}$ and $\nabla_{d_e^{(i)}}$ after computing $\frac{\partial \eta_{se}^{(i)}}{\partial \lambda_e^{(i)}}$ and $\frac{\partial \eta_{se}^{(i)}}{\partial d_e^{(i)}}$ for each $s \in S$;
        end for
    end for
    Obtaining the gradient ∇_Ψ by splicing all $\nabla_{I_{sk}^{(0)}}, \nabla_{\lambda_e^{(i)}}, \nabla_{d_e^{(i)}}, \nabla_{c_{sk}^{(i)}}, \nabla_{\rho_s}$;
    Updating the values of $\mathbf{\Theta}$ based on $\widetilde{\mathbf{\Theta}}$ and ∇_Ψ according to Adam optimization;
    Calculating distance d between ∇_Ψ and $\overline{\nabla}_\Psi$;
    $\overline{\nabla}_\Psi \longleftarrow \nabla_\Psi$; $\widetilde{\mathbf{\Theta}} \longleftarrow \mathbf{\Theta}$;
    $m++$;
end while
return $\mathbf{\Theta}$.

be computed in a recursive way according to the following formulation:

$$\mathcal{E}_{sk}^{(i)} = \mathcal{E}_{sk}^{(i-1)} * (c_{sk}^{(i-1)} + f_{sk}^{(i-1)}). \tag{24}$$

At the beginning of each iteration, the values of all $\mathcal{E}_{sk}^{(i)}$ and $\eta_{se}^{(i)}$ are computed in advance and then repeatedly used for derivatives of all possible parameters. Furthermore, the nesting of 'for' loops are carefully organized to avoid duplicate calculations

Table 4 A brief summary of MTest

Entity	Students	Tests	Exercises		Knowledge Concepts
			Objective	Subjective	
Number	69	7	56	77	88

as much as possible. In addition, the well-known Adam optimization algorithm may be adopted in order to obtain independent adaptive learning rates for different parameters.

6 Experiments and Results

6.1 Databases and Experimental Setups

The database MTest was collected specifically for data experiments. The data is from several simulated college entrance examinations in the senior year of one class of a senior high school in Shenzhen City. Teachers and students involved in the collection of such dataset have provided informed consent. The data does not involve personal privacy information. The entities contain 69 students, 7 tests and 88 knowledge concepts. Each test contains 19 exercises. The type, difficulty and item discrimination of tests and exercises are basically consistent with the Chinese National College Entrance Examination. The scores of each student on each test were collected. Specifically, in each test, exercises no. 1-8 are single-choice ones; exercises no. 9-11 are multiple-choice ones; exercises no. 12-14 are multiple-choice ones; exercises no. 15-19 are subjective comprehensive ones. The knowledge concepts are core knowledge concepts or skills clarified in the High School Mathematics Teaching Syllabus. They are consecutively numbered according to the order in which they appear in the textbook.The dataset[1] can be downloaded freely. A brief summary of the dataset is shown in Table 4.

The correlation between exercises and (explicit) knowledge concepts, and the knowledge concept graph were given by the authors. For convenience of conducting data experiments, we adopted the digraph corresponding to the preliminary relationship as the knowledge concept graph. The distribution of explicitly related knowledge concepts in the Q-matrices is shown in Fig. 6. Generally, more knowledge concepts are related to the subjective exercises than to the objective ones. Some objective exercises with lower difficulty are related to an individual knowledge concept. Most exercises, such as multiple-choice ones and comprehensive ones are related to multiple knowledge concepts. It results that the distribution of knowledge concepts in objective exercises are relatively more scattered than that in subjective exercises.

The augmenting of original Q-matrices can be seen as a preprocessing of data. Based on Algorithm 1, the Q-matrices of all tests were augmented. The updated distribution of knowledge concepts is demonstrated in Fig. 7. It can be seen from the figure that more implicit knowledge concepts of exercises were captured and the correlation degree was limited within $[0, 0.5]$. Moreover, the correlationship between

[1] Available freely at: https://pan.baidu.com/s/1-wvhK_42e3q3UbGd6UoOiw?pwd=868f

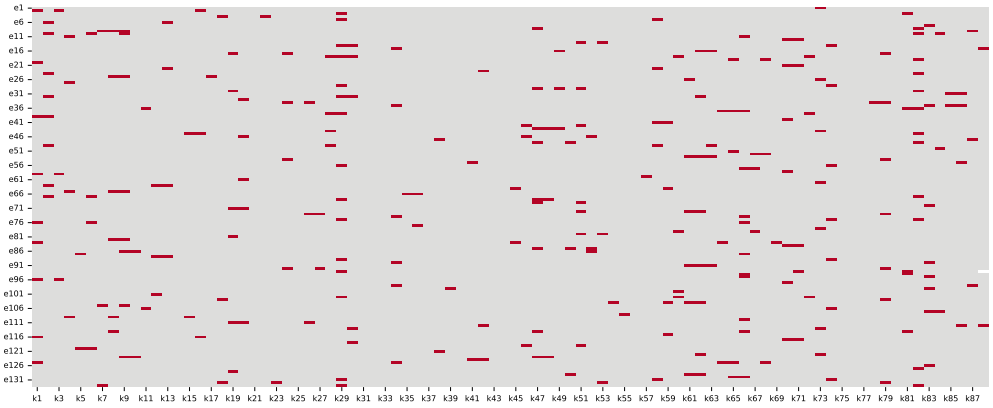


Fig. 6 Distribution of explicit knowledge concepts in original Q-matrices

exercises and knowledge concepts were quantified. The data provides a base for the experiments on knowledge tracing.

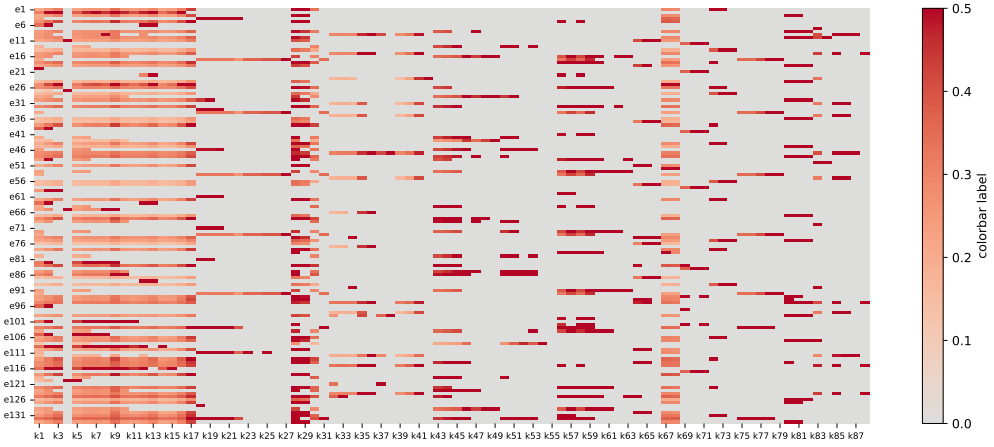


Fig. 7 Distribution of knowledge concepts in augmented Q-matrices

The dataset MTest is relatively small since only 69 students are contained. In this paper, we applied formal concept analysis (FCA) to generate a simulated data named FMTest which provides a larger scale dataset. From the viewpoint of data mining, FCA provides a method of capturing the hierarchy inherent in relational data (Wille 2009; Ganter and Wille 1999; Yao 2020). Here, we directly applied it to extend the data about scoring rate of students on exercises. Specifically, we first selected a set $\widetilde{E}$ of objective exercises. Then, a binary relation $I \subseteq S \times \widetilde{E}$ is defined as: for any $s \in S$ and $e \in \widetilde{E}$, $(s, e) \in I$ if and only if the scoring rate of s on e is 1. In terms of FCA, the triplet $(S, \widetilde{E}, I)$ serves as a (formal) context. According to FCA theory, operators

$(\cdot)^{\uparrow}$ and $(\cdot)^{\downarrow}$ are defined as follows:

$$(\cdot)^{\uparrow} : 2^{S} \to 2^{\widetilde{E}} : X \mapsto \{y \in \widetilde{E} \mid (\forall x \in X)\ (x, y) \in I\},$$
$$(\cdot)^{\downarrow} : 2^{\widetilde{E}} \to 2^{S} : Y \mapsto \{x \in S \mid (\forall y \in Y)\ (x, y) \in I\},$$

A pair (X, Y) is called a (formal) concept of $(S, \widetilde{E}, I)$ if $X^{\uparrow} = Y$ and $Y^{\downarrow} = X$. Intuitively, the extent X of a concept (X, Y) can be regarded as a group of students which have similar performance on a common set Y of exercises. Finally, for each student group X, its scoring rate on $e \in E$ can be quantified based on the scoring rates of students contained in X on the exercise e.

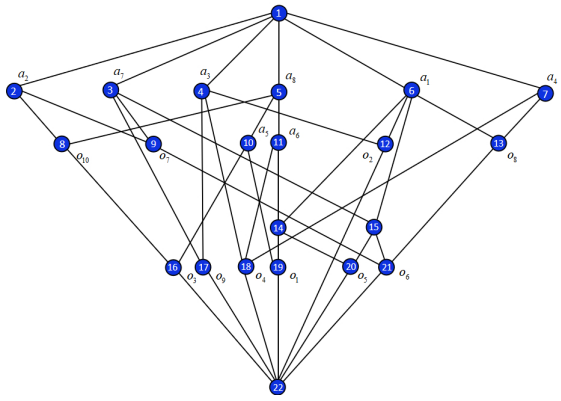


Fig. 8 The concept lattice of the illustrative context

Example 3 As an illustrative example, a context is given in Table 5. It contains 15 concepts and the corresponding concept lattice is presented in Fig. 8.

The number of concepts generally increases exponentially with the scale of the contexts. Regarding the generation of concept lattice, there exist many algorithms and tools. In this paper, we redesigned a program of the multistage algorithm for constructing concept lattices developed in (Troy et al. 2007). The central idea of the algorithm is staring from the attribute concepts (the starting stage) and repeatedly forming pairwise intersections of new generated concepts through enough stages until no new concepts are generated. The pseudo code is presented in Algorithm 3. In more detail, a container $\mathfrak{C}_{new}$ is initialized at the beginning, which stores the extents of all attribute-concepts of a context (S, E, I). The extents of all concepts are obtained through a series of while loops. In each while loop, a temporary container $\mathfrak{C}$ is used to store the results of intersecting arbitrary two elements in $\mathfrak{C}_{new}$. The container $\mathfrak{C}_{new}$ is dynamically updated, which stores the new generated concepts at each stage. The target container $\mathscr{G}$ is used to store all generated concepts. The while loop stops when $\mathfrak{C}_{new} = \emptyset$, i.e., no new concepts are generated.

Table 5 An illustrative example of a context

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
o_1	1	0	0	0	1	1	0	1
o_2	1	0	1	0	0	0	0	0
o_3	0	1	0	0	1	0	0	1
o_4	0	0	1	1	0	1	0	1
o_5	1	0	0	0	0	1	1	1
o_6	1	1	0	1	0	0	1	0
o_7	0	1	0	0	0	0	1	0
o_8	1	0	0	1	0	0	0	0
o_9	0	0	1	0	0	0	1	0
o_{10}	0	1	0	0	0	0	0	1

Algorithm 3 Generating scoring rates of student groups based on concept analysis

Require: A context (S, E, I) storing scoring rate data of students
Ensure: A context $(\mathscr{G}, E, \tilde{I})$ storing scoring rate data of (all) student groups
Initializing $\mathfrak{C}_{new}$: the extents induced by individual attributes of (S, E, I);
$\mathscr{G} \longleftarrow \mathfrak{C}_{new}$;
while $\mathfrak{C}_{new} \neq \emptyset$ **do**
    Computing $\mathfrak{C}$ by intersecting arbitrary two extents in $\mathfrak{C}_{new}$;
    $\mathfrak{C}_{new} \longleftarrow \mathfrak{C} - \mathscr{G}$;
    $\mathscr{G} \longleftarrow \mathscr{G} \bigcup \mathfrak{C}_{new}$;
end while
for each extent (student group) $X \in \mathscr{G}$ **do**
    for each exercise $e \in E$ **do**
        Computing $\tilde{I}(X, e)$ by integrating $\{I(s, e) \mid s \in X\}$;
    end for
end for
return $(\mathscr{G}, E, \tilde{I})$

We arbitrarily selected 11 objective exercises on which the scoring rates are 0 or 1. The derived concept lattice contains 1227 concepts. For each group represented by a concept and a specific exercise e, the average scoring rate on e of students contained in the group was calculated. Then, the scoring rate of a randomly selected student in the group was obtained. Subsequently, the mean of the average value and randomly selected one was taken as the scoring rate of the group on e. Moreover, on the objective exercises on which the scoring rates should be 0 or 1, a threshold was assigned in order to polarize the associated scoring rates of student groups. Furthermore, 520 groups were selected in order to generate a new dataset FMTest for data experiments. Since the scale of FMTest is much larger than MTest, it is expected to make the data experiments more comprehensive.

Three public datasets[2] FrcSub, Math1 and Math2 were downloaded for the purpose of experiments. A brief summary of them is shown in Table 6.

Table 6 A brief summary of three public datasets

Datasets	Students	Exercises		Knowledge Concepts
		Objective	Subjective	
FrcSub	536	20	0	8
Math1	4209	15	5	11
Math2	3911	16	4	16

All experiments were conducted using Visual Studio and run in an environment equipped with a Gen Intel(R) Core(TM)i5-14600HF 3.50 GHz CPU, 32 GB of RAM. Most figures for demonstration were drawn by Python. The hyper-parameters used in Algorithm 1 and Algorithm 2 are listed in Table 7.

Table 7 Hyper-parameters in Algorithm 1 and Algorithm 2

Alg.	**Algorithm 1**						**Algorithm 2**	
Par.	M	h	ϵ	M'	h'	ϵ'	M''	ϵ''
Val.	3×10^4	0.0001	1×10^{-9}	7×10^4	0.001	1×10^{-6}	20000	1×10^{-4}

The efficiency of models or their variants was indirectly assessed through its performance on prediction tasks. To compare the accuracy on prediction tasks, three evaluation metrics, root-mean-square error (RMSE), mean absolute error (MAE) and accuracy (ACC), were utilized. Particularly, ACC was applied on all exercises. Because the predicted scoring rates of the proposed model are all in the interval $[0, 1]$, in order to determine the proportion of accurately predicted samples to all samples, five different thresholds were applied. The average of ACC values associated with these thresholds was taken as the final ACC value.

6.2 Convergence Verification

Algorithm 2 was applied on the dataset MTest. To demonstrate the convergence of the algorithm, several parameters of different types were (randomly) selected. For instance, for the parameters of type $I_{sk}^{(0)}$, we randomly selected nine parameters for which the bottom labels (s, k) are $\{(5, 3), (5, 5), (5, 36), (5, 45), (15, 3), (15, 15), (15, 45), (25, 3), (25, 5)\}$. The values of these parameters as the iteration number increases were collected (see Fig. 9). It shows that the parameters' value tends to reach a stable state after enough iteration times.

To investigate the sensitivity of initial value selection to Algorithm 2, several parameters of different types were randomly selected. For each selected parameter,

[2] http://staff.ustc.edu.cn/ qiliuql/data/math2015.rar

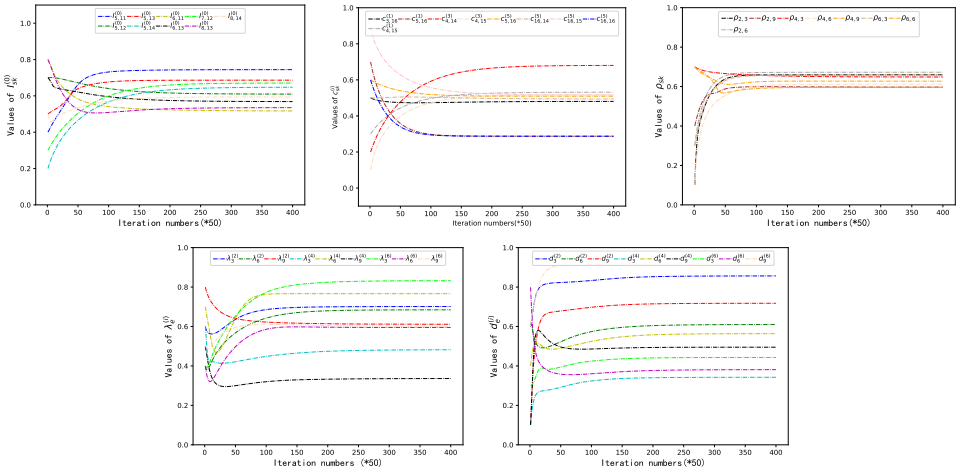


Fig. 9 Values' change trend of different types of parameters

three randomly initialized values were assigned. As the number of iterations increases, the values of each selected parameter with different initialized values finally approach the same one (see Fig. 10). This proves that the algorithm has good robustness to the initialization of parameters' values. It also shows that the proposed model has a strong robustness from the perspective of initialized value sensitivity.

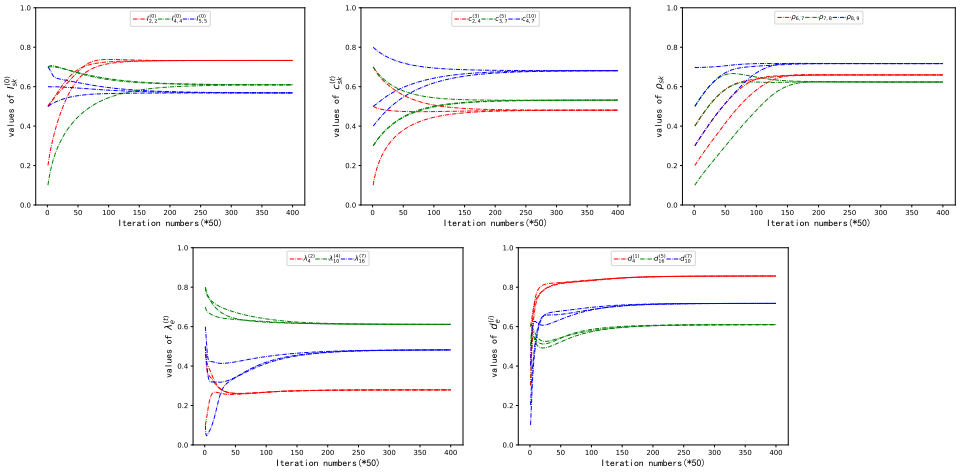


Fig. 10 Sensitivity of the model to parameters' initialized values

In order to comprehensively evaluate the convergence of the algorithm from the perspective of parameter iteration distance, we considered three types of iteration distances between parameter vectors. The first one is the Euclidean distance d_1. The second one is the mean absolute distance d_2. The third one is the root mean square

distance d_3. The iteration distances for different types of parameters were collected (see Fig. 11). It shows that the iteration distance becomes smaller as the iteration number increases. The result demonstrates the efficiency of Algorithm 2 from the perspective of convergence.

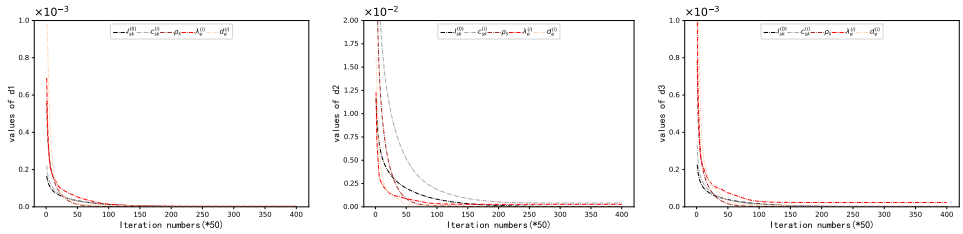


Fig. 11 Iteration distance change of different types of parameters

6.3 Comparative Experiments

6.3.1 Setting of Baselines

Comparative experiments were conducted in order to evaluate the effectiveness of the proposed model. Because the model is constructed for a new learning scenario in which the knowledge state changing has an intermittent feature, the existing cognitive diagnostic models in the literature were not fully replicated. Meanwhile, we notice that the calculation of simulated scoring rate and the recursive updating of knowledge states are the key mechanisms of the proposed model. Therefore, the following strategy is adopted for the comparative experiments design: retaining the overall framework of the proposed model, the scoring rate simulation method or the knowledge state updating mechanism is adjusted by referring to existing cognitive diagnostic models.

According to the calculation method of simulated scoring rates, the proposed model belongs to the 2-parameter IRT-based type. Therefore, other IRT-based types provide references for candidate comparative models. Furthermore, because students' proficiency on exercises can be evaluated according to the principle of fuzzy set theory, the model FuzzyCDF developed by Liu et al. (2018) is also taken into account. In this way, several models for comparison were developed as the baselines. Basic setups of these comparative models are presented in order to make these baselines more transparent.

CD1P: This model belongs to the 1-parameter IRT-based type since it only incorporates parameters $d_e^{(i)}$ on individual exercises. In order to highlight the value of the recursive updating of knowledge states, the learning rate and forgetting rate are retained in this model. The simulated scoring rate is calculated by

$$\eta_{se}^{(i)} = \frac{1}{1 + \exp[D(\theta_{se}^{(i)} - d_e^{(i)})]}.$$

The proficiency $\theta_{se}^{(i)}$ of students on individual exercises is consistent with the **INKT** model.

CD3P: This model accommodates guessing parameters on individual exercises. The scoring rate prediction is redefined as

$$\eta_{se}^{(i)} = g_e^{(i)} + \frac{(1 - g_e^{(i)})}{\exp[D\lambda_e^{(i)}(\theta_{se}^{(i)} - d_e^{(i)})]}$$

where $g_e^{(i)}$ represents students' guessing possibility on $e \in E_i$. Similar to the models **CD1P** and **INKT**, it also contains the learning rate and forgetting rate.

MPKT: This model adopts the calculation of a student's proficiency on an exercise as the arithmetic mean of the mastery across all knowledge concepts covered by the exercise. Particularly, the proficiency degree of student s on exercise $e \in E_i$ is calculated by

$$\theta_{se}^{(i)} = \frac{1}{|\widetilde{K}_e|} \sum_{k \in \widetilde{K}_e} I_{sk}^{(i)}.$$

This model also incorporates the effects of learning rate and forgetting rate. The prediction for students' scoring rates on exercises is consistent with the **INKT** model.

IFKT: In this model, the evaluation method of proficiency on exercises used in the model FuzzyCDF is accommodated. Specifically, on objective exercises, students' proficiency is determined by the minimum mastery on related knowledge concepts, i.e.,

$$\theta_{se}^{(i)} = \bigwedge_{k \epsilon \widetilde{K}_e} I_{sk}^{(i)}.$$

On subjective exercises, the proficiency is determined by the maximum mastery on related knowledge concepts, i.e.,

$$\theta_{se}^{(i)} = \bigvee_{k \epsilon \widetilde{K}_e} I_{sk}^{(i)}.$$

Other configurations are consistent with the **INKT** model. The parameter learning is realized through the well-known Monte Carlo algorithm.

In **INKT**, the parameters $c_{sk}^{(i)}$ and ρ_{sk} share identical hyper-parameters as follows:

$$\alpha_{c_{sk}^{(i)}} = \beta_{c_{sk}^{(i)}} = 1.5, \quad \alpha_{\rho_{sk}} = \beta_{\rho_{sk}} = 1.5.$$

For the parameters $d_e^{(i)}$, the associated hyper-parameters are calculated via the following formulas:

$$\alpha_{d_e^{(i)}} = 1.5, \quad \beta_{d_e^{(i)}} = \frac{0.5}{\overline{\mathbf{d}}_e^{(i)}} + 0.5.$$

Here, $\overline{\mathbf{d}}_e^{(i)} = 1 - \frac{\hat{\mathbf{d}}_e^{(i)}}{\max(\mathbf{d}_e)}$ for which $\hat{\mathbf{d}}_e^{(i)}$ represents the average score of exercise $e \in \tau_i$ across all students and $\max(\mathbf{d}_e)$ represents the full score of exercise e. For the

parameters $\lambda_e^{(i)}$, the associated hyper-parameterws are calculated via the following formulas:

$$\alpha_{\lambda_e^{(i)}} = 1.5, \quad \beta_{\lambda_e^{(i)}} = \frac{0.5}{\overline{\lambda}_e^{(i)}} + 0.5.$$

Here, $\overline{\lambda}_e^{(i)} = P_H - P_L$ is calculated based on the real scoring data. Specifically, P_H is the correct response rate of exercise e among the top 27% of students and P_L is the correct response rate of exercise e among the bottom 27% of students. All trainable parameters are randomly initialized and the step size for gradient descent in **INKT** is set as 0.0001.

6.3.2 Comparison Result Analysis

The performance of five models on the dataset MTest was collected (see Fig. 12 and Table 8). In experimental scenarios with different tests proportions and students proportions, the comparative models and **INKT** exhibit a significant and consistent divergence in performance. The overall results show that **INKT** demonstrates stable predictive superiority across all experimental conditions, which fully validates its soundness and efficiency.

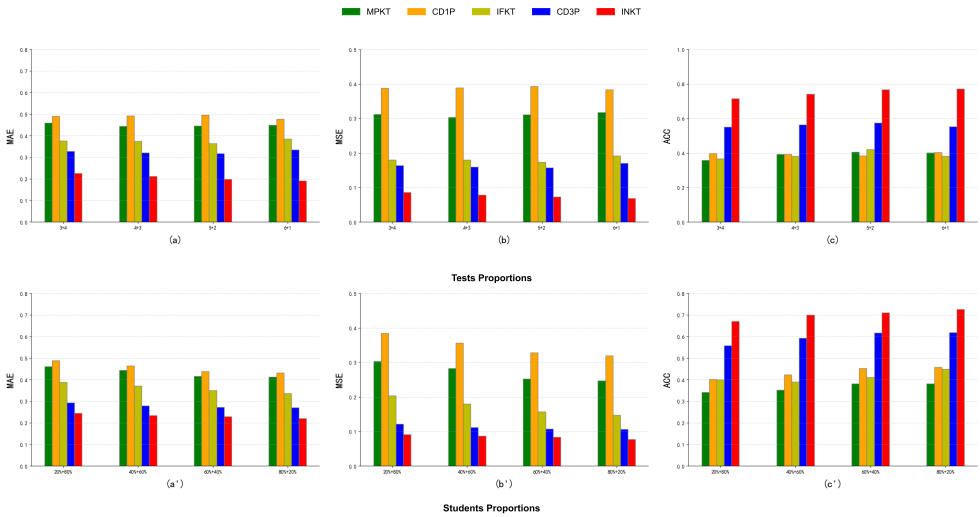


Fig. 12 Model performance comparison under different tests/students proportions on MTest

In comparative experiments with different tests proportions, the performance of **INKT** comprehensively surpasses that of **IFKT**. Quantitatively, **INKT** maintains ACC in the range of 0.71-0.77, MSE of 0.07-0.09, and MAE of 0.21-0.23. In contrast, the ACC of **IFKT** is generally below 0.60, with MSE exceeding 0.15 and MAE approaching 0.30. This gap primarily stems from the difference in model structures: **IFKT** typically adopt discrete classification logic, simplifying students' performance

Table 8 Comparative experiment results on the dataset MTest

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.3589	0.3119	0.4601	0.3922	0.3032	0.4448	0.4067	0.3106	0.4455	0.4024	0.3175	0.4500
IFKT	0.3666	0.1798	0.3772	0.3809	0.1802	0.3745	0.4211	0.1740	0.3641	0.3822	0.1922	0.3850
CD1P	0.3981	0.3878	0.4905	0.3927	0.3893	0.4921	0.3840	0.3933	0.4953	0.4041	0.3838	0.4773
CD3P	0.5505	0.1639	0.3277	0.5629	0.1594	0.3214	0.5755	0.1571	0.3173	0.5532	0.1702	0.3346
INKT	**0.7158**	**0.0858**	**0.2261**	**0.7409**	**0.0786**	**0.2129**	**0.7664**	**0.0728**	**0.1990**	**0.7723**	**0.0691**	**0.1915**

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.3411	0.3032	0.4613	0.3520	0.2831	0.4448	0.3818	0.2523	0.4159	0.3813	0.2473	0.4121
IFKT	0.4004	0.2041	0.3883	0.3903	0.1800	0.3719	0.4112	0.1575	0.3514	0.4484	0.1481	0.3370
CD1P	0.4018	0.3847	0.4884	0.4228	0.3564	0.4653	0.4523	0.3282	0.4385	0.4587	0.3200	0.4322
CD3P	0.5576	0.1219	0.2933	0.5937	0.1120	0.2794	0.61177	0.1080	0.2733	0.6194	0.1061	0.2716
INKT	**0.6713**	**0.0913**	**0.2242**	**0.7000**	**0.0870**	**0.2338**	**0.7110**	**0.0843**	**0.2295**	**0.7263**	**0.0776**	**0.2198**

into binary judgment such as "mastered/not mastered" or "high/low proficiency", which lacks fine-grained quantification of knowledge states. In contrast, **INKT** characterizes students' knowledge state through continuous output, thereby avoiding the coarse treatment of intermediate states inherent in traditional models.

Although guessing parameters are incorporated into **CD3P** besides difficulty parameters on individual exercises in **CD1P** or **MPKT**, the gaps between their performances are not large. This result indicates that under the condition of setting knowledge states unrelated to each other, adding more parameters may not bring much benefit for the model performance.

In the comparative experiments with different students proportions, the advantage of **INKT** becomes more pronounced. Under an imbalanced sample distribution of 80%+20%, **CD1P** achieves ACC of only 0.6784, with MSE and MAE as high as 0.1421 and 0.2987, respectively. In contrast, **INKT** attains ACC of 0.7263, reduces MSE to 0.0776, and optimizes MAE to 0.2198, maintaining the best performance across all metrics. Even in a balanced sample scenario of 60%+40%, **INKT**'s ACC remains 8.3% higher than that of the second-best model **CD3P**, while error metrics are reduced by approximately 35%. This result verifies the value of integrating learning and forgetting rates in **INKT**. It indicates that the recursive updating mechanism plays an important role in incorporating students' response performance at different time stages into a dynamic modeling system, effectively reducing the impact of sample distribution fluctuations on prediction stability.

The underlying rationale for these results is analyzed: traditional models are essentially "item-oriented" static paradigms that rely solely on inherent item attributes such as difficulty and guessing parameters for prediction, completely overlooking the dynamic laws of "knowledge accumulation and forgetting decay" during the learning

process. Although the **CD1P** and **CD3P** models, as improvements over traditional one-parameter and three-parameter models, take into account the dynamic changes of knowledge states, their prediction performance is still inferior to that of the **INKT** model even with the incorporation of temporal dynamic features. In contrast, our proposed model **INKT** constructs a three-dimensional "student-time-item" dynamic modeling framework. The most obvious advantage of such framework lies in its capability of simulating the temporal evolution of knowledge states by incorporating learning/forgetting parameters and meanwhile considering the item discrimination.

The proposed recursive mechanism of updating knowledge states not only breaks through the limitation of the static assumption of classical IRT-based models, but also makes up for the deficiencies in quantification precision of models such as **IFKT** and **MPKT**, thereby achieving comprehensive performance superior to that of traditional models and related improved models. These findings strongly confirm the critical role of temporal dynamic features and learning process parameters in answer prediction tasks, and provide important references for the design of personalized prediction models in the field of knowledge tracing.

In order to evaluate the performance of our proposed model on the larger-scale dataset FMTest, we conducted comparative experiments under four different students proportions: 20%+80%, 40%+60%, 60%+40% and 80%+20%. As clearly demonstrated in Fig. 13 and Table 9, the **INKT** model consistently outperforms all other baseline models across all three evaluation metrics (ACC, MSE and MAE) under every proportion. The result fully validates the performance superiority of the **INKT** framework. This improvement can be attributed to the model's explicit incorporation of item discrimination to refine the characterization of students' ability, combined with dynamic tracing of the evolutionary process of knowledge states, ultimately leading to a more accurate and robust prediction performance. Notably, the introduction of guessing parameters did not yield a significant improvement in model performance. This phenomenon indicates that a reasonable parameter configuration is sufficient to enable the model to achieve optimal efficacy. Conversely, the addition of redundant parameters not only fails to deliver performance gains but also undermines the inherent advantages of the **INKT** model.

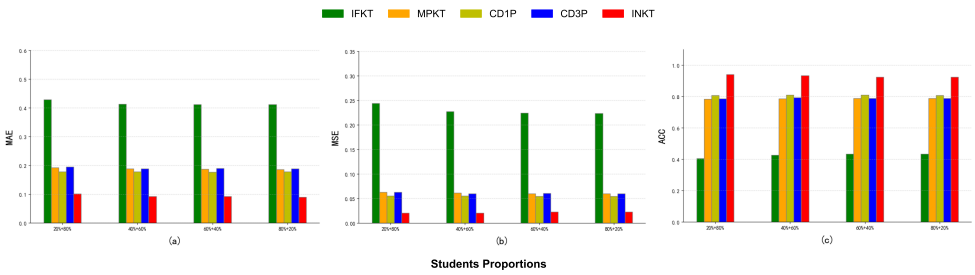


Fig. 13 Model performance comparison under different students proportions on FMTest

Table 9 Comparative experiment results on the dataset FMTest

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.7835	0.0630	0.1929	0.7856	0.0613	0.1892	0.7887	0.0603	0.1878	0.7898	0.0600	0.1866
IFKT	0.4065	0.2441	0.4297	0.4278	0.2272	0.4142	0.4344	0.2244	0.4117	0.4342	0.2241	0.4118
CD1P	0.8089	0.0555	0.1789	0.8106	0.0553	0.1781	0.8116	0.055	0.1772	0.8089	0.0552	0.1782
CD3P	0.7872	0.0634	0.1952	0.7937	0.0604	0.1889	0.7900	0.0612	0.1896	0.7885	0.0601	0.1888
INKT	**0.9426**	**0.0207**	**0.1018**	**0.9354**	**0.0210**	**0.0928**	**0.925**	**0.0230**	**0.0923**	**0.9248**	**0.0231**	**0.0908**

Comparative experiments were also conducted on public datasets Math1, Math2 and FrcSub. The experimental results of all datasets under various evaluation metrics are presented (see Tables 10,11,12). To more intuitively reflect the performance differences among different models, we further draw corresponding bar charts for each dataset (see Fig.s 14,15,16). It can be clearly observed from the results that the proposed INKT model achieves optimal or sub-optimal performance across all datasets and maintains stable overall performance.

Table 10 Comparative experiment results on the dataset Math1

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.5872	0.2108	0.3272	0.4806	0.2887	0.4053	0.3884	0.2986	0.4459	0.6641	0.0921	0.2427
IFKT	0.3659	0.2943	0.4413	0.4699	0.2443	0.3770	0.6350	0.1482	0.2781	0.9304	0.0332	0.1432
CD1P	**0.6410**	**0.1494**	**0.2788**	**0.5256**	0.2517	0.3726	0.3137	0.4107	0.5425	0.7750	0.0543	0.1854
CD3P	0.4662	0.1903	0.3558	0.5761	0.1634	0.3089	0.7426	0.0799	0.2043	0.8926	0.0349	0.1485
INKT	0.3286	0.1758	0.3731	0.4670	**0.1430**	**0.3173**	**0.7824**	**0.0646**	**0.1864**	**0.9552**	**0.0234**	**0.1275**

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.3411	0.3032	0.4613	0.3520	0.2831	0.4448	0.3818	0.2523	0.4159	0.3813	0.2473	0.4121
IFKT	0.3195	0.3070	0.46519	0.3167	0.3096	0.4622	0.3174	0.3059	0.4609	0.3162	0.3044	0.4602
CD1P	**0.6916**	0.0959	**0.2395**	**0.6903**	0.0958	**0.2393**	**0.6783**	0.1031	**0.2457**	**0.6618**	0.1138	0.2527
CD3P	0.6221	0.0969	0.2588	0.6202	0.1004	0.2583	0.6217	0.1001	0.2578	0.6200	0.1008	0.2588
INKT	0.6331	**0.0945**	0.2572	0.6476	**0.0904**	0.2484	0.6473	**0.0905**	0.2484	0.6478	**0.0910**	**0.2483**

Table 11 Comparative experiment results on the dataset Math2

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.4376	0.1994	0.3702	0.4540	0.1958	0.3612	0.4531	0.1625	0.3323	0.2795	0.1719	0.3717
IFKT	0.4746	0.2904	0.3931	0.5306	0.2919	0.3679	0.5750	0.2264	0.3184	**0.7789**	**0.0583**	**0.1495**
CD1P	0.5280	0.1648	0.3184	0.5780	0.1658	**0.3047**	0.5743	0.1553	**0.3033**	0.5458	0.1592	0.3187
CD3P	**0.5505**	0.1639	0.3277	0.5629	0.1594	0.3214	**0.5755**	0.1571	0.3173	0.5532	0.1702	0.3346
INKT	0.5312	**0.1600**	**0.3176**	**0.5784**	**0.1618**	0.3053	0.5596	**0.1531**	0.3066	0.5364	0.1475	0.3282

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.5626	0.1169	0.2867	0.5486	0.1176	0.2908	0.5494	0.1178	0.2904	0.5506	0.1169	0.2898
IFKT	0.3673	0.3042	0.4517	0.3633	0.3024	0.4504	0.3623	0.3002	0.4492	0.3625	0.2999	0.4491
CD1P	**0.6227**	**0.1010**	0.4884	**0.6240**	**0.1005**	0.4653	0.6224	**0.1006**	0.4385	**0.6258**	**0.1008**	0.4322
CD3P	0.5576	0.1219	0.2933	0.5937	0.1120	0.2794	**0.6117**	0.1080	0.2733	0.6194	0.1061	0.2716
INKT	0.6063	0.1050	**0.2692**	0.6164	0.1022	**0.2636**	0.6199	0.1019	**0.2625**	0.6212	0.1024	**0.2631**

Table 12 Comparative experiment results on the dataset FrcSub

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.6156	0.1490	0.2798	0.7020	0.1318	0.2396	0.7799	0.1054	0.1919	0.7764	**0.0718**	**0.155**
IFKT	0.1744	0.2911	0.4885	0.2267	0.3153	0.4917	0.2447	0.3098	0.4808	0.6082	0.3884	0.3910
CD1P	**0.6302**	0.1383	**0.2680**	**0.7106**	0.1267	**0.2297**	**0.7907**	0.1049	**0.1857**	**0.7767**	0.0722	0.1553
CD3P	0.6121	**0.1350**	0.2778	0.7016	**0.1245**	0.2414	0.7717	0.1066	0.2024	0.7691	0.0734	0.1674
INKT	0.6257	0.1371	0.2692	0.7078	0.1262	0.2313	0.7861	**0.1048**	0.1878	0.7753	0.0726	0.1565

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
MPKT	0.7479	0.0731	0.2000	0.7644	0.0683	0.1924	0.7655	0.0673	0.1909	**0.7811**	**0.0646**	**0.1833**
IFKT	0.3611	0.2924	0.4523	0.2881	0.2392	0.4430	0.2468	0.2352	0.4460	0.2366	0.2328	0.4458
CD1P	**0.7625**	**0.0705**	**0.1908**	**0.7741**	**0.0667**	**0.1870**	**0.7688**	0.0676	**0.1892**	0.7754	0.0648	0.1846
CD3P	0.5576	0.1219	0.2933	0.5937	0.1120	0.2794	0.61177	0.1080	0.2733	0.6194	0.1061	0.2716
INKT	0.7413	0.0744	0.2055	0.7599	0.0691	0.1949	0.7674	**0.0666**	0.1904	0.7692	0.0658	0.1890

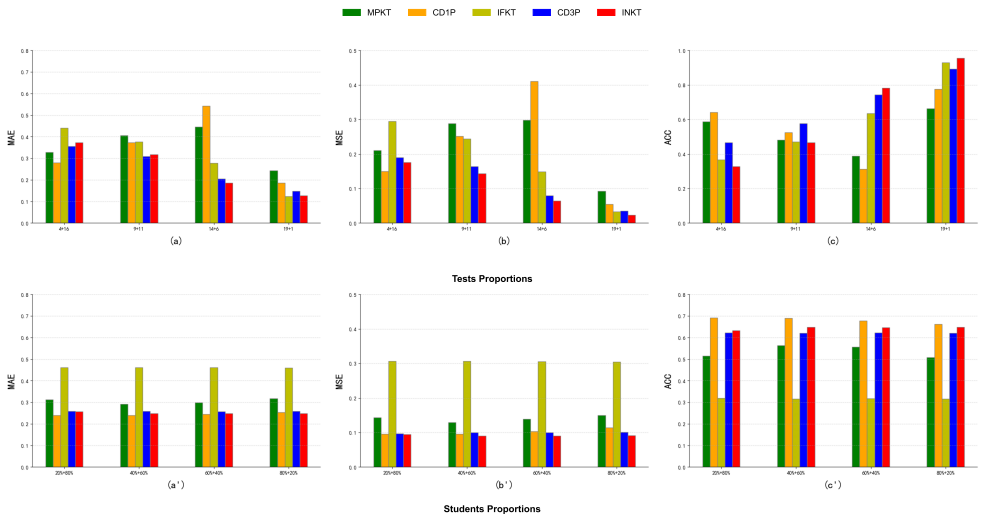


Fig. 14 Model performance comparison under different tests/students proportions on Math1

In the comparative experiments on the public Math1 dataset, the **INKT** model and the best model achieve almostly the same performance, with the differences in all evaluation metrics controlled within a small range and no obvious performance gap is observed. In the training data scenario divided by student sample ratios, however, the **INKT** model performs slightly inferior to the best model, but the performance difference between the two models is within an acceptable range, without significant performance gap. On the other two public datasets Math2 and FrcSub, the **INKT** model also performs slightly lower than the best model. Nevertheless, all evaluation metrics of **INKT** still maintain a good level with no sharp decline in performance, which demonstrates its strong stability. It should be specially noted that all three public datasets (Math1, Math2, FrcSub) adopted in this experiment have certain limitations: they only include single independent exercises, without covering complete test scenarios and continuous answering processes. Therefore, in the experimental analysis, this paper treats each single exercise as an independent test sample to ensure the rationality and objectivity of the experimental evaluation. In addition, the adjustment of student sample ratio and the change of student number further verify the superiority of the **INKT** model: as the number of students participating in the experiment gradually increases, the model's characterization of students' knowledge states becomes more and more accurate, and the comprehensive prediction performance shows a steady upward trend. This observation indirectly confirms a good adaptability of the **INKT** model in large-scale student groups.

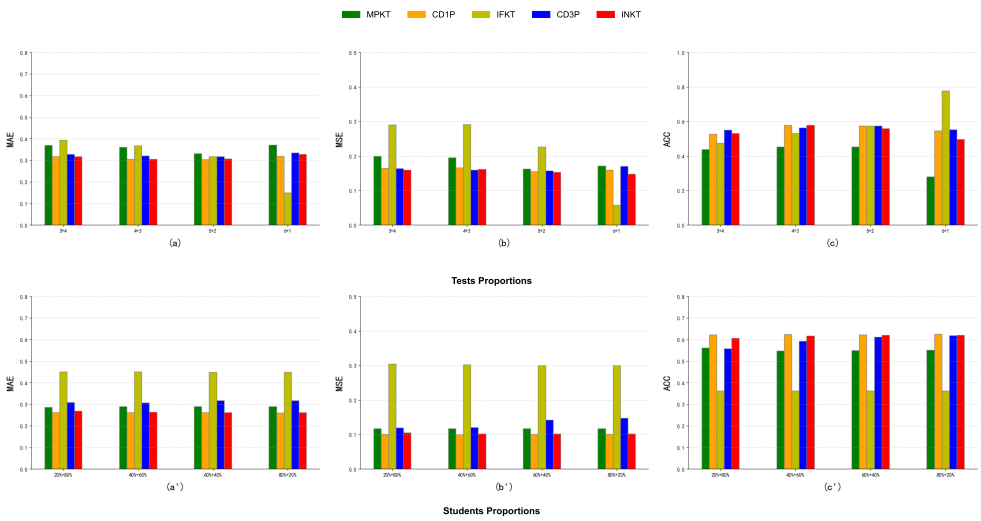


Fig. 15 Model performance comparison under different tests/students proportions on Math2

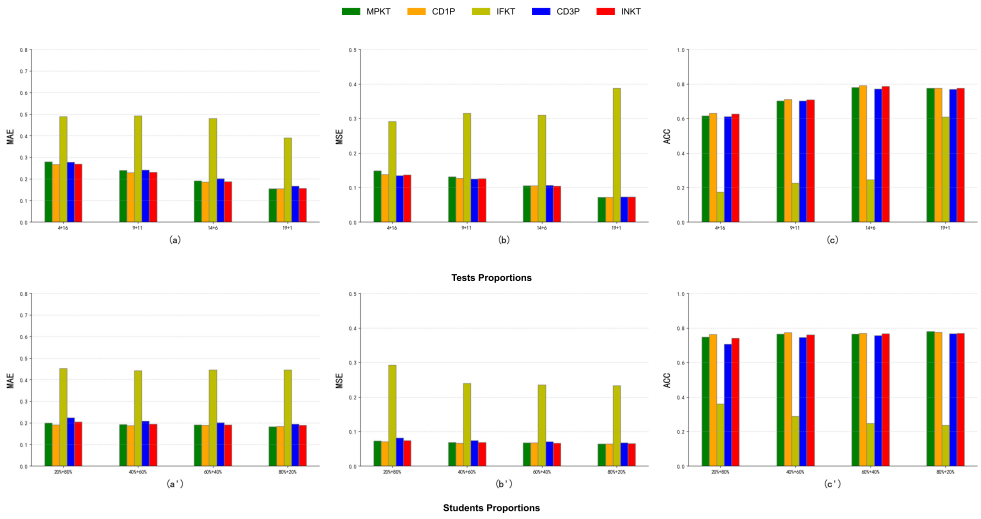


Fig. 16 Model performance comparison under different tests/students proportions on FrcSub

6.4 Experiments for Ablation

6.4.1 Setting for Ablation

The ablation experiments were conducted in order to study the impact of various components to the model performance. Particularly, six variants were developed through

adjusting modules or parameters contained in **INKT**. By comparing their performance, the indispensability of several components was verified. The detail setting of the variants are as follows:

Var$_1$: The parameters ρ_{sk} are removed from the module of knowledge tracing. The transition probability from $I_{sk}^{(i)}$ to $I_{sk}^{(i+1)}$ depends entirely on the learning rate $c_{sk}^{(i)}$.

Var$_2$: The parameters $c_{sk}^{(i)}$ are removed from the module of knowledge tracing. The transition probability from $I_{sk}^{(i)}$ to $I_{sk}^{(i+1)}$ depends entirely on the forgetting rate ρ_{sk}.

Var$_3$: The augmented Q-matrix is substituted with the original Q-matrix, while all other experimental configurations remain identical to those used in **INKT**.

Var$_4$: All non-zero elements in the augmented Q-matrix are binarized to 1 and serve as a replacement for the augmented Q-matrix. All other settings are consistent with **INKT**.

Var$_5$: The parameters ρ_{sk} are replaced by ρ_s. All other parameters remain consistent with **INKT**.

The following strategy is adopted for the ablation experiments. Firstly of all, the training of each variant is conducted on the entire data. Values of all parameters, denoted by $\tilde{\lambda}_e^{(i)}, \tilde{d}_e^{(i)}, \tilde{c}_{sk}^{(i)}, \tilde{I}_{sk}^{(0)}$ and $\tilde{\rho}_{sk}$ for convenience, are stored as reusable data for subsequent data experiments. In detail, such subsequent experiments is divided into the following parallel parts.

(1) The entire dataset is split into two categories according to different tests proportions. Under each proportion, one part is for training and the other is for testing. Based on the data for training, parameters $I_{sk}^{(0)}$ and ρ_{sk} are trained with other parameters $\lambda_e^{(i)}, d_e^{(i)}, c_{sk}^{(i)}$ unchanged (i.e., $\lambda_e^{(i)} = \tilde{\lambda}_e^{(i)}, d_e^{(i)} = \tilde{d}_e^{(i)}, c_{sk}^{(i)} = \tilde{c}_{sk}^{(i)}$). Consequently, the scoring rates of all students on the exercises contained in the data for testing are predicted.

(2) The entire dataset is randomly split into two subsets according to different students proportions. Under each split proportion, one subset is used for training and the other for testing. Based on the training data, parameters $\lambda_e^{(i)}$ and $d_e^{(i)}$ are trained while other parameters $I_{sk}^{(0)}$, ρ_{sk} and $c_{sk}^{(i)}$ remain unchanged (i.e., $I_{sk}^{(0)} = \tilde{I}_{sk}^{(0)}$, $\rho_{sk} = \tilde{\rho}_{sk}$, $c_{sk}^{(i)} = \tilde{c}_{sk}^{(i)}$). Consequently, the scoring rates of the other students on the exercises in the testing data are predicted.

6.4.2 Ablation Result Analysis

Based on dataset MTest, we evaluated the performance of the variant models on evaluation metrics ACC, MSE and MAE respectively. Particularly, for the variant **Var**$_1$, five assignments of topological index on knowledge concepts were randomly given and then the average of the results was calculated. For **Var**$_2$, the original Q-matrices were directly used for the module of knowledge tracing. The experimental results are present in Fig. 17 and Table 13.

The results demonstrate that removing either forgetting rate $f_{sk}^{(i)}$ (in **Var**$_1$) or learning rate $c_{sk}^{(i)}$ (in **Var**$_2$) results in significantly worse performance on all metrics across all proportions on tests or students. It verifies the rationality of incorporating

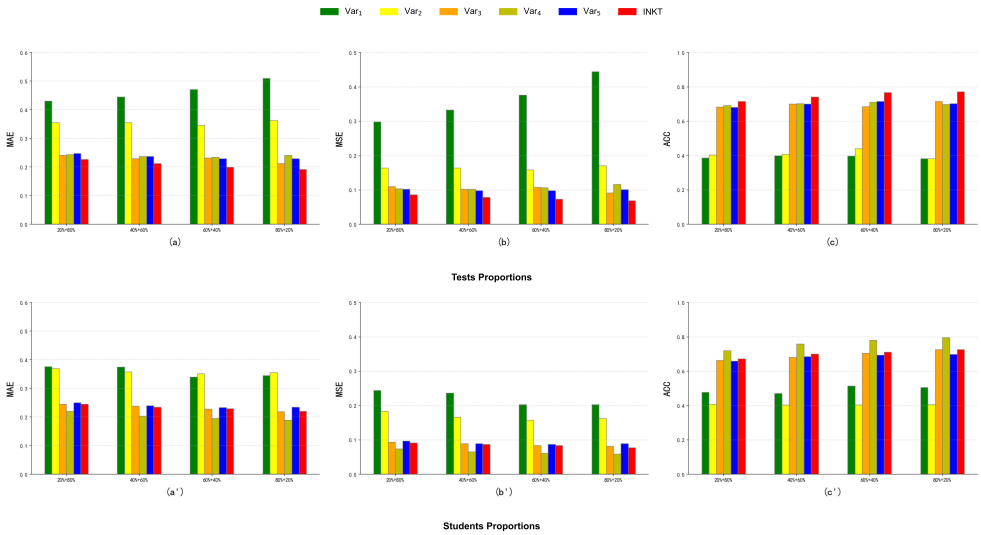


Fig. 17 Ablation experiment results under different tests/students proportions on MTest

Table 13 Ablation experiment results on the dataset MTest

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.3867	0.2981	0.4303	0.3984	0.3327	0.4452	0.3964	0.3756	0.4703	0.3821	0.4444	0.5101
Var_2	0.4035	0.1634	0.3556	0.4086	0.1636	0.3554	0.4397	0.1583	0.3460	0.3823	0.1700	0.3633
Var_3	0.6906	0.1033	0.2435	0.7029	0.1020	0.2368	0.7116	0.1065	0.2339	0.6979	0.1162	0.2413
Var_4	0.6813	0.1025	0.2469	0.7008	0.0978	0.2370	0.7157	0.0975	0.2291	0.7016	0.1006	0.2289
Var_5	0.6829	0.1093	0.2405	0.7003	0.1022	0.2290	0.6846	0.1077	0.2315	0.7152	0.0913	0.2127
Var_6	0.7138	0.0964	0.2297	0.7324	0.0894	0.2181	0.7364	0.0856	0.2112	0.7129	0.0928	0.2199
INKT	**0.7158**	**0.0858**	**0.2261**	**0.7409**	**0.0786**	**0.2129**	**0.7664**	**0.0728**	**0.1990**	**0.7723**	**0.0691**	**0.1915**

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.4766	0.2438	0.3761	0.4709	0.2368	0.3745	0.5144	0.2026	0.3393	0.5044	0.2033	0.3443
Var_2	0.4077	0.1830	0.3694	0.4041	0.1657	0.3579	0.4044	0.1569	0.3506	0.4051	0.1629	0.3551
Var_3	**0.7194**	**0.0743**	**0.2195**	**0.7583**	**0.0653**	**0.2031**	**0.7811**	**0.0608**	**0.1948**	**0.7943**	**0.0588**	**0.1894**
Var_4	0.6599	0.0964	0.2493	0.6847	0.0897	0.2391	0.6947	0.0868	0.2337	0.6978	0.0893	0.2344
Var_5	0.6630	0.0939	0.2452	0.6796	0.0895	0.2383	0.7042	0.0841	0.2191	0.7026	0.0812	0.2242
Var_6	0.6773	0.0927	0.2393	0.7128	0.0825	0.2226	0.7200	0.0833	0.2211	0.7306	0.0805	0.2177
INKT	0.6713	0.0913	0.2242	0.7000	0.087	0.2338	0.7110	0.0843	0.2295	0.7263	0.0776	0.2198

both types of rates in the **INKT** model. With the mechanisms of forgetting and learning, the model better aligns with students' real learning processes and delivers accurate judgments.

Notably, the model variants $\mathbf{Var}_3$ and $\mathbf{Var}_4$ performe prominently in the model evaluation, demonstrating strong predictive performance, with $\mathbf{Var}_3$ standing out particularly. Specifically, $\mathbf{Var}_3$ adopts the original Q-matrix without enhancement for model training. In the experimental scenario where training data and test data are divided according to student sample ratios, all evaluation metrics of $\mathbf{Var}_3$ reach the optimal level, showing its excellent predictive capability. Even so, the **INKT** model still possesses significant and irreplaceable advantages. These advantages are reflected not only in the accuracy and stability of model prediction, but also in the interpretability of the model. Compared with variant models such as $\mathbf{Var}_3$, the **INKT** model has a clearer modeling logic and more theoretically sound parameter settings, which can clearly explain the dynamic evolution process of students' knowledge states. The detailed discussion on this point has been elaborated in the relevant previous sections of this paper.

Replacing only ρ_{sk} by ρ_s (in $\mathbf{Var}_5$) causes only slight performance decline. It indicates that disconnecting the forgetting rate or learning rate with knowledge concepts has limited impact on overall model performance. Notably, **INKT** outperforms all other variants comprehensively when data is divided according to tests proportions. For example, under the 3+4 tests proportion, **INKT** achieves ACC of 0.7158 and MSE of 0.0858, both of which are optimal.

The results of ablation experiments on three public datasets (Math1, Math2, FrcSub) are presented in Fig.s B1, B2 and B3 and Tables B1, B2 and B3 (see Appendix B). Obviously, the **INKT** model has consistently maintained a good performance level, demonstrating strong generalization ability and stability. Admittedly, the **INKT** model performs slightly worse than $\mathbf{Var}_3$ in some evaluation metrics, but the performance gap between the two is within an acceptable range. Moreover, the **INKT** model's performance does not show obvious fluctuations and always remains at a high level, being able to adapt to the characteristics of different public datasets. This fully proves that the model has good generalization performance and can maintain stable prediction results in different data scenarios.

6.5 Case Study

6.5.1 Cases of Knowledge Proficiency

For the case study of the proposed model **INKT** on knowledge proficiency, two representative students, s_{10} and s_{33}, were selected from the dataset MTest. In order to elucidate the nuanced impact of learning and forgetting effects on long-term knowledge retention, we empirically delineated the dynamic evolutionary trajectories of their knowledge proficiency across distinct knowledge concepts over sequential tests. For demonstration purposes, the mastery of student s_{10} and s_{33} on several specific knowledge concepts is presented in Fig. 18.

s_{10}: On most knowledge concepts, the proficiency of s_{10} sustains at a relatively high level. The overall result is consistent with the relatively better observable performance

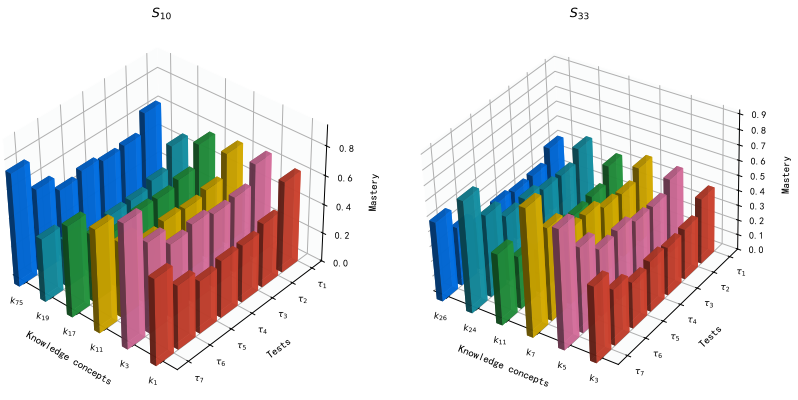


Fig. 18 Mastery on several specific knowledge concepts of students s_{10} and s_{33}

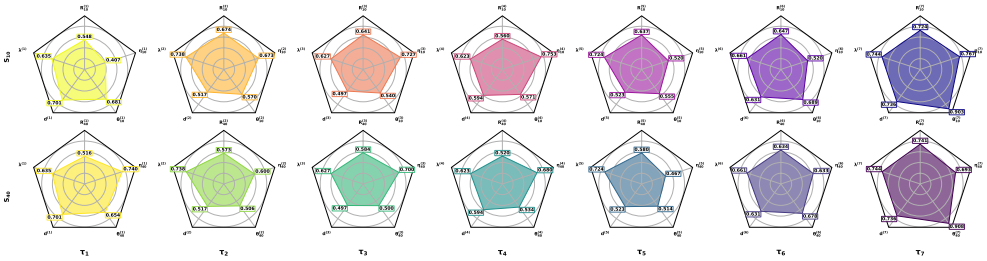


Fig. 19 Dynamic learning states of student s_{10} and s_{40}

of s_{10}. Moreover, a salient deficiency is identified in the proficiency on knowledge concept k_{19}. Particularly, such proficiency persistently remains below 0.4 throughout the time window from τ_3 to τ_6. It demonstrates the rationality of the proposed model in the sense that even relatively excellent students may perform poorly on certain knowledge concepts. The reason for this phenomenon may be partly attributed to the knowledge-context-aware augmentation of Q-matrices which brings a refinement on the connection between exercises and knowledge concepts.

s_{33}: By contrast, the overall knowledge proficiency of s_{33} is marginally inferior than that of s_{10}. Meanwhile, on several specific knowledge concepts, s_{33} possesses better mastery than s_{10}. Moreover, a pronounced upward trend can be observed in proficiency on k_5, k_7 and k_{24} over successive test assessments. Such pattern exemplifies the dynamic equilibrium between learning and forgetting effects in simulating students' knowledge state changing. In more detail, knowledge proficiency is not static but evolves dynamically in response to practice on exercises. It is consistent with a fact that students with an initial low knowledge proficiency can obtain measurable

improvement in subsequent evaluative outcomes via targeted and sustained practice on exercises. The effect of the augmentation of Q-matrices can be also speculated.

In a brief summary, the visualization outcome unequivocally reveals divergent and informative patterns of knowledge concept proficiency. The result not only unveils the inherent complexity of the learning process but also validates the efficiency of the proposed model in capturing students' latent knowledge states. Meanwhile, the updated Q-matrices further refine the exercise-knowledge component mapping, enabling the precise capture of the subtle yet consequential variations in students' knowledge proficiency.

6.5.2 Cases of Learning States

For the case study on learning states, two representative students, s_{10} and s_{40}, were selected from the dataset MTest. Radar charts in Fig. 19 depict their dynamic evolution of the learning states across seven tests. By integrating the test difficulty $d^{(i)}$ and test discrimination $\lambda^{(i)}$, a systematic analysis is given towards the intrinsic correlation among students' latent ability $\theta^{(i)}$, actual scoring rates $R^{(i)}$ and predicted scoring rates $\eta^{(i)}$ on tests. Distinct evolutionary characteristics are observed in the learning states of students s_{10} and s_{33}, which verifies the advantage of the proposed model. On the whole, the latent ability $\theta^{(i)}$ and actual/predicted scoring rates ($R^{(i)}/\eta^{(i)}$) show a consistent upward trend. Specifically, a surge-type growth on student s_{10} in high-discrimination environment is observed, and meanwhile a progressive improvement on student s_{33} under stable test difficulty is demonstrated. The dynamic variations in test difficulty and discrimination further amplifies such individual differences.

In the early time windows from $\tau^{(1)}$ to $\tau^{(4)}$, small deviation between actual and simulated scoring rates is observed and gentle growth in latent ability is shown on both students s_{10} and s_{40}. The proposed model achieves a high consistency between simulated scoring rates and actual ones through relatively accurate estimation of initial learning and forgetting rates. In the middle time windows from $\tau^{(5)}$ to $\tau^{(6)}$, with the increase of test discrimination, the scoring rates and latent ability of both students rise synchronously. By dynamically updating the learning rate, the proposed model efficiently captures the difference in learning gains between two students, improving the prediction accuracy. In the final time window $\tau^{(7)}$ in which test difficulty and discrimination reach their peaks, the latent ability of both students is significantly enhanced. The dynamic correction mechanism for the forgetting rate in the proposed model effectively offset the knowledge decay effect in long-term testing, avoiding the prediction bias of traditional models and further verifying the robustness of the proposed model.

In a brief summary, high-discrimination tests benefit the identification of students' latent ability and thus amplify individual difference among students' learning performance. Through the dynamic learning and forgetting rates, the proposed model not only efficiently captures the quantitative correlation between the evolutionary characteristics of students' latent ability and test attributes, but also reduces the prediction error. Its capability of estimating personalized parameters can adapt to the distinct learning characteristics of students and realize rational tracing of their knowledge states.

7 Conclusion, Limitations and Future Works

Towards the tasks of tracing students' knowledge state with characteristics of stability in short time range and variability across different stages, this paper proposed a new knowledge tracing model **INKT**. Firstly, an approach of topologically sorting knowledge concepts was proposed by exploiting the complete paths contained in the knowledge concept graph. Secondly, a method was proposed for the purpose of realizing the knowledge-context-aware augmentation of Q-matrices by requiring the updated Q-matrices have best compatibility with the topological sorting on knowledge concepts. Thirdly, based on the augmented Q-matrices, an intermittently knowledge tracing model was developed by exploring Bayesian formulas and item response theory. By integrating learning and forgetting parameters, the proposed model is expected to accurately capture the temporal evolution of knowledge mastery, which not only overcomes the static assumption limitation of classical IRT models but also addresses the quantification precision shortcomings of classical cognitive diagnostic models. Ablation and comparative experiments were conducted, verifying the efficiency and rationality of the proposed model.

Both the linear order on knowledge concepts and the preliminary relationship between knowledge concepts can be seen as special knowledge graphs (Hofer et al. 2024; Pan et al. 2024). An obvious limitation is that the preliminary relationship is provided by experts, which has a certain degree of subjective bias. Therefore, methods of automatically obtaining knowledge graphs from data is worth researching. Moreover, it has been shown that there exist other various relationship between knowledge concepts from the perspective of cognitive learning (Wang et al. 2024; Liu et al. 2023). In this sense, an interesting work in the future is the study of other linear sorting methods or nonlinear sorting strategies. Furthermore, in the updating of Q-matrices, only the compatibility with the linear order on knowledge concepts is considered. However, there exists more complex logical structures on knowledge concepts. Therefore, a more comprehensive Q-matrix update strategy is a valuable topic worth studying. In addition, the recursive updating mechanism on knowledge states is based on the assumption that the current knowledge state is only related to the previous one. However, from the perspective of attention mechanisms, such assumption has certain limitations (Vaswani et al. 2017). Therefore, improving the recursive updating mechanism by incorporating attention mechanism is a meaningful work in the future.

Funding Declaration This research was financially supported by the National Natural Science Foundation of China (Grant No. 12171149, 12571495) and Research Project of Degree and Postgraduate Teaching Reform in Hunan Province (2023JGYB098). All funders provided financial support for the experimental implementation, data collection, data analysis, and manuscript writing of this research without any involvement in the study design, data interpretation, or decision to publish the results. The authors declare that there are no conflicts of interest related to the funding received for this work.

References

Ouyang, F., Jiao, P.: Artificial intelligence in education: The three paradigms. Computers and Education: Artificial Intelligence **2**, 100020 (2021). https://doi.org/10.1016/j.caeai.2021.100020

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

Kerimbayev, N., Adamova, K., Shadiev, R., Altinay, Z.: Intelligent educational technologies in individual learning: a systematic literature review. Smart Learning Environments **12**(1), 1 (2025). https://doi.org/10.1186/s40561-024-00360-3

Liu, Q., Shen, S., Huang, Z., Chen, E., Zheng, Y.: A survey of knowledge tracing (2021). https://doi.org/10.48550/arXiv.2105.15106

Abdelrahman, G., Wang, Q., Nunes, B.: Knowledge tracing: A survey. ACM Computing Surveys **55**(11), 1–37 (2023). https://doi.org/10.1145/3569576

Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction **4**(4), 253–278 (1994). https://doi.org/10.1007/BF01099821

De La Torre, J.: Dina model and parameter estimation: A didactic. Journal of educational and behavioral statistics **34**(1), 115–130 (2009). https://doi.org/10.3102/1076998607309474

Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User modeling and user-adapted interaction **27**(3), 313–350 (2017). https://doi.org/10.1007/s11257-017-9193-2

Käser, T., Klingler, S., Schwing, A.G., Gross, M.: Dynamic bayesian networks for student modeling. IEEE Transactions on Learning Technologies **10**(4), 450–462 (2017). https://doi.org/10.1109/TLT.2017.2689017

Cen, H., Koedinger, K., Junker, B.: Learning factors analysis–a general method for cognitive model evaluation and improvement. In: International Conference on Intelligent Tutoring Systems, pp. 164–175. Springer, ??? (2006). https://doi.org/10.1007/11774303_17

Pavlik Jr, P.I., Cen, H., Koedinger, K.R.: Performance factors analysis–a new alternative to knowledge tracing. In: International Conference on Artificial Intelligence in Education, pp. 531–538 (2009). https://doi.org/10.3233/978-1-60750-028-5-531

Liu, Z., Guo, T., Liang, Q., Hou, M., Zhan, B., Tang, J., Luo, W., Weng, J.: Deep learning based knowledge tracing: A review, a tool and empirical studies. IEEE

Transactions on Knowledge and Data Engineering (2025). https://doi.org/10.1109/TKDE.2025.3552759

Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., Chen, Z., Hu, G.: Fuzzy cognitive diagnosis for modelling examinee performance. ACM Transactions on Intelligent Systems and Technology (TIST) **9**(4), 1–26 (2018). https://doi.org/10.1145/3168361

Nakagawa, H., Iwasawa, Y., Matsuo, Y.: Graph-based knowledge tracing: modeling student proficiency using graph neural network. In: IEEE/WIC/aCM International Conference on Web Intelligence, pp. 156–163 (2019). https://doi.org/10.1145/3350546.3352513

Ye, Y., Shan, Z.: Hgkt: Hypergraph-based knowledge tracing for learner performance prediction. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2023). https://doi.org/10.1109/IJCNN54540.2023.10191844 . IEEE

Xu, T., Zhao, Y., Zhang, J., Deng, J., Zhao, Z., Han, D., Wang, F., Wei, X.: Sqgkt: Student-question interaction graph-based knowledge tracing. Expert Systems with Applications, 129175 (2025). https://doi.org/10.1016/j.eswa.2025.129175

Yang, H., Qi, T., Li, J., Guo, L., Ren, M., Zhang, L., Wang, X.: A novel quantitative relationship neural network for explainable cognitive diagnosis model. Knowledge-Based Systems **250**, 109156 (2022). https://doi.org/10.1016/j.knosys.2022.109156

Cui, C., Yao, Y., Zhang, C., Ma, H., Ma, Y., Ren, Z., Zhang, C., Ko, J.: Dgekt: a dual graph ensemble learning method for knowledge tracing. ACM Transactions on Information Systems **42**(3), 1–24 (2024). https://doi.org/10.1145/3638350

Huo, Y., He, M., Tan, X., Chen, K.: Gcn and gat-based interpretable knowledge tracing model. Complex & Intelligent Systems **11**(7), 300 (2025). https://doi.org/10.1007/s40747-025-01921-w

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

Anderson, J.R.: Cognitive modelling and intelligent tutoring. Technical report, ERIC (1986)

Leighton, J., Gierl, M.: Cognitive Diagnostic Assessment for Education: Theory and Applications, (2007). https://doi.org/10.1017/CBO9780511611186

Guo, D., Xu, W., Ding, W., Yao, Y., Wang, X., Pedrycz, W., Qian, Y.: Concept-cognitive learning survey: Mining and fusing knowledge from data. Information Fusion **109**, 102426 (2024). https://doi.org/10.1016/j.inffus.2024.102426

Harris, D.: Comparison of 1-, 2-, and 3-parameter irt models. Educational measurement: issues and practice **8**(1), 35–41 (1989). https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Cai, L., Choi, K., Hansen, M., Harrell, L.: Item response theory. Annual Review of Statistics and Its Application **3**(1), 297–321 (2016). https://doi.org/10.1177/0011000099273004

Khan, A., Ghosh, S.K.: Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and information technologies **26**(1), 205–240 (2021). https://doi.org/10.1007/s10639-020-10230-3

Zwinderman, A.H.: A generalized rasch model for manifest predictors. Psychometrika **56**(4), 589–600 (1991). https://doi.org/10.1007/BF02294492

Christensen, K.B.: From rasch scores to regression. Journal of Applied Measurement **7**(2), 184 (2006)

Lord, F.M.: Applications of Item Response Theory to Practical Testing Problems, 1st edn. (2012). https://doi.org/10.4324/9780203056615

Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests, (1980). https://doi.org/10.1007/978-94-011-2732-5

Birnbaum, A.: Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores (1968). https://doi.org/http://dx.doi.org/

Reckase, M.D.: The past and future of multidimensional item response theory. Applied Psychological Measurement **21**(1), 25–36 (1997). https://doi.org/10.1177/0146621697211002

Bradshaw, L., Templin, J.: Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. Psychometrika **79**(3), 403–425 (2014). https://doi.org/10.1007/s11336-013-9350-4

Zhan, P., Jiao, H., Liao, D.: Cognitive diagnosis modelling incorporating item response times. British Journal of Mathematical and Statistical Psychology **71**(2), 262–286 (2018). https://doi.org/10.1111/bmsp.12114

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

Tsutsumi, E., Guo, Y., Kinoshita, R., Ueno, M.: Deep knowledge tracing incorporating a hypernetwork with independent student and item networks. IEEE Transactions on Learning Technologies **17**, 951–965 (2023). https://doi.org/10.1109/TLT.2023.3346671

Guo, L., Zhang, J., Ma, G., Dai, J.: A deep knowledge tracing model based on cognitive assimilation and item response theory with better interpretability. Applied Soft Computing **180**, 113331 (2025). https://doi.org/10.1016/j.asoc.2025.113331

Delen, D., Topuz, K., Eryarsoy, E.: Development of a bayesian belief network-based dss for predicting and understanding freshmen student attrition. European journal of operational research **281**(3), 575–587 (2020). https://doi.org/10.1016/j.ejor.2019.03.037

Bulut, O., Shin, J., Yildirim-Erbasli, S.N., Gorgun, G., Pardos, Z.A.: An introduction to bayesian knowledge tracing with pybkt. Psych **5**(3), 770–786 (2023). https://doi.org/10.3390/psych5030050

Liu, Z., Liu, Q., Chen, J., Huang, S., Gao, B., Luo, W., Weng, J.: Enhancing deep knowledge tracing with auxiliary tasks. In: Proceedings of the ACM Web Conference 2023, pp. 4178–4187 (2023). https://doi.org/10.1145/3543507.3583866

Wu, T., Ling, Q.: Fusing hybrid attentive network with self-supervised dual-channel heterogeneous graph for knowledge tracing. Expert Systems with Applications **225**, 120212 (2023). https://doi.org/10.1016/j.eswa.2023.120212

Ning, Q., Guo, C., Liu, K., Tang, J., Zhang, S., Zeng, W., Zhao, X.: Dual sequence modeling for knowledge tracing. Data Science and Engineering **10**(4), 653–664 (2025). https://doi.org/10.1007/s41019-025-00294-x

Pandey, S., Srivastava, J.: Rkt: relation-aware self-attention for knowledge tracing. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 1205–1214 (2020). https://doi.org/10.1145/3340531.3411994

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., Hu, G.: Ekt: Exercise-aware knowledge tracing for student performance prediction. IEEE Transactions on Knowledge and Data Engineering **33**(1), 100–115 (2019). https://doi.org/10.1109/TKDE.2019.2924374

Abdelrahman, G., Wang, Q.: Deep graph memory networks for forgetting-robust knowledge tracing. IEEE Transactions on Knowledge and Data Engineering **35**(8), 7844–7855 (2022). https://doi.org/10.1109/TKDE.2022.3206447

Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., Su, Y.: Neuralcd: A general framework for cognitive diagnosis. IEEE Transactions on Knowledge and Data Engineering **35**(8), 8312–8327 (2023). https://doi.org/10.1109/TKDE.2022.3201037

Park, S., Lee, D., Park, H.: Enhancing knowledge tracing with concept map and response disentanglement. Knowledge-Based Systems **302**, 112346 (2024). https://doi.org/10.1016/j.knosys.2024.112346

Li, Q., Yuan, X., Liu, S., Gao, L., Wei, T., Shen, X., Sun, J.: A genetic causal explainer for deep knowledge tracing. IEEE Transactions on Evolutionary Computation **28**(4), 861–875 (2023). https://doi.org/10.1109/TEVC.2023.3324461

Zhu, J., Ma, X., Huang, C.: Stable knowledge tracing using causal inference. IEEE Transactions on Learning Technologies **17**, 124–134 (2023). https://doi.org/10.1109/TLT.2023.3319760

Huang, C., Wei, H., Huang, Q., Jiang, F., Han, Z., Huang, X.: Learning consistent representations with temporal and causal enhancement for knowledge tracing. Expert Systems with Applications **245**, 123128 (2024). https://doi.org/10.1016/j.eswa.2023.123128

Lu, F., Li, Y., Bao, Y.: Deep knowledge tracing integrating temporal causal inference and pinn. Applied Sciences **15**(3), 1504 (2025). https://doi.org/10.3390/app15031504

Lu, Y., Chen, P., Pian, Y., Zheng, V.W.: Cmkt: Concept map driven knowledge tracing. IEEE Transactions on Learning Technologies **15**(4), 467–480 (2022). https://doi.org/10.1109/TLT.2022.3196355

Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: International Conference on Formal Concept Analysis. Lecture Notes in Computer Science, vol. 5548, pp. 314–339. Springer, ??? (2009). https://doi.org/10.1007/978-3-642-01815-2_23

Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Lecture Notes in Computer Science, vol. 1500 (1999). https://doi.org/10.1007/978-3-642-59830-2

Yao, Y.: Three-way granular computing, rough sets, and formal concept analysis. International Journal of Approximate Reasoning **116**, 106–125 (2020). https://doi.org/10.1016/j.ijar.2019.11.002

Troy, A.D., Zhang, G.-Q., Tian, Y.: Faster concept analysis. In: International Conference on Conceptual Structures, pp. 206–219 (2007). https://doi.org/10.1007/978-3-540-73681-3_16 . Springer

Hofer, M., Obraczka, D., Saeedi, A., Köpcke, H., Rahm, E.: Construction of knowledge graphs: Current state and challenges. Information **15**(8), 509 (2024). https://doi.org/10.3390/info15080509

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering **36**(7), 3580–3599 (2024). https://doi.org/10.1109/TKDE.2024.3352100

Wang, J., Xu, W., Ding, W., Qian, Y.: Multiview fuzzy concept-cognitive learning

with high-order information fusion of fuzzy attributes. IEEE Transactions on Fuzzy Systems **32**(12), 6965–6978 (2024). https://doi.org/10.1109/TFUZZ.2024.3470794

Liu, Z., Li, J., Zhang, X., Wang, X.-Z.: Incremental incomplete concept-cognitive learning model: A stochastic strategy. IEEE Transactions on Neural Networks and Learning Systems **36**(1), 393–406 (2023). https://doi.org/10.1109/tnnls.2023.3333537

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv (2017). https://doi.org/10.48550/arXiv.1706.03762

Appendix A

A.1 Joint Posterior Distribution

$$
\begin{aligned}
P(\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho} | \mathbf{Q}, \mathbf{R}) &= \frac{P(\mathbf{Q}, \mathbf{R} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}) P(\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho})}{P(\mathbf{Q}, \mathbf{R})} \\
&= \frac{P(\mathbf{R} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}, \mathbf{Q}) P(\mathbf{Q} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}) P(\mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho})}{P(\mathbf{Q}, \mathbf{R})} \\
&= \frac{P(\mathbf{R} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}, \mathbf{Q}) P(\mathbf{Q} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}) P(\mathbf{I}^{(0)}) P(\boldsymbol{\lambda}) P(\mathbf{d}) P(\mathbf{c}) P(\boldsymbol{\rho})}{P(\mathbf{Q}, \mathbf{R})} \\
&\propto P(\mathbf{R} | \mathbf{I}^{(0)}, \boldsymbol{\lambda}, \mathbf{d}, \mathbf{c}, \boldsymbol{\rho}, \mathbf{Q}) P(\mathbf{I}^{(0)}) P(\boldsymbol{\lambda}) P(\mathbf{d}) P(\mathbf{c}) P(\boldsymbol{\rho})
\end{aligned}
$$

A.2 Gradient Derivations

$$
\begin{aligned}
\nabla_{\lambda_e^{(i)}} &= \frac{\alpha_{\lambda_e^{(i)}} - 1}{\lambda_e^{(i)}} - \frac{\beta_{\lambda_e^{(i)}} - 1}{1 - \lambda_e^{(i)}} + \sum_{s \in S} \left[-\frac{(\eta_{se}^{(i)} - R_{se}^{(i)})}{\sigma^2_{\eta_{se}^{(i)}}} \frac{\partial \eta_{se}^{(i)}}{\partial \lambda_e^{(i)}} \right], \\
\nabla_{d_e^{(i)}} &= \frac{\alpha_{d_e^{(i)}} - 1}{d_e^{(i)}} - \frac{\beta_{d_e^{(i)}} - 1}{1 - d_e^{(i)}} + \sum_{s \in S} \left[-\frac{(\eta_{se}^{(i)} - R_{se}^{(i)})}{\sigma^2_{\eta_{se}^{(i)}}} \frac{\partial \eta_{se}^{(i)}}{\partial d_e^{(i)}} \right], \\
\nabla_{I_{sk}^{(0)}} &= \frac{\alpha_{I_{sk}^{(0)}} - 1}{I_{sk}^{(0)}} - \frac{\beta_{I_{sk}^{(0)}} - 1}{1 - I_{sk}^{(0)}} + \sum_{i=1}^{N} \sum_{e \in E_i} \left[-\frac{(\eta_{se}^{(i)} - R_{se}^{(i)})}{\sigma^2_{\eta_{se}^{(i)}}} \frac{\partial \eta_{se}^{(i)}}{\partial I_{sk}^{(0)}} \right], \\
\nabla_{c_{sk}^{(i)}} &= \frac{\alpha_{c_{sk}^{(i)}} - 1}{c_{sk}^{(i)}} - \frac{\beta_{c_{sk}^{(i)}} - 1}{1 - c_{sk}^{(i)}} + \sum_{j=i+1}^{N} \sum_{e \in E_i} \left[-\frac{(\eta_{se}^{(j)} - R_{se}^{(j)})}{\sigma^2_{\eta_{se}^{(j)}}} \frac{\partial \eta_{se}^{(j)}}{\partial c_{sk}^{(i)}} \right], \\
\nabla_{\rho_{sk}} &= \frac{\alpha_{\rho_{sk}} - 1}{\rho_{sk}} - \frac{\beta_{\rho_{sk}} - 1}{1 - \rho_{sk}} + \sum_{i=1}^{N} \sum_{e \in E_i} \left[-\frac{(\eta_{se}^{(i)} - R_{se}^{(i)})}{\sigma^2_{\eta_{se}^{(i)}}} \frac{\partial \eta_{se}^{(i)}}{\partial \rho_{sk}} \right].
\end{aligned}
$$

$$
\mathcal{E}_{sk}^{(i)} = \prod_{j=0}^{i-1} (c_{sk}^{(j)} + f_{sk}^{(j)}) = \mathcal{E}_{sk}^{(i-1)} (c_{sk}^{(i-1)} + f_{sk}^{(i-1)}).
$$

$$
\begin{aligned}
\frac{\partial \eta_{se}^{(i)}}{\partial \lambda_e^{(i)}} &= D(\theta_{se}^{(i)} - d_e^{(i)}) \eta_{se}^{(i)} (\eta_{se}^{(i)} - 1), \\
\frac{\partial \eta_{se}^{(i)}}{\partial d_e^{(i)}} &= -D \lambda_e^{(i)} \eta_{se}^{(i)} (\eta_{se}^{(i)} - 1), \\
\frac{\partial \eta_{se}^{(i)}}{\partial I_{sk}^{(0)}} &= D \lambda_e^{(i)} \tilde{q}_{ek}^{(i)} \eta_{se}^{(i)} (\eta_{se}^{(i)} - 1) \prod_{j=0}^{i-1} (c_{sk}^{(j)} + f_s^{(j)}), \\
\frac{\partial \eta_{se}^{(j)}}{\partial c_{sk}^{(i)}} &= D \lambda_e^{(j)} I_{sk}^{(0)} \tilde{q}_{ek}^{(j)} \eta_{se}^{(j)} (\eta_{se}^{(j)} - 1) \frac{\mathcal{E}_{sk}^{(j)}}{c_{sk}^{(i)} + f_s^{(i)}} \quad (i < j), \\
\frac{\partial \eta_{se}^{(i)}}{\partial \rho_{sk}} &= D \lambda_e^{(i)} \eta_{se}^{(i)} (\eta_{se}^{(i)} - 1) \tilde{q}_{ek}^{(i)} I_{sk}^{(0)} \sum_{j=0}^{i-1} \left[-\Delta t^{(j)} f_{sk}^{(j)} \frac{\mathcal{E}_{sk}^{(i)}}{c_{sk}^{(j)} + f_{sk}^{(j)}} \right].
\end{aligned}
$$

Appendix B

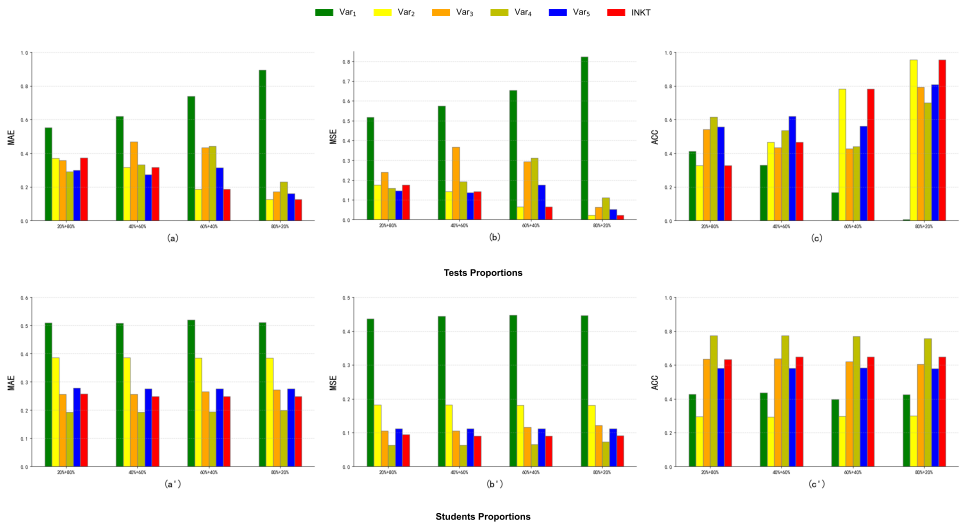


Fig. B1 Ablation experiment results under different tests/students proportions on Math1

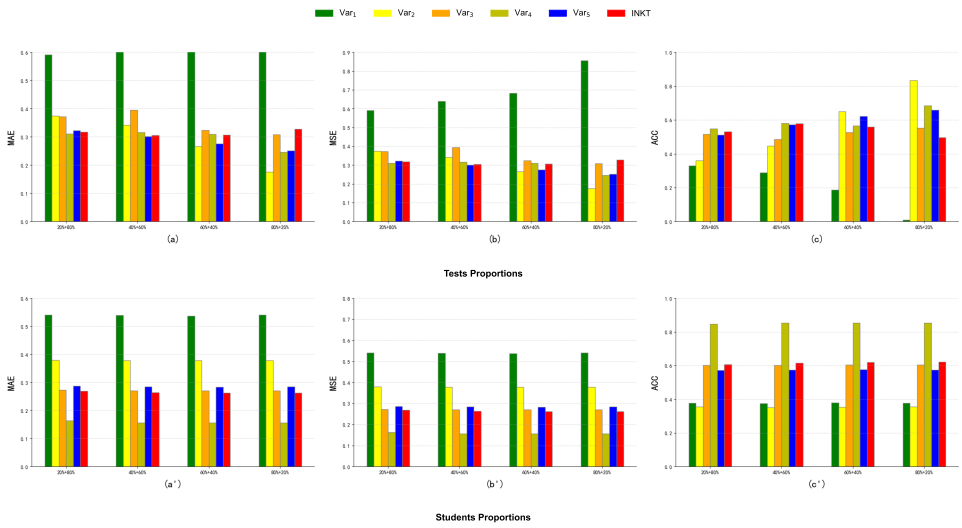


Fig. B2 Ablation experiment results under different tests/students proportions on Math2

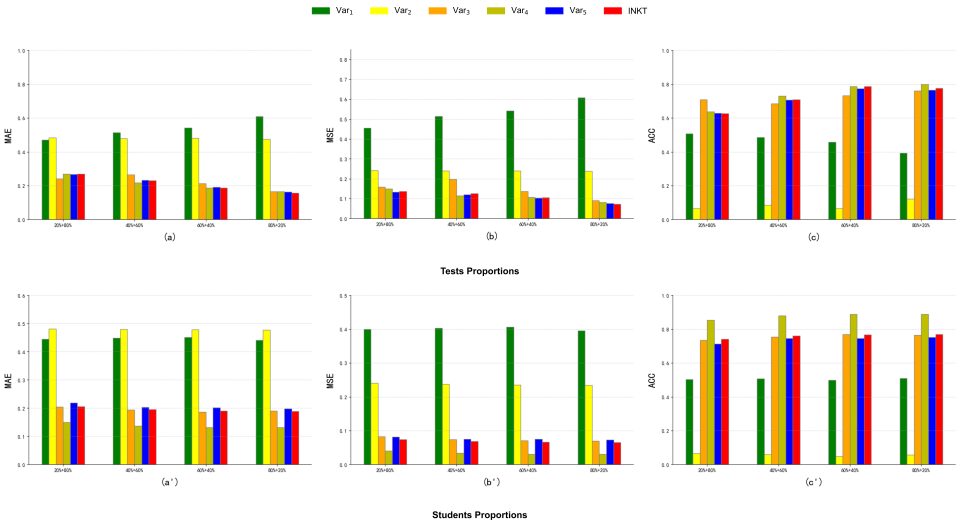


Fig. B3 Ablation experiment results under different tests/students proportions on FrcSub

Table B1 Ablation experiment results on the dataset Math1

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.4116	0.5182	0.5535	0.3290	0.5739	0.6201	0.1683	0.6339	0.3785	0.0076	0.8228	0.8940
Var_2	0.3286	0.1749	0.3721	0.4670	0.1430	0.3173	0.7824	0.0646	0.1864	0.9551	0.0234	0.1275
Var_3	0.6156	0.1598	0.2915	0.5346	0.1918	0.3313	0.4408	0.3116	0.4417	0.7004	0.1106	0.2315
Var_4	**0.6813**	**0.1025**	**0.2469**	**0.7008**	**0.0978**	**0.2370**	0.7157	0.0975	0.2291	0.7016	0.1006	0.2289
Var_5	0.5416	0.2399	0.3584	0.4334	0.3677	0.4679	0.4280	0.2939	0.4340	0.7928	0.0628	0.1728
INKT	0.3286	0.1758	0.3731	0.4670	0.1430	0.3173	**0.7824**	**0.0646**	**0.1864**	**0.9552**	**0.0234**	**0.1275**

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.4283	0.4362	0.5095	0.4350	0.4439	0.5086	0.3971	0.4469	0.5195	0.4251	0.4461	0.5110
Var_2	0.2960	0.1822	0.3860	0.2941	0.1819	0.3857	0.2969	0.1817	0.3854	0.2989	0.1817	0.3852
Var_3	**0.7729**	**0.0636**	**0.1931**	**0.7744**	**0.0632**	**0.1926**	**0.7693**	**0.0650**	**0.1944**	**0.7557**	**0.0727**	**0.1987**
Var_4	0.6599	0.0964	0.2493	0.6847	0.0897	0.2391	0.6947	0.0868	0.2337	0.6978	0.0893	0.2344
Var_5	0.6360	0.1051	0.2570	0.6375	0.1055	0.2566	0.6211	0.1162	0.2651	0.6057	0.1214	0.2715
INKT	0.6331	0.0945	0.2572	0.6476	0.0904	0.2484	0.6473	0.0905	0.2484	0.6478	0.0910	0.2483

Table B2 Ablation experiment results on the dataset Math2

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.3290	0.5460	0.5941	0.2894	0.5971	0.6404	0.1878	0.6041	0.6834	0.0103	0.7781	0.8563
Var_2	0.3609	0.1803	0.3744	0.4475	0.1611	0.3415	**0.6506**	**0.1154**	**0.2662**	**0.8342**	**0.0448**	**0.1764**
Var_3	**0.5495**	0.1670	**0.3111**	**0.5805**	0.1859	0.3163	0.5663	0.1772	0.3099	0.6857	0.1263	0.2459
Var_4	0.5119	0.1585	0.3227	0.5734	**0.1545**	**0.3017**	0.6222	0.1325	0.2758	0.6582	0.1200	0.2515
Var_5	0.5166	0.2580	0.3724	0.4850	0.2758	0.3948	0.5277	0.1701	0.3239	0.5524	0.1412	0.3083
INKT	0.5312	**0.1600**	0.3176	0.5784	0.1618	0.3053	0.5596	0.1531	0.3066	0.4964	0.1475	0.3282

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.3766	0.4720	0.5405	0.3761	0.4711	0.5398	0.3808	0.4689	0.5373	0.3767	0.4724	0.5405
Var_2	0.3558	0.1832	0.3793	0.3511	0.1827	0.3789	0.3540	0.1829	0.3788	0.3560	0.1830	0.3790
Var_3	**0.8463**	**0.0398**	**0.1636**	**0.8537**	**0.0379**	**0.1564**	**0.8539**	**0.0377**	**0.1561**	**0.8539**	**0.0378**	**0.1562**
Var_4	0.5721	0.1175	0.2869	0.5746	0.1165	0.2848	0.5761	0.1159	0.2836	0.5740	0.1168	0.2850
Var_5	0.6026	0.1070	0.2727	0.6033	0.1063	0.2712	0.6050	0.1059	0.2703	0.6055	0.1063	0.2705
INKT	0.6063	0.1050	0.2692	0.6164	0.1022	0.2636	0.6199	0.1019	0.2625	0.6212	0.1024	0.2631

Table B3 Ablation experiment results on the dataset FrcSub

(1) Evaluation metrics under different tests proportions

	3+4			4+3			4+2			6+1		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.5069	0.4557	0.4703	0.4866	0.5134	0.5133	0.4583	0.5417	0.5416	0.3917	0.6082	0.6082
Var_2	0.0665	0.2413	0.4830	0.0858	0.2401	0.4792	0.0662	0.2409	0.4809	0.1216	0.2383	0.4757
Var_3	0.6367	0.1492	0.2692	**0.7297**	**0.1143**	**0.2165**	**0.7875**	0.1071	**0.1862**	**0.7988**	0.0820	0.1662
Var_4	0.6289	**0.1330**	**0.2667**	0.7056	0.1196	0.2323	0.7741	**0.1037**	0.1920	0.7660	0.0757	0.1625
Var_5	**0.7093**	0.1599	0.2413	0.6846	0.1968	0.2649	0.7317	0.1368	0.2126	0.7615	0.0901	0.1664
INKT	0.6257	0.1371	0.2692	0.7078	0.1262	0.2313	0.7861	0.1048	0.1878	0.7753	**0.0726**	**0.1565**

(2) Evaluation metrics under different students proportions

	20%+80%			40%+60%			60%+40%			80%+20%		
Model	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓	ACC↑	MSE↓	MAE↓
Var_1	0.5037	0.3994	0.4452	0.5081	0.4027	0.4480	0.4991	0.4060	0.4518	0.5104	0.3953	0.4408
Var_2	0.0661	0.2404	0.4808	0.0590	0.2376	0.4798	0.0494	0.2350	0.4784	0.0574	0.2340	0.4772
Var_3	**0.8529**	**0.0401**	**0.1499**	**0.8795**	**0.0335**	**0.1371**	**0.8878**	**0.0312**	**0.1323**	**0.8881**	**0.0311**	**0.1320**
Var_4	0.7140	0.0816	0.2182	0.7448	0.0754	0.2026	0.7450	0.0750	0.2020	0.7514	0.0726	0.1981
Var_5	0.7337	0.0823	0.2050	0.7535	0.0744	0.1943	0.7684	0.0703	0.1869	0.7657	0.0699	0.1896
INKT	0.7413	0.0744	0.2055	0.7599	0.0691	0.1949	0.7674	0.0666	0.1904	0.7692	0.0658	0.1890