Deep Knowledge Tracing Based on Behaviour in the Item Response Theory Framework

by

Chaojun Ma

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

in
Measurement, Evaluation, and Data Science

Department of Educational Psychology
University of Alberta

Abstract

Knowledge tracing involves assessing students' performance based on their learning interaction records and providing them with personalized learning paths. Deep learning methods model students' knowledge states by exploring extensive exercise records, with the Deep-IRT approach considered superior to others. However, Deep-IRT overlooks relevant behavioural features that could assist in modelling students' knowledge states. Therefore, this study introduces a new method for modelling student behaviour, combining behavioural features with learning records to enhance knowledge tracking performance. Experimental results on two real-world benchmark datasets indicate that the proposed model significantly outperforms the Deep-IRT model in predicting future students' abilities. Limitations of the current study and potential directions for future research are also discussed.

Preface

This thesis is an original work by Chaojun Ma. No part of this thesis has been previously published.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr.Ying Cui. Your support and guidance were indispensable in completing the thesis research. Your patient mentorship not only contributed to the development of my programming skills but also ignited my interest in the fields of quantitative psychology and deep learning.

Additionally, I want to thank all my friends who have consistently provided encouragement and companionship. Special thanks to my parents for their unwavering love and support throughout this journey.

I would also like to express my heartfelt appreciation to my cat, 'Duo Duo.' It was during my most challenging times that 'Duo Duo' stood by me, providing companionship and comfort.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

As a result of the development of computing and information technologies, intelligent tutoring systems (ITS) are now being widely implemented, making online education possible on a massive scale. During the pandemic, a large number of schools were obliged to deliver learning content through online ITS platforms (Abdelrahman et al., 2023; Ni et al., 2022). Unlike traditional online learning platforms, ITS can personalize effective learning paths for students and track their current knowledge mastery based on their answers, thereby enabling tailored instruction (Feng et al., 2009). However, these technology-enhanced learning environments bring with them not only a range of advantages but also challenges (Abdelrahman et al., 2023). One of the predominant challenges encountered in implementing ITS involves evaluating and representing the student's level of knowledge as this relies on potentially uncertain data (AlShaikh & Hewahi, 2021). As a result, there is an urgent need to use more effective scientific methods for targeted analysis and exploration (Feng et al., 2009). This has led to an increasing number of researchers attempting to leverage the rich data available on ITS to model the students' learning behaviours through machine learning or deep learning algorithms. Appropriate models should be able to predict a student's future performance based on analysis of their behaviour and past assessment results (Liu, 2022). Knowledge tracing (KT) is a particularly effective way of accomplishing this, so it has become an integral component of ITS (Fanzhi et al., 2022). KT uses educational data mining for the purposes of knowledge concepts analysis. This, in turn, depends upon computer-assisted large-scale data processing capabilities to track students' learning activities (Ni et al., 2022).

Background

The notion of "knowledge tracing" was first introduced by Corbett and Anderson in 1995. Since then, it has evolved into a well-established field of research in educational artificial

intelligence (Liu, 2022). There are two popular approaches: traditional knowledge tracing; and deep knowledge tracing. Early KT models focused on using student-related learning parameters to predict their performance. At this stage, the most popular algorithm was the Bayesian network tracing (BKT) (Pelánek, 2017). BKT employs probabilistic graphical theory, which considers a student's knowledge state as a latent variable and models it to track changes in the student's level of knowledge (Abdelrahman et al., 2023). However, researchers started to realize that other external factors could affect students' knowledge acquisition, such as an items' difficulty or forgetting. These factors are difficult to be incorporated into BKT (Liu, 2022). In 2000, a logistic model was proposed that offered better levels of interpretability and that was capable of including complex features for the purposes of knowledge-tracing tasks (Pavlik et al., 2021). Logistic models have attracted an increasing amount of attention and various variants have now been proposed (Cen et al., 2006; Deonovic et al., 2018). BKT and logistic models can both be considered traditional statistical approaches because they rely on statistical machine learning algorithms and student interaction data for the modelling process. However, statistical approaches struggle to model concepts with complex relationships and often necessitate the modelling of each concept individually before any relationships can be established. This consumes enormous amounts of time and resources (Song et al., 2022).

A deep knowledge tracking (DKT) model was introduced by Piech et al. (2015). They suggested applying recurrent neural networks (RNNs) to knowledge tracing because RNNs can convert raw data into high-dimensional and continuous representations, making them well-suited to complex data structures (Piech et al., 2015). In contrast to the traditional approaches, especially BKT, deep learning-based KT models can draw upon students' temporal interactions and consider the relationships between different knowledge components to build models more efficiently (Liu,

2022). A number of researchers recognized the promise in Piech et al.'s work, which has led to the development of a range of deep learning-based KT models. Some of the examples include dynamic key-value memory networks (DKVMN) (Yeung, 2019; He et al., 2021; Xiao et al., 2022), which are based on memory networks; Graph-based knowledge tracing (GKT) (Nakagawa et al., 2019; Abdelrahman & Wang, 2021), which is based on graph theory; and Exercise-enhanced recurrent neural networks (EERNNA) (Su et al., 2018; Shin et al., 2021), which are based on attention mechanisms. A number of deep learning-based KT models have now been deployed in the real world to provide services to students (Li et al., 2023). Of particular interest, is a novel deep knowledge tracing model called Deep-IRT (Yeung, 2019), which incorporates aspects of item response theory (IRT) into the DKVMN model. This helps, in particular, to improve the model's interpretability. Unfortunately, Deep-IRT and its variants do not give sufficient consideration to the potential impact of student behaviour on their model, which can have an important negative influence on its accuracy (Xu et al., 2021). This thesis looks at a specific way in which the performance of Deep-IRT might be enhanced.

Purpose of Current Study

There are several ways in which students' behaviour can be introduced into the Deep-IRT model. In previous studies, researchers have attempted to improve the model's performance by directly introducing student behaviour data as additional input (Xiao et al., 2022; Wang et al., 2021). However, this method is open to being influenced by data sparsity, which can lead to inconsistent prediction results (Sarsa et al., 2022). Some prior research has also sought to formalize the relationship between student behaviour and deep models (He et al., 2021), which opens up the possibility of incorporating student behaviour data into Deep-IRT. However, no study has yet attempted to actually enact this proposition by using independent neural network layers to pre-

process the student behaviour data. This study, therefore, aimed to construct an enhanced version of the original Deep-IRT model that is based upon educational psychology theory (He et al., 2021). It also examined a way of reshaping the difficulty and ability network to improve its overall performance. The performance of the new modified the Deep-IRT model was investigated and compared with baseline models under a variety of conditions.

The core research questions this study set out to address were:

1) Does introducing student behaviour information via independent neural networks improve existing Deep-IRT model performance?
2) Does the proposed algorithm have better predictive performance than Deep-IRT and DKVMN models?

Chapter 2: Literature Review

In this chapter, to situate this thesis work on developing an enhanced version of Deep-IRT, this thesis goes into greater detail regarding the development of knowledge tracing. This chapter begins by giving an overview of the idea of knowledge tracing, and then moving on to various traditional approaches that have been developed in the past, including Bayesian and logistic models. Finally, this chapter then reviews the use of deep learning, as well as various approaches that have sought to fuse traditional methods and deep learning.

An overview of knowledge tracing

KT can be formulated as a supervised sequential learning problem, where the goal is to predict the probability of a student giving a correct response to the next exercise, given their historical sequence of interaction (Zou et al., 2020). Figure 1 shows a typical student learning scenario to illustrate how KT is applied. The student is given a sequence of questions $(q_1, q_2, q_3 \ldots q_t)$ from a question set, Q. The student is required to respond to each question separately, $(r_1, r_2, r_3 \ldots r_t)$. Each question, q, is associated with one or multiple knowledge concepts, denoted as k (e.g., equality and equations). Importantly, these knowledge concepts are not independent of each other. For instance, to master the concept of linear equations (k_3), one must first have an understanding of equations (k_2) and equality (k_1). In that case, the interdependency between knowledge points is an important part of what needs to be considered in KT tasks (Abdelrahman et al., 2023). The student's behaviour might also be recorded during the interaction (e.g., their response time, number of hints, engagement, etc.). On the basis of the above information, KT predicts the student's future performance and estimates the level of concepts mastered by the student. Beyond this, estimates of the current state of the student's knowledge can help build learning strategies that will optimize student learning efficiency (Liu et al., 2023).

However, there are several challenges to overcome when seeking to capture the state of a student's knowledge. These include: (1) the fact that each question can require the mastery of more than one concept, which increases the complexity of tracking their knowledge state (Song et al., 2022); (2) students can be forgetful, impacting the estimated results (Abdelrahman et al., 2023).

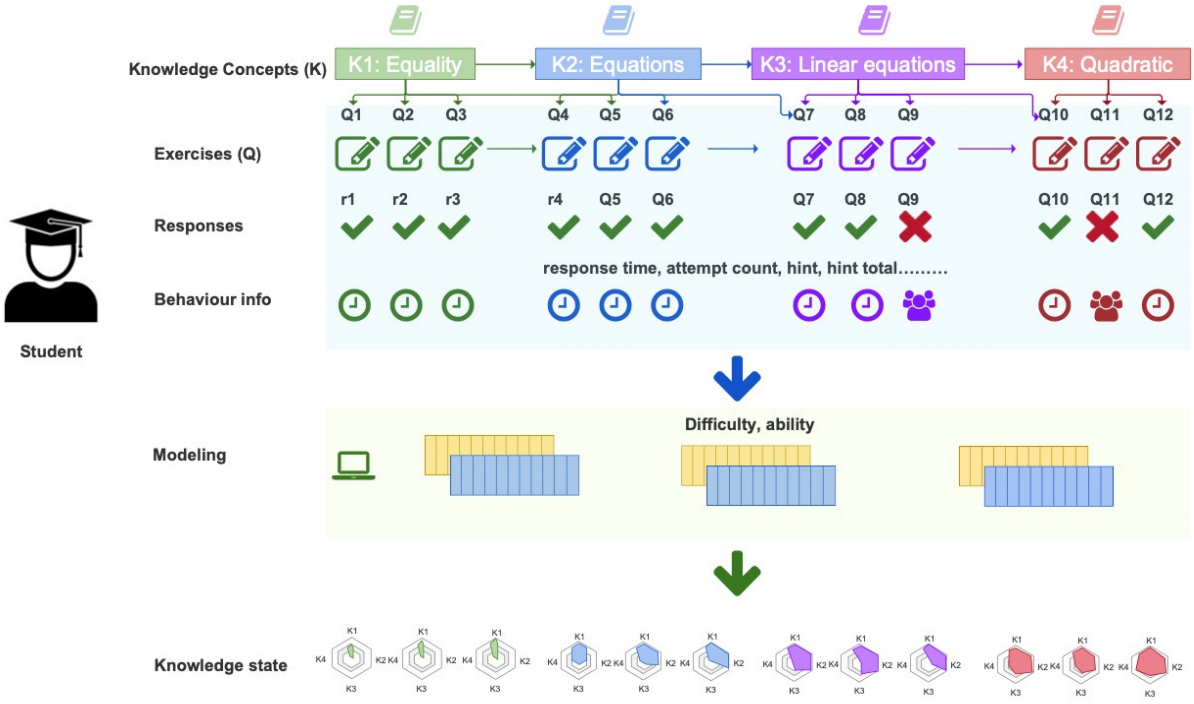

Figure 1. An overview of knowledge tracing.

Original Approaches

The first use of 'Knowledge tracing' (KT) as a concept was in a publication by Anderson et al. in the *Artificial Intelligence* journal in 1990 (Anderson et al., 1990). Corbett & Anderson (1995) subsequently gave a specific definition of KT, which is to track and estimate students' knowledge states through the use of machine learning algorithms. This was followed by their first outline of how to apply Bayesian network to the field of KT. Bayesian Knowledge Tracing (BKT) simulates a student's knowledge level, models the current status of their knowledge, and

establishes the potential transitions between subsequent different knowledge states (Fenzhi et al., 2022). BKT models use probabilistic theory to model students' learning trajectories, for instance by means of hidden Markovs (Abdelrahman et al., 2023). Hidden Markov models treat a student's knowledge state as a hidden variable and seek to infer their level of knowledge mastery from their responses (i.e., correct, or incorrect) when answering questions (Ma et al., 2022). BKT models adhere to three critical assumptions: (1) all students have the same background knowledge; (2) each question involves just one concept, with each concept being independent; (3) forgetting phenomena will never occur after a student has mastered a concept (Corbett & Anderson, 1995). Let P_{L0} denote a student's level of knowledge mastery level and P_T its prior probability. The student's updated state of knowledge mastery will then be:

$$P(L_{n+1}) = P(L_n|obs) + \big(1 - P(L_n|obs)\big) * P(T) \quad (1)$$

where $P(L_{n+1})$ is the sum of two probabilities: (1) that the knowledge concept (KC) has already been mastered; (2) that the current knowledge state can be converted to mastery (Fanzhi et al., 2022). Although BKT primarily relies on a particular knowledge concept to infer a student's learning performance, a hidden state can represent the knowledge concept. This is the way in which BKT can be seen to inform later DKT models (Fanzhi et al., 2022)

Unfortunately, BKT assumes that all students have the same level of prior knowledge and that each question only involves one skill (Abdelrhman et al., 2023). It also struggles to capture the interdependency between knowledge concepts. To address these issues, an individualized BKT model (Lee & Brunskill, 2012) and a dynamic BKT model (Kaser et al., 2014) have been proposed. Individualized BKT takes into account the fact that the standard BKT model may underestimate the learning performance of students who are above average level and overestimate the learning performance of students who are below average. Individualized BKT therefore introduces two

separate parameters into the model to correct the biases present in standard BKT, namely the student's initial knowledge mastery state and learning ability (Lee & Brunskill, 2012). Dynamic BKT, however, jointly models dependencies between multiple skills and different skill levels. Its goal is to capture a hierarchical structure of prerequisite skills within a single model. For example, if a certain skill is a prerequisite for mastering another, then the latter skill is treated as being conditionally dependent on the former skill (Kaser et al., 2014).

Later studies in this area focused on practical problems, such as introducing emotional states to improve the prediction performance (Spaulding & Breazeal, 2015) or using a dynamic learning rate to capture a student's improvement (Agarwal et al., 2020). BKT and its variations are simple to construct and offer strong interpretability. However, the relevant models mainly rely on knowledge states and cannot capture rich data features such as student's behaviour limiting their application in practical scenarios (Fenzhi et al., 2022).

Another method for tracing students' knowledge states is through logistic models, which can simplify KT tasks by calculating the probability of students correctly answering the exercises. The key idea is that the probability of correctly answering an exercise can be represented by a mathematical function of student behaviour and certain knowledge state parameters (Liu, 2022). On the basis of this, researchers have proposed using KT-related logistic functions to model students and predict answers where student performance is a dependent variable and a set of learning model parameters is derived from historical data (Fenzhi et al., 2022). One of the most famous logistic models was built on Item Response Theory (IRT). IRT is the basis of the modern psychometric theory. The 'item' in this case, relates to a question in a test, so 'item response' is a student's correct (or incorrect) answer to a specific question. IRT is a collection of statistical models developed for analyzing responses to tests, surveys, and similar instruments. These models

have enabled researchers and measurement experts to gain insights into how well individual examinees perform in specific areas of evaluation (Finch & French, 2019). IRT uses a 1-PL (parameter) model, a 2-PL model, or a 3-PL model, depending on the number of parameters involved. As a 1-PL model, the Rasch model is the simplest, and this simplicity has served to attract the attention of KT researchers (Abdelrhman et al., 2023). According to the Rasch model, given a student's potential ability (θ) and the item difficulty level (β), the probability that the student will give a correct response to given item is:

$$P\left(x_{ji}=1\middle|\theta_i,\beta_j\right)=\frac{e^{\theta_i}-\beta_j}{1+e^{\theta_i-\beta_j}} \tag{2}$$

where, x_{ji} indicates the response to item j by individual i; b_j indicates the difficulty of item j; and θ_i indicates the level of the individual's ability. In other words, if the student demonstrates high levels of ability in relation to an item, the correct response probability will also be high. If the item's difficulty level exceeds the students' expectations, they will be more likely to give an incorrect answer. Unfortunately, even though the Rasch model can predict student performance, item difficulty, and student ability, it has inbuilt assumptions that limit its applicability (Abdelrahman et al., 2023). The Rasch model assumes all items have the same discrimination parameter, implying that all items contribute equally to measuring the KC. In reality, this assumption may not hold for all items. Therefore, many researchers have drawn more fully upon IRT with the goal of forging a deeper connection between KT and IRT (Liu, 2022).

IRT places emphasis on students' latent abilities. The characteristics of the factors in a student's learning interactions (e.g., attempt count) can play a significant role in predicting their future learning performance (Embretson & Reise, 2013). In one study, for instance, Wilson et al. (2016) found that a Bayesian model using IRT-based extensions was capable of outperforming DKT. This model considered both time-related features and learning factors. Later, researchers

turned to using IRT to improve the interpretability of deep learning KT models (Yeung, 2019) or combined deep learning algorithms with IRT to estimate a student's potential ability more accurately (Tsutsumi et al., 2021).

Following on from the initial work on IRT, Learning Factors Analysis (LFA) emerged as a way of addressing the known IRT drawbacks of one item only containing one concept and it treats a student's learning ability as unchangeable (Fanzhi et al., 2022). Here, the knowledge tracing machine (KTM) proposed by Vie and Hisashi exhibits the best performance (Vie & Kashima, 2018; Abdelrahman et al., 2023). KTM uses factorization machines algorithms to model a student's performance, which it can accomplish rapidly even when confronted with sparse data. However, KTM relies on the repeated learning of latent concepts and, for some concepts that are not frequently practiced, its prediction performance can be no better than that of IRT (Fanzhi et al., 2022)

To sum up, although logistic regression is more robust than BKT, it ignores a great deal of student information during the modelling process, especially response time and the forgetting rate (Song et al., 2022). This has prevented logistic regression models from entirely replacing BKT (Fanzhi et al, 2022).

Deep Learning-based Approaches

As data processing rates have improved, deep learning, which relies on big data and complex models, has started to attract the attention of researchers. Most deep learning models for knowledge tracing are based on BKT or logistic models, which are then combined with neural networks to predict student performance (Fanzhi et al., 2022). The deep learning models can capture more intricate representations of student knowledge and uncover and utilize information about the interrelationships between concepts (Song et al., 2022). They are also able to handle the

fact that a student's current knowledge state can be influenced by the progression of their knowledge state over time. This makes them more realistic than BKT, where a student's knowledge state is solely modelled upon their final response (Xiao et al., 2022). The first application of recurrent neural networks (RNNs) to knowledge tracing tasks was by Piech et al. in 2005. This is now regarded as the seminal work in Deep learning-based knowledge Tracing.

Inspired by the Markov process present in BKT models and recent developments in RNNs, Piech et al. (2005) proposed an innovative new model for knowledge tracing called Deep knowledge Tracing model (DKT) that relies exclusively upon deep learning. The DKT are deep-learning models that can capture sequential dynamics through recurrent connections (Zhang et al., 2021). This property can reflect the proximate causality effect in learning science and facilitate the retention of information relating to learning trajectories (Song et al., 2022). To that end, Piech et al. (2005) converted raw observations of student learning into a fixed-learning vector $(x_1, x_2 \ldots x_t)$. Through one-shot encoding, $\mathbf{x}_t$ can be transformed into an input vector. Then, using the hidden state, $\mathbf{h}_t$, in an RNN to represent a student's knowledge state, $\mathbf{h}_t$ can be passed through a sigmoid-activated linear layer to obtain an ultimate predicted result, $\mathbf{y}_t$. For example, in the most common instantiations of KT, an interaction x_t can be formed as a tuple of (q_t, a_t). q_t represents the question and a_t is the student's response ($a_t \in \{0,1\}$). So, q_1 could be a question about square root problems (e.g., where the problem number is 100) where a student answers correctly. In that case, x_t is (100, 1). If q_2 is a question about linear intercept problems (where the problem number is 101) and the answer is incorrect, x_t will be (101, 0). As these student learning interactions are discrete and do not have numerical significance (e.g., using 100 to represent square root problems and 101 to represent the intercept problems), to represent these features reasonably it is necessary to encode these categorical features in a different way. One-hot encoding is a commonly used data

pre-processing method for this kind of situation. It is based on there being N category values. It then constructs an N-dimensional vector of zeros, where each dimension represents a category feature value. The corresponding position in the N-dimensional vector is set to 1 for a specific category feature value (Ai, 2019).

Each element represents the predicted probability of a student answering a question correctly for a corresponding KC (Song et al., 2022). At each time step, t, the model calculates, $\mathbf{h}_t$ and $\mathbf{y}_t$ as follows:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{hx}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h) \quad (3)$$

$$\mathbf{y}_t = Sigmoid(\mathbf{W}_{yh}\mathbf{h}_t + \mathbf{b}_y) \quad (4)$$

In term of performance prediction, DKT models outperform BKT model by twenty percent, even without any human expert annotated data (Liu, 2022; Song et al., 2022). Overall, they have perhaps shown the best applicability to time-series-related tasks (Liu et al., 2023). However, DKT has several limitations: (1) it uses uninterpretable hidden states to represent a student's knowledge state, which makes it impossible to determine a student's level of mastery (Abdelrahman & Wang, 2019; Zou et al., 2020); (2) it focuses on student responses and sets aside other rich learning trajectory information (such as number of students attempts or elapsed time) (Liu, 2022). So, critiques of DKT point to its poor capacity for interpretation and limited learning features.

Researchers have tried a number of approaches to overcome the problems associated with DKT, such as creating new structure, redesigning the loss functions, and introducing attention mechanisms (Liu et al., 2023). Out of these, the most effective way is to extend the model's input features by designing new structures. Memory networks, for instance, mimic human memory mechanisms and forgetting phenomena. Neural networks that introduce external memory into their structure are generally referred to as Memory Augmented Neural Networks (MANNs). A crucial

characteristic of MANNs is their ability to achieve good classification prediction even when confronted with sparse data (Santoro et al., 2016). Inspired by a MANN framework proposed by Graves et al. (2016), Zhang et al. (2017) improved the DKT model by introducing an external memory structure that moves beyond basic MANNs by using dynamic rather than static memory to capture temporal dependencies and evolving patterns in a student's learning process (Zhang et al., 2017). This improved model was called a Dynamic Key-Value Memory Network (DKVMN). DKVMN is now widely used for KT tasks and a read-write mechanism has now been added to predict a student's performance and update the interactions. DKVMN assumes there are N latent traits underlying a sequence of exercises. These are stored in a *concept matrix*. The corresponding latent traits and student's mastery levels at timestamp t are stored in an *ability matrix* (Zhang et al., 2017). The reading mechanism follows two steps to predict student performance: attention; and reading. The attention step embeds an input feature and multiplies it with the concept matrix to calculate the number of latent traits involved in a particular question. After that, the results are passed through ability matrix to give a weighted sum and through the fully connected layer to get the relative weight, at which point the process finishes with the reading step. According to a student's response to question q_t, the writing process can be seen to transform the student's latent level of mastery from state V_{t-1} at time $t-1$ to V_t at time t, with the ability matrix then being updated (Zou et al., 2020).

Although DKVMN remains the most popular memory network, some other approaches have been proposed. Graph networks, for instance, offer another way of enhancing the input features. Here, the focus is on knowledge concept relationships, such as their similarity and the correspondence between exercises and existing concepts. Nakagawa et al. (2019) were the first to suggest Graph-based Knowledge Tracing (GKT). They redefined the knowledge tracing problem

as a time series node-level classification problem (Abdelrahman et al., 2023). GKT allows researchers to get a better understanding of students' interactions and knowledge (Liu, 2022)

Aside from this, Su et al. (2018) have argued for the extraction of textual features from exercises by using an attention-based exercise-enhanced recurrent neural network. This models the student learning process by embedding textural features within it so as to obtain exercise representations. This was the first attempt to introduce attention mechanisms into knowledge tracing (Fanzhi et al., 2022). Pandey & Karypis (2019), meanwhile, have applied a Transformers model to KT tasks and have introduced the use of NLP algorithms to knowledge tracing (Xu et al., 2021). To tackle the problem of interpretability, some researchers have tried to visualize and model student behaviour in higher dimensions and thereby exhibit the interaction between different knowledge concepts (e.g., Ding & Larson, 2021).

While existing graph-based algorithms or attention mechanisms can enhance the performance of DKT (Fanzhi et al., 2022), DKVMN is still currently considered the best knowledge-tracing model due to its lower probability of overfitting and its capacity to automatically discover correlations between exercise questions and concepts (Sun et al., 2021; Zou et al., 2020). Many researchers have therefore sought to further improve and build upon DKVMN. This includes redesigning the structure or introducing new modules. Ai et al. (2019) were the first to redesign the DKVMN model. They did this by treating questions as a set of concept tags, which are usually available in online ITS platforms. Concept tags are typically used to label each question or item with the relevant knowledge points they involve. This allows questions to be associated with their respective knowledge points. The suggestion was that question concept tags could then be used to modify the concept matrix structure and improve the performance of a tutoring system. Other research has attempted to extract textual features from the questions themselves. One such

approach applied a multi-head self-attention mechanism to question texts and combined this with other behavioural features to redesign the DKVMN model (Xiao et al., 2022). Some researchers have argued that the input features in the DKVMN model are limited and do not make sufficient use of the rich information available on online education platforms. In relation to this, Sun et al. (2021) have proposed a versatile knowledge-tracing algorithm that combines learning ability and behavioural features. This algorithm passes information about the number of attempts made by a student and the number of hints requested through an Xgboost layer to obtain a preliminary prediction, which then serves as part of a DKVMN model's input. In a distinct approach, Yeung (2019) attempted to incorporate IRT-related modules to enhance the performance of the model. Using difficulty and ability neural network layers can lead to more meaningful estimates and increase a network model's interpretability while preserving the DKVMN model's predictive capability. However, Yeung's model lacks theoretical support. Although it does seem to improve the performance and interpretability of DKVMN, it does not explain why introducing new network layers would have such an effect. Nor does it indicate where the new network layers should be placed.

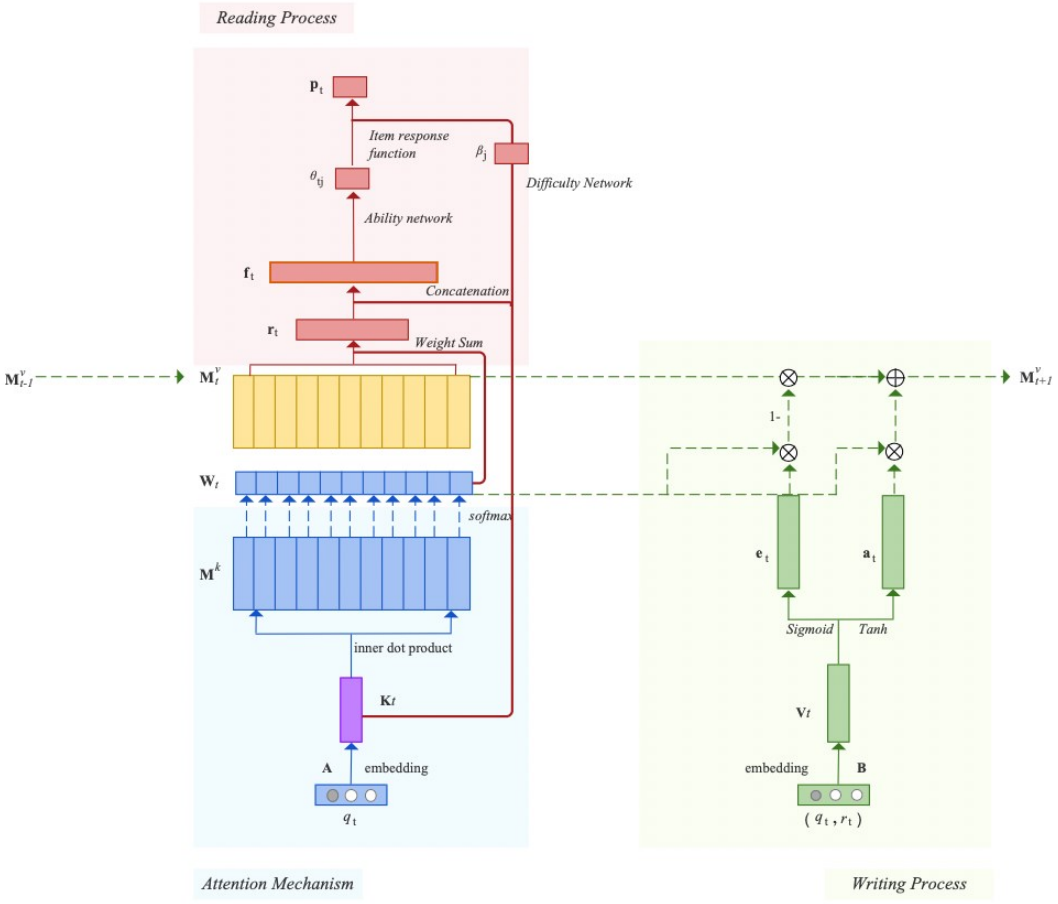

Figure 2. Architecture for Deep-IRT.

He et al. (2021) were the first researchers to provide an extensible deep knowledge tracing theory (EDKT) that could explain how adding neural networks has an impact on DKVMN performance. Drawing upon human learning theory and deep learning mechanisms, they suggested that learning factors and knowledge growth should be considered essential aspects of a DKVMN architecture. This laid a solid theoretical foundation for subsequent related research and enhanced the interpretability of DKVMN model. Meanwhile, Wang et al. (2021) proposed a deep multi-type knowledge tracing model that made use of students' non-assessed data to improve the prediction performance by changing the knowledge growth structure, which achieved promising results.

Alongside this, Tsutsumi et al. (2021) used a hyper network to optimize Deep-IRT in order to represent students' knowledge states by rebalancing historical information and the current interaction. However, to date, there has been little research seeking to expand upon the Deep-IRT model proposed by Yeung. There is a particular need to use EDKT theory to re-examine and potentially re-design the Deep-IRT model (shown in Figure 2). The basic Deep-IRT model relies solely on a neural network to estimate difficulty and ability, but this overlooks any valuable information that may be present in a student's behavioural data. EDKT provides a way of incorporating student behavioural information into Deep-IRT models. Proceeding in this fashion can enhance their predictive power while maintaining interpretability. This reflects the call from a large number of studies of the relevant literature, which point out that future studies need to explore more effective ways of integrating learning theories into deep learning. Numerous authors also emphasize the need to incorporate more relevant information in the model (e.g., Abdelrahman et al., 2023; Fanzhi et al., 2022; Song et al., 2022). This study therefore offers a positive contribution to the existing research by moving forward in this direction.

Chapter Summary

Knowledge Tracing is a supervised sequential learning problem that aims to predict the likelihood of students giving correct answers to questions, given their historical sequence of interaction. Challenges, here, are the complexity arising from questions requiring mastery of more than one concept and the tendency of students to forget information. The first use of KT was by Anderson et (Anderson et al., 1990). Corbett & Anderson (1995) subsequently sought to apply Bayesian networks to the field of KT through Bayesian Knowledge Tracing (BKT. BKT is limited by its assumptions that all students have the same level of prior knowledge, that each question only involves one skill, and that knowledge concepts are potentially interdependent. This has led to a

range of refinements, but these continue to rely on knowledge states and cannot capture rich data features. Another method is the use of logistic models, where the key idea is that the probability of correctly answering a question can be represented by a mathematical function expressing student behaviour and certain knowledge state parameters. Some of the most famous logistic models have been built using Item Response Theory (IRT), which is a collection of statistical models developed for analysing responses. Being the simplest, the Rasch model has attracted the most attention in KT. This, however, is limited in its application, so various researchers have attempted to improve it through the use of deep learning (e.g., Deep-IRT (Yeung et al., 2019)). Deep learning can model student knowledge in greater detail, make use of information about the interrelationships between concepts, and handle changes in student knowledge states over time. Piech et al. (2005) were the first to use recurrent neural networks for knowledge tracing in a model called Deep Knowledge Tracing (DKT). DKT significantly outperforms BKT but is limited by its use of hidden states to represent student knowledge and lack of attention to student learning trajectories. Zhang et al. (2017) improved the DKT model by introducing an external memory structure that uses dynamic rather than static memory to capture temporal dependencies and evolving patterns in a student's learning process This improved model, Dynamic Key-Value Memory Network (DKVMN), is still the most popular memory-based deep learning approach in KT, though other approaches have also been proposed, such as Graph-based Knowledge Tracing (GKT) (Nakagawa et al., 2019) and attention-based exercise-enhanced recurrent neural networks (Su et al., 2018). A number of attempts have been made to improve and build upon DKVMN, one of the most promising being Yeung et al.'s (2019) Deep-IRT. However, Deep-IRT lacks grounding in educational psychology and relies on neural networks to estimate difficulty and ability, rather

than attempting to deduce this from student behavioural data. The Enhanced Deep-IRT model proposed in this thesis seeks to rectify this limitation.

Chapter 3: Methods and Study design

Overview of the design

The current study used a previous research framework based on the Deep-IRT architecture to implement the proposed model (See Figure 3). The proposed model has three main modules: an 'embed input' module (that combines the input with student behavioural information and question tags); a 'prediction' module (that calculates the probability of the student giving a correct answer to a specific question); and an 'update' module (that takes the modelled result and actual student response to update the student's knowledge state). Each module is composed of multiple neural network layers.

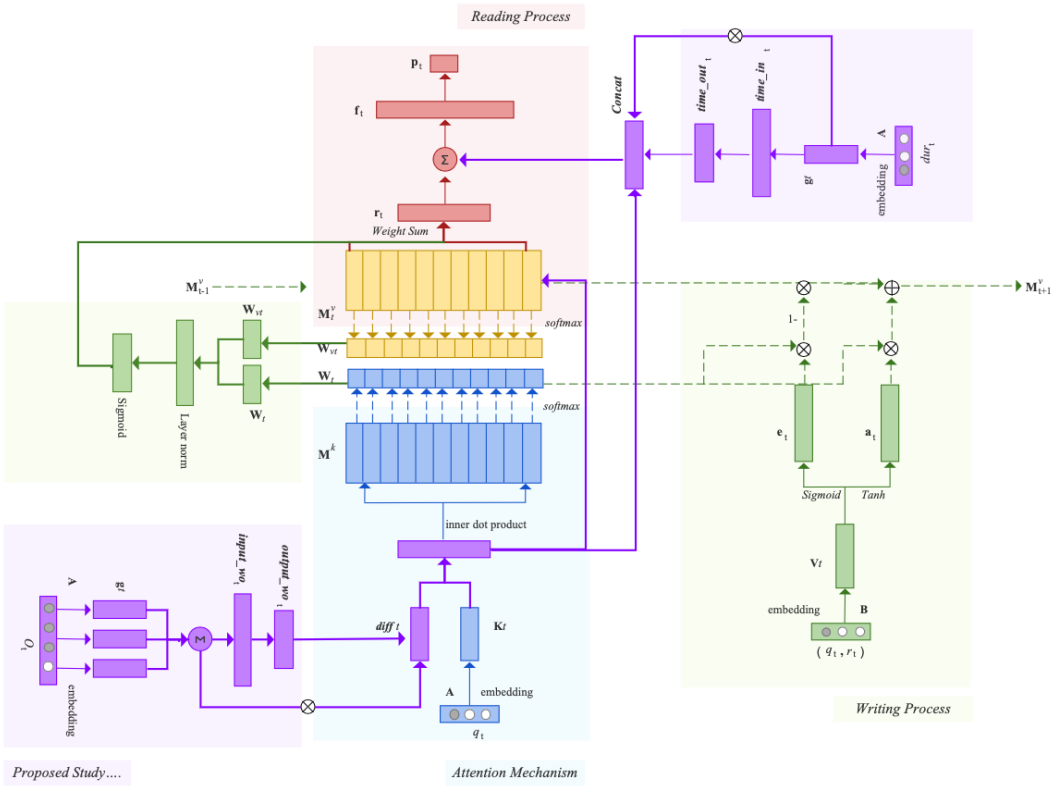

Figure 3. Architecture for proposal model (Enhanced Deep-IRT)

The model was implemented using Pytorch and Pandas. The training and validation sets contained 200 items for each iteration. These were randomly extracted from the 2009 and 2017 ASSISTments student learning trajectory datasets. The test set was the same length as the training set and randomly extracted items from the same datasets. A standard cross entropy loss function was applied for the model evaluation. As the Deep-IRT architecture is fully differentiable, the proposed model was trained using stochastic gradient descent (Zhang et al., 2017).

Before the experiment began, the raw data was pre-processed using Pandas, then loaded into memory using a Dataloader to improve the efficiency of the model's execution. The pre-processing and memory loading operations used the same methods as those used in previous studies (e.g., Sarsa et al., 2022). The experimental process involved several steps. First, the input exercise, q_t, and student behavioural data, b_t, were multiplied separately by an embedding matrix, $\mathbf{A}$, to obtain the continuous embedding vector, $\mathbf{k}_t$, and the input vector, $\mathbf{qb}_t$ for the model. The student behavioural data, $\mathbf{qb}_t$, was run through a difficulty neural network to obtain the difficulty vector, $\mathbf{Dif}_t$, for the current exercise. The difficulty vector was added to the exercise vector, $\mathbf{k}_t$, and the resulting vectors, $\mathbf{v}_t$, were used separately to calculate the similarity between the concept matrix, $\mathbf{M}^k(i)$, and the ability matrix, $\mathbf{M}_t^{\mathrm{v}}(i)$. This similarity calculation helped to determine the knowledge points and skill requirements involved in the current exercise. After this, the similarity scores, $w_{t(i)}$ and $w_{vt(i)}$, were used to calculate the estimated knowledge state, which was further processed through a fully connected layer and a sigmoid activation function. This process ultimately yielded the probability of a student correctly answering a question, indicating mastery of the corresponding knowledge points. Finally, depending on the student's actual performance, the previous ability matrix was erased and updated using a forget gate and a long short-term

memory network. This step makes it possible to capture changes in a student's ability as they practice different exercises. Table 1 shows more details about the notation used in this study.

Table 1

Notation definitions

Notation	Description
q_t	exercise record
$\mathbf{k}_t$	exercise record feature
b_t	learning behavior record
$\mathbf{qb}_t$	behavior record feature
$\mathbf{Dif}_t$	exercise difficulty feature
$\mathbf{v}_t$	cross-feature combining exercise and estimated difficulty
$\mathbf{M}^k$	concept matrix
$\mathbf{M}_t^{\mathrm{v}}$	ability matrix at time step t
$\boldsymbol{W}_{dif}$	weight of difficulty feature
$w_{t(i)}$	correlation weight of question-concept
$w_{vt(i)}$	correlation weight of question-skill
$\boldsymbol{Abi}_t$	student ability feature at time step t
$\mathbf{r}_t$	level of knowledge concept mastery
$\mathbf{Dur}_t$	duration feature
$\mathbf{f}_t$	level of overall knowledge concept mastery
p_t	probability of correct response
$\mathbf{e}_t$	erase vector
$\mathbf{a}_t$	add vector

Datasets

The study drew on two commonly used knowledge tracing datasets (ASSIST2009 and ASSIST2017) to evaluate the model's performance (see Table 2). There were several reasons for choosing these two datasets. First, they contain the most comprehensive student information currently available and have been widely used for evaluating KT models in recent years (Liu, 2022). In addition, this study compared the model's performance with two benchmark models that also made use of the same datasets (Li et al., 2023; Wang et al., 2022; Sarsa et al., 2022).

Table 2

Basic Description of the Datasets

Datasets	Students	Knowledge concepts	Records
ASSIST2009	4,217	124	401,756
ASSIST2017	1,709	102	918,106

The ASSISTments 2009 dataset (Feng et al., 2009) was collected from the ASSISTments online tutoring platform. There are two versions of this dataset. The older version contains duplicated records and is therefore not usable for the evaluation of deep learning models (Zhang et al., 2017). The current experiment used an updated 'skill-builder' version that is smaller than the original dataset. The data covers 4,217 students, 401,756 answer records, and 124 knowledge concepts.

The ASSIST2017 dataset (Patikorn et al., 2017) was also collected from the ASSISTments online tutoring platform. It contains 1,709 students, 918,106 answer records, and 102 knowledge concepts. The ASSIST2017 dataset is the richest dataset available in terms of the average number of records per student (Wang et al., 2022).

Baseline models

In order to evaluate the proposed model's performance, two typical KT models were selected to provide a baseline:

(1) *DKVMN:* This model uses memory-augmented neural networks to model students' latent knowledge concepts and dynamically updates their knowledge state over time. DKVMN is known to be an elegant architectural model that makes effective use of additional information to enhance its prediction performance.

(2) *Deep-IRT*: This is a variant of DKVMN model that differs by introducing one-parameter logistic item response theory (1PL-IRT) to the DKVMN architecture. This model provides for better interpretability and reduces the overfitting problems associated with the original DKVMN model.

Implementation details

The raw data coming from the datasets could not be input directly into the model, so Pandas was used to reformat the data. The outliers were detected using the Tukey method, which calculates the difference from the interquartile range (IQR) to determine their presence. After the pre-processing, the dataset consisted of $7n$ rows, where n represents the total number of questions answered by each student. Each question was associated with a question label, an indicator of whether the response was correct (1 means correct, 0 means incorrect), and four behavioural elements (see Table 3).

Table 3

Data features after pre-processing

Feature name	Description
user_id	Student ID
problem_id	Exercise number
correct	Response results 1 = correct 0 = incorrect
first_response (2009) / time_taken (2017)	Time spent by a student on an exercise
hint	Whether the students requested a hint during the first attempt
attempt_count	Number of attempts at an exercise
hint_count	Number of hints for an exercise

This thesis proposed embed input module has two separate neural network layers, one to estimate the difficulty of the question and the other to calculate correlations between the question, associated concepts, and ability. The difficulty estimation layer takes the available behavioural information about the students as input. b_t denotes their learning behaviour at time step t. After their behavioural information is embedded into matrix $\boldsymbol{A}$**,** it undergoes a summation operation.

$$b_t = hint_t + attemptCount_t + hintCount_t \tag{5}$$

It then passes through a fully connected layer with Tanh activation to get a summary vector, $\boldsymbol{W}_{dif}$:

$$\boldsymbol{W}_{dif} = tanh(\mathbf{W}_{wdif}[\mathbf{qb}_t] + \mathbf{b}_t) \tag{6}$$

Next, an attention mechanism is introduced because the weight of the behaviour relating to each exercise's difficulty is not fixed. The embedded sum of the student's behavioural data, $\mathbf{qb}_t$, is multiplied by $\boldsymbol{W}_{dif}$ to obtain the final difficulty vector, $\mathbf{Dif}_t$:

$$Attention(\mathbf{Dif}_t) = \mathbf{qb_t} * Sgmoid\big(\mathbf{W}_{attention}\big[\mathbf{W}_{dif}\big] + \mathbf{b}_t\big) \tag{7}$$

The correlation layer concatenates the current exercise difficulty vector, $\mathbf{Dif}_t$, and the input exercise vector, $\mathbf{q}_t$, to obtain the cross-feature, $\mathbf{v}_t$. This is then passed to the concept matrix, $\mathbf{M}^k(i)$, and the ability matrix, $\mathbf{M}_t^{\mathrm{v}}(i)$, to calculate the inner products. After that, it is passed through a softmax layer and the corresponding correlation weight vectors are obtained for the prediction module, i.e.:

$$w_{t(i)} = softmax\left(\mathbf{M}_i^k \mathbf{v}_t\right) \tag{8}$$

$$w_{vt(i)} = softmax(\mathbf{M}_i^v \mathbf{v}_t) \tag{9}$$

Previous research has indicated a correlation between the accuracy of students' answers and their working memory capacity (Darolia & Varshney, 2015). Inspired by these findings, the following algorithm incorporates different relationships in the student ability network to represent working memory capacity. $w_{t(i)}$ is used to denote the question-concept relationship, and $w_{vt(i)}$ is used to denote the question-skill relationship.

$$\boldsymbol{Abi}_t = Sigmoid\left(\mathbf{W}_{Abi}\left(\| w_{t(i)} \| + \| w_{vt(i)} \|\right) + \mathbf{b}_{Abi}\right) \tag{10}$$

Once the corresponding correlation weight, $\mathbf{Abi}_t(i)$, has been obtained, the model uses it to perform a weighted sum of all the ability matrix slot vectors, $\mathbf{M}_t^{\mathrm{v}}$. This represents the read content (Yeung, 2019; Zhang et al., 2017), r_t, which is the student's level of mastery of the associated knowledge concept, i.e.:

$$\mathbf{r}_t = \sum_{i=1}^{N} \mathbf{Abi}_t(i) \mathbf{M}_t^v(i)^T \tag{11}$$

As that the time required for a student to respond to an exercise is also related to their ability, the more complex the question, the longer it will take them to answer, and vice versa. Previous research has shown that the time taken also has an impact on the learning outcome. In that case, drawing on EDKT theory, the time taken was also introduced as a learning factor within

the model. Duration, here, is denoted by the independent feature, dur_t, and is constructed as follows:

$$\mathbf{Dur}_t = Sigmoid(\mathbf{W}_{dur}[dur_t] + \mathbf{b}_t) \quad (12)$$

$\mathbf{r}_t$, $\mathbf{v}_t$, $\boldsymbol{Abi}_t$and $\mathbf{Dur}_t$ are now combined and passed through a fully connected layer with tanh activation to obtain the overall concept mastery (Zhang et al., 2017), $\mathbf{f}_t$:

$$\mathbf{f}_t = tanh(\mathbf{W}_f[\mathbf{r}_t, \mathbf{v}_t, \mathbf{Abi}_t, \mathbf{Dur}_t] + \mathbf{b}_f) \quad (13)$$

Finally, $\mathbf{f}_t$ is passed through another fully connected layer with Sigmoid activation, which outputs the probability of the student answering q_t correctly (Yeung, 2019; Zhang et al., 2017):

$$p_t = P(q_t) = Sigmoid\big(\mathbf{W}_p\mathbf{f}_t + \mathbf{b}_p\big) \quad (14)$$

After the student has given their response (correct or incorrect), the question q_t and response a_t are taken together to update the value matrix $\mathbf{M}_t^v$. First, the tuple (q_t, a_t) is embedded within the matrix $\mathbf{B} \in \mathbb{R}^{2Q*d_v}$ to obtain the vector $\mathbf{u}_t \in \mathbb{R}^{d_v}$, which represents the change in the student's knowledge after working on the exercise. Next, drawing upon the idea of the input and forget gates in a long short-term memory network, part of the memory is erased using the erase vector, $\mathbf{e}_t \in \mathbb{R}^{d_v}$. This step aims to mimic the potential impact of forgetting behaviour on learning growth (Zhang et al., 2017):

$$\mathbf{e}_t = sigmoid(\mathbf{W}_e\mathbf{u}_t + \mathbf{b}_e) \quad (15)$$

$\mathbf{e}_t$ is a column vector in the range (0,1). After erasure, an add vector, $\mathbf{a}_t \in \mathbb{R}^{d_v}$, is used to update each memory slot in $\mathbf{M}_t^v$ (Zhang et al., 2017):

$$\mathbf{a}_t = tanh(\mathbf{W}_a\mathbf{u}_t + \mathbf{b}_a) \quad (16)$$

Finally, the new value matrix, $\mathbf{M}_t^v$, can be formulated as follows (Yeung, 2019; Zhang et al., 2017):

$$\mathbf{M}_{t+1}^v = \tilde{\mathbf{M}}_t^v(i) + w_{ti}\mathbf{a}_t^T \quad = \mathbf{M}^v\,(i) \otimes (1 - w_{ti}\mathbf{e}_t)^T + w_{ti}\mathbf{a}_t^T \quad (17)$$

The experiment adopted a fivefold cross-validation method to evaluate the model's performance. For each dataset, 70% was used for training and validation and 30% for testing. For each run-through, the validation set was used to fine-tune the hyperparameters. The experiments were conducted five times for each dataset, and the average of the five results was taken as the final result. Unlike most previous studies (Wang et al., 2021; Zhang et al., 2017), in which the concept and ability matrix has been initialized in advance, this study used a Kaiming normal distribution to initiate the model, thereby avoiding gradient explosion. The memory size, d_0, was set at 20 dimensions for each dataset, and both d_k and d_v were set at 200. The model used an Adam optimizer, with an initial learning rate of 0.01. The learning rate was reduced if the training loss increased. The calculation framework used for the experiments was Pytorch and the experiments were conducted on an NVIDIA RTX 3060.

This study followed other empirical research that has suggested using a cross-entropy loss function to optimize the model. All the learnable parameters in the model were jointly learned by minimizing the cross-entropy loss between the real label, a_t, and the predicted value, p_t, as follows:

$$\ell = -\sum_t (a_t log p_t + (1 - a_t)\log(1 - p_t)) \tag{18}$$

Chapter Summary

This thesis reports on a study that used student behavioural information to redesign the Deep-IRT neural network. The proposed model uses question-concept and question-skill relationships to build a new network to represent student ability. The model's architecture is constructed around three network modules: 'embed input'; 'prediction'; and 'update'. Two KT datasets (ASSIST 2009 and ASSIST 2017) were used to tune the model's hyperparameters and evaluate its performance. The datasets were pre-processed in Pandas. Two baseline models,

DKVMN and Deep-IRT, were compared with the proposed model to assess its relative performance. The results of the various experiments undertaken are reported in the next chapter.

Chapter 4: Results

This chapter compares the performance of DKVMN, Deep-IRT, and the proposed model when applied to the two large evaluation datasets. The chapter concludes with the results of an ablation test, which was undertaken to assess whether each new factor helped to improve the model's performance.

Overall model performance

According to the previous empirical study, the Area under the Curve (AUC) is a commonly used metric in the knowledge tracing model (Sarsa et al., 2022). It corresponds to the likelihood that the classifier will rank a randomly selected positive instance higher than a randomly selected negative instance. Moreover, some research apply AUC as the performance index. Hence the current study uses the AUC to report the proposed model performance (see Figure 4).

Two experiments, one using ASSIST 2009 and the other ASSIST 2017, were conducted for each of the three models. These revealed a number of differences in their performance. For ASSIST 2009, the AUC value for the proposed model was 98.33%, which was 15.44% and 16.70% higher than DKVMN and Deep-IRT, respectively. For ASSIST 2017, the AUC value for the proposed model was 86.12%, which was 19.24% and 20.83% higher than DKVMN and Deep-IRT, respectively. Compared with state-of-the-art models (AUC: 0. 919), the proposed model improved the AUC values by 6.43% for the ASSIST 2009 dataset (Sun et al., 2021). Table 4 provides an overview of the above results.

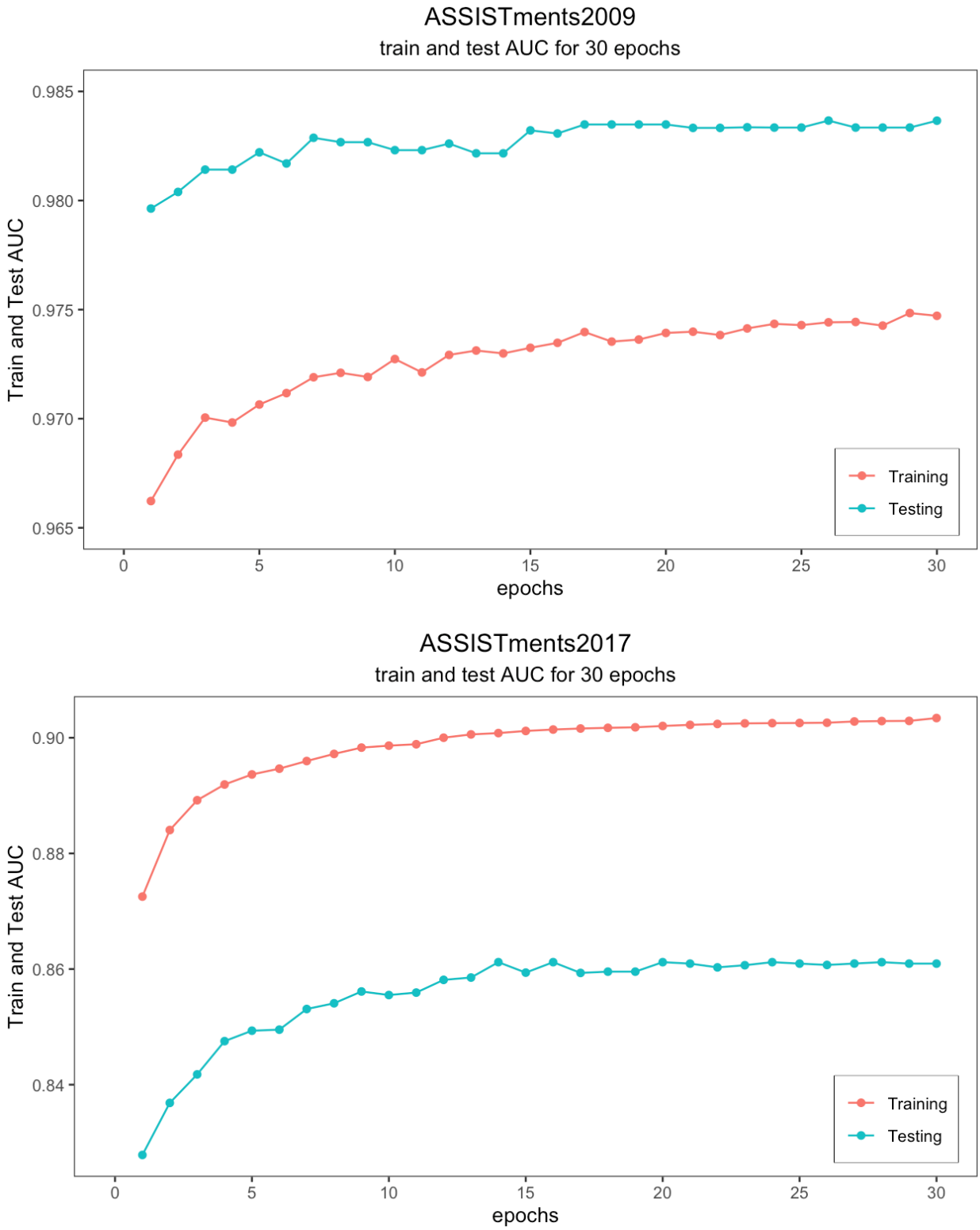

Figure 4. Comparative performance for the training and test data when applying the proposed model to the ASSIST 2009 and ASSIST 2017 datasets.

Model performance and comparison of the results

Experiment 1 compared the performance of DKVMN with the proposed model using the same hyperparameters across the two evaluation datasets. The AUC of DKVMN was 66.87% for the ASSIST 2017 dataset and 82.89% for the ASSIST 2009 dataset (see Figure 5). As the proposed model enhances the input features more than DKVMN, the DKVMN model was also modified so that it was using the same features as the proposed model. The AUC of the modified DKVMN was 85.16% for the ASSIST 2017 dataset and 96.98% for the ASSIST 2009 dataset. However, the proposed model outperformed the modified DKVMN by 2.05% for ASSIST 2017 and by 0.96% for ASSIST 2009.

Experiment 2 investigated the differences in performance of Deep-IRT and the proposed model for the same two datasets. The Deep-IRT model's AUC was 65.29% for ASSIST 2017 and 81.63% for ASSIST 2009 (see Figure 5). As with Experiment 1, the input structure of Deep-IRT was also modified so that it used the same features as the proposed model. The AUC of the modified Deep-IRT was 84.96% for ASSIST 2017 and 97.43% for ASSIST 2009. The proposed model outperformed Deep-IRT by 1.16% for ASSIST 2017 and 0.9% for ASSIST 2009.

Table 4

Overall performance of the different models when applied to the two datasets

Model	ASSIST 2009	ASSIST 2017
DKVMN	0.8289	0.6687
Baseline Deep-IRT	0.8163	0.6529
Enhanced Deep-IRT (current)	0.9833	0.8612

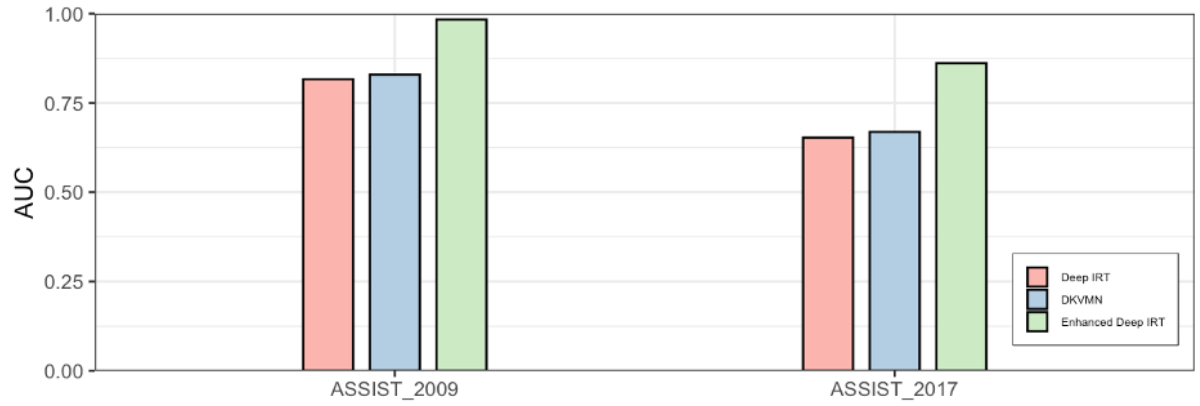

Figure 5. Performances of all models on all datasets.

Ablation study results

This research adopts a technique inspired by neuroscience, specifically the ablation studies approach, to explore how a single factor influences the model and examine the effects of various feature combinations on the model's performance. In traditional neuroscience studies, researchers tend to damage specific neural areas to investigate different neural tissue influences on the brain's capabilities to perform a specific task, which can obtain insight into the brain's structure and organization of processing. Since ablation studies have proven valuable in examining complex neural systems such as primate brains (Meyes et al., 2019), exploring their potential for advancing state-of-the-art artificial neural systems is reasonable. Moreover, most deep knowledge tracing model research utilized ablation studies to confirm that their neuro layer design can positively contribute to the model performance (He et al., 2021; Xiao et al., 2022). Hence, to better understand the contribution of each factor to the proposed model's final performance, an ablation study was undertaken that involved several different experiments using the datasets. The proposed model centres around three key factors: difficulty, ability, and time taken. The cumulative totals for these are shown in Table 5.

By adding the new modules to Deep-IRT, its AUC was generally enhanced, as was shown for both the 2017 and 2009 datasets. However, by combining different factors, the model's performance may be improved to different degrees. As can be seen, a combination of attention to difficulty and ability yielded a higher AUC than just using the difficulty. This was the case for both the 2009 and 2017 datasets. A combination of all three factors (difficulty, ability, and time taken) produced the best overall performance.

Table 5

Ablation study results for different model configurations

Model	ASSIST 2009	ASSIST 2017	Compare Baseline
Baseline (Deep-IRT)	0.8163	0.6529	-
Enhanced Deep-IRT (*ability*)	0.8184	0.6543	+ 0.0021 (2009) + 0.0014 (2017)
Enhanced Deep-IRT (*time*)	0.8168	0.6554	+ 0.0005 (2009) + 0.0025 (2017)
Enhanced Deep-IRT (*difficulty*)	0.9808	0.8573	+ 0.1645 (2009) + 0.2044 (2017)
Enhanced Deep-IRT (*difficulty+ability*)	0.9829	0.8587	+ 0.1666 (2009) + 0.2058 (2017)
Enhanced Deep-IRT (*difficulty+time*)	0.9813	0.8588	+ 0.1650 (2009) + 0.2059 (2017)
Enhanced Deep-IRT (*difficulty+ability+time*)	0.9833	0.8612	+ 0.1670 (2009) + 0.2083 (2017)

Chapter summary

The performance of the proposed model was evaluated in comparison to two other baseline models, DKVMN and Deep-IRT, using the ASSIST 2009 and ASSIST 2017 datasets. The proposed model (enhanced deep-IRT) outperformed DKVMN by between 15.44% and 19.25%

and Deep-IRT by between 16.70% and 20.83%, depending on the specific dataset and modifications used. An ablation study revealed that the factors drawn upon within each dataset can also influence the performance results. It was established that the proposed Enhanced Deep-IRT model using features relating to difficulty, ability, and time taken will produce the best results. Overall, the experiments confirmed that an improvement in model performance relies on introducing students' behaviour information and redesigning the network architecture in the ways proposed in Chapter 3.

Chapter 5: Discussion

Since the introduction of the first deep learning model specifically designed for knowledge tracing in 2015 (Piech et al., 2015), various types of deep learning models have been proposed. These have contributed new insights regarding the tracking and prediction of student performance, given the prospects offered by the processing of big data (Liu et al., 2023). Out of all the different models developed, DKVMN has perhaps generated the most interest amongst researchers due to its interpretability and elegant and scalable structure (Zhang et al., 2017). While many algorithms have now been proposed that enhance some aspect of the DKVMN model, there has been relatively little work done so far regarding the incorporation of theoretical insights coming out of educational psychology. An exception, here, is Deep-IRT, which is a pioneering model that attempts to combine IRT theory with DKVMN. However, in the Deep-IRT model, the estimation of difficulty and ability relies entirely on what is already present in the neural networks, thereby missing the opportunity to build in additional important information (Yeung, 2019). Inspired by the Transformation design, I have attempted to enhance the performance and interpretability of the Deep-IRT model by incorporating attention mechanisms and normalization layers. In practical terms, this has involved not only looking at how to enhance the model's input features but also modification of the model by adding in difficulty and ability layers.

As indicated at the outset, the principal goal of the research reported in this thesis was to investigate ways of improving the Deep-IRT model's performance. My first step towards accomplishing this was to enhance the input features based on the features of student learning interactions (i.e., the total number of attempts, the number of hints provided, and the time elapsed) instead of relying only on questions and student responses as the input. The next step was to develop new algorithms that could estimate an item's difficulty and the ability required to be able

to respond to it correctly. For estimation of the difficulty, This study developed a difficulty neural network that introduces attention mechanisms that dynamically model student's perception of a question's difficulty. For the assessment of student ability, I created an ability neural network that considers not only the relationship between questions and knowledge concepts (Q-C), but also the relationship between questions and skills (Q-S). An integrated relationship between knowledge, concepts, and skills is obtained through a normalization layer, which estimates a student's current ability state. In addition, an independent elapsed time network is used to balance the influence of the different difficulty and ability estimations to predict the student's responses. This latter element was built in because it is important to determine the impact of time-related factors on student performance predictions by taking into account the time taken to answer a question and how this relates to the question's difficulty and the student's ability.

The secondary aspect of the research was to compare the proposed model's performance with baseline models (DKVMN and Deep-IRT) and to investigate the extent to which the redesigned aspects of the algorithm were contributing to the proposed model's performance. To achieve this, the present study utilized small datasets (from two extensive datasets) to execute multiple replicated experiments to identify appropriate hyperparameters for both the baseline and proposed models. Additionally, an ablation study was conducted to evaluate the performance of individual algorithms applied to each neural layer. This facilitated observation of the respective contributions of each neural layer to the proposed model's AUC value.

Compared with the baseline model (Deep IRT), the current proposed model introduces an independent neural network for estimating difficulty, redesigns the algorithm for updating the ability matrix, and incorporates elapsed time as a separate feature into the model. Given the use of

the same dataset, the algorithm for the item input layer remains consistent with the previous approach, serving as the input feature for the model.

The original research questions were:

1) Does introducing student behavioural information via independent neural networks improve the baseline model's performance?
2) Does the proposed algorithm have better predictive performance than other baseline models?

The impact of introducing student behavioural information

The ablation study confirmed that the proposed algorithms have outperform the baseline model. Specifically, introducing a difficulty estimation network improved the Deep -IRT baseline model by 16.70% for the ASSIST 2009 dataset and 20.83% for the ASSIST 2017. As mentioned earlier, there are two ways of estimating question difficulty: absolute difficulty assessment; and relative difficulty assessment (Xu et al., 2021). Here, student behavioural information in the datasets was used to build a relative difficulty estimation network, instead of relying on just the character of a specific question. This approach is more open to generalization. For the ability assessment network, there was an improvement on the baseline model of 0.21% for the 2009 dataset and 0.14% for the 2017 dataset. Unlike baseline models that only use a single fully connected layer and a tanh activation function to describe student ability, the proposed model uses the question-concept and question-skill relationship to model student ability. This is a novel approach to knowledge tracing. These features are combined using a fully connected layer and normalization layer to showcase student ability. For the network focused upon the time elapsed, there were improvements over the baseline model of 0.05% for the ASSIST 2009 dataset and of 0.25% for the ASSIST 2017 dataset. This finding is consistent with previous research, which

suggests that duration may contribute to model performance, although the impact may not be particularly pronounced (Xiao et al., 2022). A possible reason is that, although research has shown a relationship between the time taken to answer a question and a student's mastery of latent concepts (Pelánek & Jarušek, 2015), individual differences in students' emotional and cognitive patterns during problem-solving, makes any prediction of student performance based on duration challenging (Ofelia et al., 2013).

The predictive performance in comparison to established baseline models

The study reported here used AUC scores to measure each model's performance. The results of this exercise suggest that enhancing the input features does, indeed, lead to an improvement in model performance. After enhancing the input, the AUC score was improved by 18.67% for the ASSIST 2009 dataset and 4.01% for the ASSIST 2017 dataset. Similarly, for the Deep-IRT model, adding in the enhanced input features resulted in an improvement of 17.52% for the ASSIST 2009 dataset and 2.88% for the 2017 ASSIST dataset. A somewhat surprising result here is that, although the same features were enhanced for both datasets (such as the number of attempts, the response time, the number of requests for help, and the initial attempt behaviour), the resulting improvement was better for the 2009 dataset. A likely reason for this is that the 2009 dataset was not over-parameterized and had incorporated other factors (He et al., 2021). Other prior research has found the same phenomenon. The state-of-the-art model for ASSIST 2009 prior to the current study achieved an AUC score of 91.9%. In addition, as a report by He et al. (2021) has already suggested, the structure of DKVMN and its variants enables it to leverage neural networks in ways in which it can automatically discover the relationship between skills and latent concepts. As a result, enhancing the input features does not lead to over-parameterization or the

overfitting issues that have been seen with other deep knowledge tracing models (i.e., DKT). Hence, this study provides additional evidence for the validity of these previous hypotheses.

Chapter Summary

An enhanced Deep-IRT model has been proposed that takes forward the state of the art by incorporating attention mechanisms and normalization layers. In particular, this has sought to build in inspiration from educational psychology with regard to the relationship between working memory and question-concept (or skill) relationships. The research associated with this model has set out to address two research questions, the first regarding whether building in network layers relating to difficulty, ability and duration can augment Deep-IRT's performance, and the second regarding whether the resulting improvements enable the model to outperform existing baseline models. In relation to the first question, it was found that difficulty and ability factors had the most notable effect on performance, but that duration also had a slight effect. In the case of the second question, it has been demonstrated that the proposed model offers advantages over existing models, though it made a more significant difference for the ASSIST 2009 dataset than for the ASSIST 2017 dataset.

Chapter 6: Conclusion

The study reported in this thesis focused on the development of an enhanced Deep-IRT model. The model is a knowledge-tracing algorithm that is able to incorporate and build predictions upon information about the behaviour of students when engaged in learning exercises. The results of experiments conducted on two real-world datasets indicate that the proposed model outperformed baseline models such as DKVMN and Deep-IRT. In particular, the proposed model helped to improve the interpretability of the baseline model. Drawing upon theories relating to EDKT, a set of distinct neural network layers was created, and an appropriate overall architecture was designed. Specific innovations in the model that significantly contributed to its success were the use of relative difficulty theory (Xu et al., 2021), research on working memory (Darolia & Varshney, 2015), and question-concept (or skill) relationships (Ma et al., 2022) derived from educational psychology. The relative difficulty perspective on questions suggests that the difficulty of a question reflects the individual knowledge level of students and that the difficulty of a question should be quantified through the interactive performance between students and the question (Xu et al., 2021). The model's design provides for a more accurate and explainable assessment of student abilities and experienced levels of difficulty. Generally, the model facilitates a better understanding of students' learning processes and performance.

Limitations of the Study and the Directions for Future Research

Inevitably, the study reported here had some limitations. First, the experiments were conducted using specific datasets, which may introduce data bias and limit the model's generalizability. The publicly available education datasets suitable for knowledge tracing model evaluation mainly derive from the ASSISTment Data project and competition datasets geared towards the development of artificial intelligence. The ASSIST 2009 and 2017 datasets contain

the most information about student behaviour. However, these datasets only record students' responses to mathematical questions, so they are limited with regard to what they may say about learning in other subjects (Song et al., 2022). Indeed, predicting student performance in other subjects using the current research model would be challenging. Future studies therefore need to explore how best to use datasets relating to other subjects to improve the generalisability of the existing models.

Second, the current study used the simplest Rasch model to build the proposed neural networks. As a result, the proposed model does not consider other potential factors that might impact a student's response, such as guessing. Ding and Larson (2020) identify a number of areas for improvement in existing deep knowledge tracing models with regard to the handling of uncertain student behaviour (including guessing). They suggest introducing a regularization mechanism to adjust the loss function, so that it can capture uncertain student behaviour more effectively. This regularization mechanism can help a model to handle phenomena such as guessing when students are responding to questions, thereby improving the model's predictive performance and reducing the impact of the uncertainty. Future research should therefore explore ways of incorporating other algorithms into the model's loss function, so that the impact of uncertain behaviour can be properly taken into account when making predictions. The first step towards this would be to test a range of regularization mechanisms and evaluate their effect on the model's performance. Ultimately, this feature will enhance the model's robustness and generalizability and make it more reliable and effective in real-world applications.

Finally, while the proposed model achieved outstanding predictive performance and enhanced interpretability, the network architecture is relatively complex. This led to slow convergence during training. It is therefore worth exploring how to integrate new deep learning

algorithms, such as Transformer, that might help to enhance the model's computational efficiency. This, too, will improve its effectiveness for real-world applications.

By pursuing the various new avenues of exploration proposed above, the work commenced in this thesis can continue to contribute to new advances in the field of deep knowledge tracing, as well as leaving it open to being inspired by new developments in the field.

References

Abdelrahman, G., & Wang, Q. (2019). *Knowledge Tracing with Sequential Key-Value Memory Networks*. https://doi.org/10.1145/3331184.3331195

Abdelrahman, G., & Wang, Q. (2021, August 18). *Deep Graph Memory Networks for Forgetting-Robust Knowledge Tracing*. ArXiv.org. https://doi.org/10.48550/arXiv.2108.08105

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

Agarwal, D., Baker, R. S., & Muraleedharan, A. (2020). Dynamic Knowledge Tracing through Data Driven Recency Weights. In *ERIC*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED607821

Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-Aware Deep Knowledge Tracing and Exercise Recommendation in an Online Learning System. In *ERIC*. International Educational Data Mining Society. https://eric.ed.gov/?id=ED599194

Cen, H., Koedinger, K. R., & Junker, B. W. (2006). Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *Springer EBooks*, 164–175. https://doi.org/10.1007/11774303_17

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, *4*(4), 253–278. https://doi.org/10.1007/bf01099821

Darolia, C. R., & Varshney, N. (2015). On the Relationship between Working Memory and Fluid:Crystallized Intelligence. *Indian Journal of Health and Wellbeing*, *6*(12), 1229–1231. https://www.i-scholar.in/index.php/ijhw/article/view/147619

Deonovic, B., Yudelson, M., Bolsinova, M., Attali, M., & Maris, G. (2018, October 12). *Learning meets Assessment: On the relation between Item Response Theory and Bayesian Knowledge Tracing*. ArXiv.org. https://doi.org/10.48550/arXiv.1803.05926

Ding, X., & Larson, E. C. (2020). Incorporating uncertainties in student response modeling by loss function regularization. *Neurocomputing*, *409*, 74–82. https://doi.org/10.1016/j.neucom.2020.05.035

Ding, X., & Larson, E. C. (2021). On the Interpretability of Deep Learning Based Models for Knowledge Tracing. *ArXiv:2101.11335 [Cs]*. https://arxiv.org/abs/2101.11335

Embretson, S. E., & Steven Paul Reise. (2013). *Item response theory for psychologists*. L. Erlbaum Associates.

Fanzhi, Z., Luqian, X. U., Yan, Z., Yuexia, Z., & Junwei, L. (2022). Review of Knowledge Tracing Model for Intelligent Education. *Jisuanji Kexue Yu Tansuo*, *16*(8), 1742–1763. https://doi.org/10.3778/j.issn.1673-9418.2111054

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. User Modeling and User-Adapted Interaction, 19(3), 243–266. https://doi.org/10.1007/s11257-009-9063-7

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. Routledge.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476. https://doi.org/10.1038/nature20101

He, L., Li, X., Tang, J., & Wang, T. (2021). *EDKT: An Extensible Deep Knowledge Tracing Model for Multiple Learning Factors*. 340–355. https://doi.org/10.1007/978-3-030-73194-6_23

Käser, T., Klingler, S., Schwing, A. G., & Gross, M. (2014). Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks. *Lecture Notes in Computer Science*, 188–198. https://doi.org/10.1007/978-3-319-07221-0_23

Lee, J.-I., & Brunskill, E. (2012). The Impact on Individualizing Student Models on Necessary Practice Opportunities. *Educational Data Mining*, 118–125.

Li, Z., Yu, S., Lu, Y., & Chen, P. (2023). Plastic gating network: Adapting to personal development and individual differences in knowledge tracing. *Information Sciences*, *624*, 761–776. https://doi.org/10.1016/j.ins.2023.01.011

Liu, C., & Li, X. (2021). *Memory Attentive Cognitive Diagnosis for Student Performance Prediction*. 79–90. https://doi.org/10.1007/978-981-16-8143-1_8

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2023). A Survey of Knowledge Tracing. *ArXiv:2105.15106 [Cs]*. https://arxiv.org/abs/2105.15106

Liu, T. (2022). Knowledge tracing: A bibliometric analysis. *Computers and Education: Artificial Intelligence*, *3*, 100090. https://doi.org/10.1016/j.caeai.2022.100090

Ma, Y., Han, P., Qiao, H., Cui, C., Yin, Y., & Yu, D. (2022). SPAKT: A Self-supervised Pre-trAining method for Knowledge Tracing. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2022.3187987

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019, October 1). *Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network*. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8909656

Meyes, R., Lu, M., Constantin, & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. *ArXiv (Cornell University)*

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2021). Graph-based knowledge tracing: Modeling student proficiency using graph neural networks. *Web Intelligence*, *19*(1-2), 87–102. https://doi.org/10.3233/web-210458

Ni, Q., Wei, T., Zhao, J., He, L., & Zheng, C. (2022). *HHSKT: A learner–question interactions based heterogeneous graph neural network model for knowledge tracing*. *215*, 119334–119334. https://doi.org/10.1016/j.eswa.2022.119334

Ofelia, M., Baker, R. S., Gowda, S. M., & Heffernan, N. T. (2013). *Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System*. 41–50. https://doi.org/10.1007/978-3-642-39112-5_5

Pandey, S., & Karypis, G. (2019). A Self-Attentive model for Knowledge Tracing. *ArXiv:1907.06837 [Cs, Stat]*. https://arxiv.org/abs/1907.06837

Patikorn, T., Baker, R. S., & Heffernan, N. T. (2017). ASSISTments Longitudinal Data Mining Competition Special Issue: A Preface. Journal of Educational Data Mining, 12(2). https://doi.org/10.5281/zenodo.4008048

Pavlik, P. I., Eglington, L. G., & Harrell-Williams, L. M. (2021). Logistic Knowledge Tracing: A Constrained Framework for Learner Modeling. *IEEE Transactions on Learning Technologies*, *14*(5), 624–639. https://doi.org/10.1109/tlt.2021.3128569

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*(3-5), 313–350. https://doi.org/10.1007/s11257-017-9193-2

Pelánek, R., & Jarušek, P. (2015). Student Modeling Based on Problem Solving Times. *International Journal of Artificial Intelligence in Education*, *25*(4), 493–519. https://doi.org/10.1007/s40593-015-0048-x

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep Knowledge Tracing. *ArXiv:1506.05908 [Cs]*. https://arxiv.org/abs/1506.05908

Santoro, A., Sergey Bartunov, Botvinick, M., Daan Wierstra, & Lillicrap, T. (2016). Meta-Learning with Memory-Augmented Neural Networks. *PMLR*, 1842–1850. http://proceedings.mlr.press/v48/santoro16.html

Sarsa, S., Leinonen, J., & Hellas, A. (2022). Empirical Evaluation of Deep Learning Models for Knowledge Tracing: Of Hyperparameters and Metrics on Performance and Replicability. *ArXiv:2112.15072 [Cs]*. https://arxiv.org/abs/2112.15072

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2021). SAINT+: Integrating Temporal Features for EdNet Correctness Prediction. *ArXiv:2010.12042 [Cs]*. https://doi.org/10.1145/3448139.3448188

Song, X., Li, J., Cai, T., Yang, S., Yang, T., & Liu, C. (2022). A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, *258*, 110036. https://doi.org/10.1016/j.knosys.2022.110036

Spaulding, S., & Breazeal, C. (2015). Affect and Inference in Bayesian Knowledge Tracing with a Robot Tutor. *MIT Web Domain*. https://dspace.mit.edu/handle/1721.1/109395

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C., Wei, S., & Hu, G. (2018). Exercise-Enhanced Sequential Modeling for Student Performance Prediction.

Proceedings of the AAAI Conference on Artificial Intelligence, *32*(1). https://doi.org/10.1609/aaai.v32i1.11864

Sun, X., Zhao, X., Li, B., Ma, Y., Sutcliffe, R., & Feng, J. (2021). Dynamic Key-Value Memory Networks With Rich Features for Knowledge Tracing. *IEEE Transactions on Cybernetics*, 1–7. https://doi.org/10.1109/tcyb.2021.3051028

Tsutsumi, E., Kinoshita, R., & Ueno, M. (2021). Deep Item Response Theory as a Novel Test Theory Based on Deep Learning. *Electronics*, *10*(9), 1020. https://doi.org/10.3390/electronics10091020

Vie, J.-J., & Kashima, H. (2018). Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing. *ArXiv:1811.03388 [Cs, Stat]*. https://arxiv.org/abs/1811.03388

Wang, C., Zhao, S., & Sahebi, S. S. (2021). *Learning from Non-Assessed Resources: Deep Multi-Type Knowledge Tracing*. Semantic Scholar. https://www.semanticscholar.org/paper/Learning-from-Non-Assessed-Resources%3A-Deep-Tracing-Wang-Zhao/7433411d47d4a58a92121df072b81b7f50f50925

Wang, X., Zheng, Z., Zhu, J., & Yu, W. (2022). What is wrong with deep knowledge tracing? Attention-based knowledge tracing. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03621-1

Wilson, K. R., Karklin, Y., Han, B., & Chaitanya Ekanadham. (2016). *Back to the Basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation*.

Wright, R. (1995). Manuel Alvar, “Estudios léxicos: segunda serie” (Book Review). *Bulletin of Hispanic Studies*, *72*(2), 217–217. https://doi.org/10.3828/bhs.72.2.217a

Xiao, Y., Xiao, R., Huang, N., Hu, Y., Li, H., & Sun, B. (2022). Knowledge tracing based on multi-feature fusion. *Neural Computing and Applications*, *35*(2), 1819–1833. https://doi.org/10.1007/s00521-022-07834-w

Xu, J., Wei, T., Yu, G., Huang, X., & Lyu, P. (2021). Review of Question Difficulty Evaluation Approaches. *Journal of Frontiers of Computer Science and Technology*, 735–759. https://doi.org/10.3778/j.issn.1673-9418.2108086

Yeung, C.-K. (2019). Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. *ArXiv:1904.11738 [Cs, Stat]*. https://arxiv.org/abs/1904.11738

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into Deep Learning. *ArXiv:2106.11342 [Cs]*. https://arxiv.org/abs/2106.11342

Zhang, J., Shi, X., King, I., & Yeung, D.-Y. (2017). Dynamic Key-Value Memory Networks for Knowledge Tracing. *Proceedings of the 26th International Conference on World Wide Web*. https://doi.org/10.1145/3038912.3052580

Zhao, W., Xia, J., Jiang, X., & He, T. (2023). A novel framework for deep knowledge tracing via gating-controlled forgetting and learning mechanisms. *Information Processing & Management*, *60*(1), 103114. https://doi.org/10.1016/j.ipm.2022.103114

Zou, Y., Yan, X., & Li, W. (2020). Knowledge Tracking Model Based on Learning Process. *Journal of Computer and Communications*, *8*(10), 7–17. https://doi.org/10.4236/jcc.2020.810002