

Transition-Aware Multi-Activity Knowledge Tracing

1st Siqian Zhao

Computer Science Department
University at Albany, SUNY
Albany, NY 12222, USA
szhao2@albany.edu

2nd Chunpai Wang

Computer Science Department
University at Albany, SUNY
Albany, NY 12222, USA
cwang25@albany.edu

3rd Shaghayegh Sahebi

Computer Science Department
University at Albany, SUNY
Albany, NY 12222, USA
ssahebi@albany.edu

Abstract—Accurate modeling of student knowledge is essential for large-scale online learning systems that are increasingly used for student training. Knowledge tracing aims to model student knowledge state given the student’s sequence of learning activities. Modern Knowledge tracing (KT) is usually formulated as a supervised sequence learning problem to predict students’ future practice performance according to their past observed practice scores by summarizing student knowledge state as a set of evolving hidden variables. Because of this formulation, many current KT solutions are not fit for modeling student learning from non-assessed learning activities with no explicit feedback or score observation (e.g., watching video lectures that are not graded). Additionally, these models cannot explicitly represent the dynamics of knowledge transfer among different learning activities, particularly between the assessed (e.g., quizzes) and non-assessed (e.g., video lectures) learning activities. In this paper, we propose Transition-Aware Multi-activity Knowledge Tracing (TAMKOT), which models knowledge transfer between learning materials, in addition to student knowledge, when students transition between and within assessed and non-assessed learning materials. TAMKOT is formulated as a deep recurrent multi-activity learning model that explicitly learns knowledge transfer by activating and learning a set of knowledge transfer matrices, one for each transition type between student activities. Accordingly, our model allows for representing each material type in a different yet transferrable latent space while maintaining student knowledge in a shared space. We evaluate our model on three real-world publicly available datasets and demonstrate TAMKOT’s capability in predicting student performance and modeling knowledge transfer.

Index Terms—knowledge tracing, knowledge transfer, multi-activity, massive open online courses, student knowledge modeling, multiple learning material types

I. INTRODUCTION

The popularity and necessity of online education have increased the use of online education systems in recent years. Given the amount of data produced as a result of student interactions, automatically understanding individual students’ knowledge and learning process is essential for the success of such online education systems. Knowledge Tracing (KT) models aim to quantify students’ state of knowledge at each point of the learning period. Many modern KT models are formulated as a supervised sequence learning problem to predict students’ future practice performance according to their past performances in learning activities [1], [1]–[8]. Students learn from activities such as solving problems, taking tests, reviewing worked examples, and watching video lectures [9], [10] (multi-activity). Among these activities, the ones that can

be used to assess student performance, like problem-solving and test-taking, serve as relatively reliable measures of student knowledge. Unlike these assessed activities, the non-assessed activities, such as reading worked examples, cannot provide explicit indicators of knowledge in students. Until recently, modeling the assessed activity types has been the main focus of these supervised sequence learning KT methods, and these models overlook essential aspects of the student learning process by ignoring the non-assessed learning activities.

Indeed, research shows that non-assessed learning activities can help students to learn better [11], [12]. But, the realization and attainment of the gained knowledge from the assessed and non-assessed learning materials can be different. For example, Hou et al. conclude that practice activities are useful for student success in projects, but they do not help as much in exam preparation [13]. Instead, they show that reviewing practice quizzes could help for exam preparation. In other words, the knowledge that is gained from one learning material type (e.g., video lectures) can be transferred to another (e.g., solving problems). However, the dynamics and realization of this transfer depend on the transition order of learning materials. For example, consider a student who is learning about “summation” and “multiplication” concepts by watching video lectures and practicing problems. Since the multiplication concept can be explained as an extension of summation, the student can learn multiplication easier from a video lecture after practicing summation problems. Meaning that the summation knowledge that is gained by solving problems can be transferred to help achieve better multiplication knowledge using video lectures. However, the reverse sequence may not be as helpful. A student may not be able to solve summation problems just by watching multiplication videos if they do not have background knowledge in summation.

As a result, explicitly modeling knowledge transfer between different learning material types, particularly both assessed and non-assessed ones, is essential to accurately understand student learning processes. Recently, a handful of works have sought to model both assessed and non-assessed learning activities [14]–[16]. However, none of these approaches model how student knowledge transfers from one learning activity type to another. In this paper, we propose Transition-Aware Multi-activity Knowledge Tracing (TAMKOT) to explicitly model knowledge transfer every time a student transitions between different learning activity types. TAMKOT models student

knowledge states in a set of latent variables at every step in the student learning sequence. Every time a student transitions from one learning material to another, TAMKOT uses a transition-specific matrix to transfer the student's knowledge according to the type of involved learning activities. Unlike previous KT models, our formulation allows for unlimited transitions between different learning activity types and does not limit sequence lengths for any of the material types. This is realized via the simple, yet efficient, formulation of transition identifiers in TAMKOT that activate one transition-specific matrix at a time. Our model provides the flexibility for different material types to have different latent representation spaces that are mapped to a shared student knowledge space.

We evaluate TAMKOT on three real-world datasets. The experiments show that TAMKOT performs significantly better than state-of-the-art supervised knowledge tracing models in predicting student performance. Furthermore, despite its simplicity, TAMKOT performs better than the existing multi-activity knowledge tracing models in datasets with granular learning materials. More importantly, our analysis demonstrates that knowledge transfer can be different depending on the transition order between learning material types, especially in complex learning materials. Finally, we showcase the interpretability of the learned student knowledge states.

II. RELATED WORK

In general, our work relates to the problem of KT, which quantifies a student's knowledge state at each learning time step based on the student's sequence of previous learning activities. This quantification of student knowledge is usually defined over a set of knowledge components or concepts of the learning materials. For example, an Algebra problem can include knowledge components related to the Summation and Multiplication concepts and the students learn those concepts by attempting that problem. Traditionally, many KT models, such as Bayesian knowledge tracing (BKT) [17], and regression-based KT models [2], [18]–[20], rely on a *predefined mapping* between the learning materials and knowledge components as their input. This mapping could be achieved by human experts labeling the learning materials. However, while labeling the concepts for each learning material can be done in smaller educational systems, it is not feasible for large-scale systems that present thousands of learning materials to students.

Later, some KT models were designed to avoid requiring such a predefined mapping. These KT models learn the underlying *latent concepts* presented in the learning materials for this purpose [7], [8], [21]. This includes state-of-the-art deep learning-based KT methods. For example, Zhang et al. proposed dynamic key-value memory network (DKVMN) that uses a variant of memory augmented neural networks for modeling the underlying learning material concepts and student knowledge state [7]. DKT, the first deep learning-based KT method, utilizes the Recurrent Neural Networks (RNN) to model student knowledge state [1]. Later, attention mechanisms have been used to improve KT models. For

instance, Pandey et al. introduced SAKT, an attention-based method that models the interdependencies among student interactions at each attempt [22]. As another example, AKT is an attentive method that learns context-aware question and student performance representations and utilizes the attention mechanism to model student knowledge state [23].

However, all the approaches mentioned above ignore students' non-assessed learning activities and only consider the assessed learning materials. In a few recent approaches, non-assessed learning activities are used as extra features when modeling assessed learning activities [24], [25]. While these models successfully improve student performance prediction, they do not explicitly represent student knowledge gain and knowledge state when interacting with non-assessed learning materials. To the best of our knowledge, MA-Elo [15], MA-FM [26], MVKM [14], and DMKT [16] are the only *multi-activity KT models* that explicitly model student knowledge state from multiple types of learning activities. MA-Elo is a multivariate Elo-based [27] learner model that adjusts student knowledge state in a predefined set of knowledge concepts according to the difference between the student's predicted and observed next activity [15]. MA-FM is based on factorization machines and models student knowledge state in predefined knowledge concepts by a weighted count of the student's previous successes, failures, and activities with different learning material types [26]. Both MA-FM and MA-Elo require the predefined mapping between the learning materials and concepts, and cannot be used in systems that do not include this information. MVKM and DMKT are the only multi-activity models that do not require this information. MVKM models student learning activities of different material types as separate tensors and uses tensor factorization to capture latent students' features, student knowledge, and learning material latent concepts [14]. DMKT extends a KT model for assessed activities, DKVMN [7], and uses key-value memory networks to model student knowledge gain over both assessed and non-assessed learning activities [16]. Since DMKT has a fixed architecture that does not adjust to the learning activity types, it can only model a fixed number of non-assessed activities between every two assessed ones. None of the above methods explicitly represent the dynamics of knowledge transfer among different learning activities. Moreover, MVKM and DMKT assume the exact same latent concept spaces for different types of materials, while different types of learning materials could have different knowledge concepts.

III. PROBLEM FORMULATION

KT is usually evaluated by the task of student performance prediction, where students' upcoming performances are predicted, given their past learning activities. KT methods predominantly focus on assessed learning activities as students' past activities. Specifically, a student's interaction at each time step t is denoted as (q_t, r_t) , where q_t represents the assessed learning material (e.g., problem) that the student interacts with at the time step t , and r_t denotes the student's performance (e.g., score, correctness, or grade) in q_t . Given the previous

performance records of a student as $\{(q_1, r_1), \dots, (q_t, r_t)\}$, KT aims to predict the student's future performance r_{t+1} in a problem q_{t+1} at time step $t + 1$.

Our goal is to trace students' knowledge at each time step t as they learn from both assessed and non-assessed learning material types, explicitly model the knowledge transfer from each learning material type to another, and predict student performance on future assessed learning materials. Additionally, we would like the model to represent unlimited student transitions between different activity types with no particular order. Without loss of generality, assume an education system with one assessed learning material type (e.g., problems) and one non-assessed learning material type (e.g., video lectures). Each student only interacts with one learning material, either an assessed or a non-assessed one, at each time step t . We represent student activity type using an indicator $d_t \in \{0, 1\}$, where 0 represents the assessed learning material type, and 1 represents the non-assessed type. We also denote a student's activity at each time step t as a tuple $\langle i_t, d_t \rangle$, where

$$i_t = \begin{cases} (q_t, r_t) & \text{if } d_t = 0 \\ l_t & \text{if } d_t = 1 \end{cases}$$

Here (q_t, r_t) shows that the student interacts with the problem q_t at time step t with performance r_t , and l_t represents the video lecture that the student watches at time step t . This formulation allows us to represent student learning activities with both learning material types. Eventually, we represent a student's whole trajectory of activities with different learning materials types as a sequence of tuples $\{\langle i_1, d_1 \rangle, \dots, \langle i_t, d_t \rangle\}$.

To achieve our goal of predicting student performance, given their assessed and non-assessed learning activity history, we assume that students gain knowledge in a set of latent concepts or topics that are presented in learning materials. However, the realization of student knowledge can vary in different material types. We also assume that the knowledge gained using one learning material can be transferred to another learning material when students switch between them.

IV. TAMKOT MODEL

In this section, we introduce our model Transition-Aware Multi-view Knowledge Tracing (TAMKOT). We build TAMKOT into three layers: (1) the embedding layer that maps each learning activity to the latent embedding space, (2) the hidden layer to model and transfer the knowledge between assessed and non-assessed interactions at each time step, and (3) the prediction layer to predict student's performance on an upcoming assessed learning material. We formulate TAMKOT by building a transition-aware multi-activity component on top of LSTM [28]. An overview of TAMKOT's architecture is presented in Figure 1. In the following, we introduce the details of each layer.

Notations. We use lowercase letters to denote scalars, e.g., q_t ; boldface capital letters for matrices, e.g., $\mathbf{A}^q$; and boldface lowercase letters for vectors, e.g., $\mathbf{q}_t$.

A. Embedding Layer

The goal of this layer is to learn the embedding vector of each learning activity $\langle i_t, d_t \rangle$ as the input to hidden knowledge transfer layer for estimating the student's knowledge hidden state $\mathbf{h}_t$, using the latent representation of its learning material (q_t and l_t) and student response (r_t). The few existing multi-activity KT methods model both assessed and non-assessed learning materials in the same latent space with the same dimensionality [14], [16]. Unlike these works, we assume that the assessed and non-assessed learning materials can have different latent spaces. This allows TAMKOT to be flexible in having more (or less) fine-grained representation for each learning material type. Having problems as assessed and video lectures as non-assessed learning materials, we first map all problems into the problem latent space and video lectures into the video latent space and achieve their underlying latent concepts matrices $\mathbf{A}^q \in \mathbb{R}^{Q \times d_q}$ (for problems) and $\mathbf{A}^l \in \mathbb{R}^{L \times d_l}$ (for video lectures). Here, Q and L are the number of questions and video lectures, and d_q and d_l are latent concept sizes for questions and video lectures respectively. For student performance r_t in assessed learning materials, we use another embedding matrix $\mathbf{A}^r$ that maps student performance into the latent space. When modeling binary student performance outcomes (e.g., success or failure in solving problems), $\mathbf{A}^r \in \mathbb{R}^{2 \times d_r}$, where d_r is the performance embedding size. For modeling numerical performance outcomes (e.g., exam scores between 0 and 1), we use a linear mapping $f(r_t) = r_t \mathbf{A}^r$ that maps the numerical performance into higher dimension, and $\mathbf{A}^r \in \mathbb{R}^{d_r}$.

At each time step t , TAMKOT looks up latent learning material and student performance representations for the learning activity $\langle i_t, d_t \rangle$ to create its embedding vector. For the problem activity $i_t = (q_t, r_t)$, it looks up latent representation $\mathbf{q}_t \in \mathbb{R}^{d_q}$ for the problem q_t , and $\mathbf{r}_t \in \mathbb{R}^{d_r}$ for the student performance outcome r_t . It then concatenates them as $\mathbf{x}_t = [\mathbf{q}_t \oplus \mathbf{r}_t]$ to create the activity i_t 's embedding. For the video lecture activity $i_t = l_t$, it looks up lecture l_t 's latent representation $\mathbf{l}_t \in \mathbb{R}^{d_l}$ as activity i_t 's embedding.

B. Hidden Knowledge Transfer Layer

The hidden knowledge transfer layer is designed to represent the student's knowledge state $\mathbf{h}_t$ and learn knowledge transfer while they are freely interacting with and transitioning between assessed and non-assessed learning material types. Similar to LSTM, TAMKOT is composed of a memory cell, an input gate, an output gate, and a forget gate. However, unlike LSTM which is invariant to activity types, TAMKOT models various activity types and their transitions by considering the current and previous activity type as an extra input, and adopting the internal gate formulations to appropriate activity type transitions. This results in a different formulation for each of TAMKOT's gate compared to LSTM and provides an explicit between-type knowledge transfer model.

In particular, we assume a different knowledge transfer pattern for each transition between learning material types. For each of these transitions, we propose a set of indicators

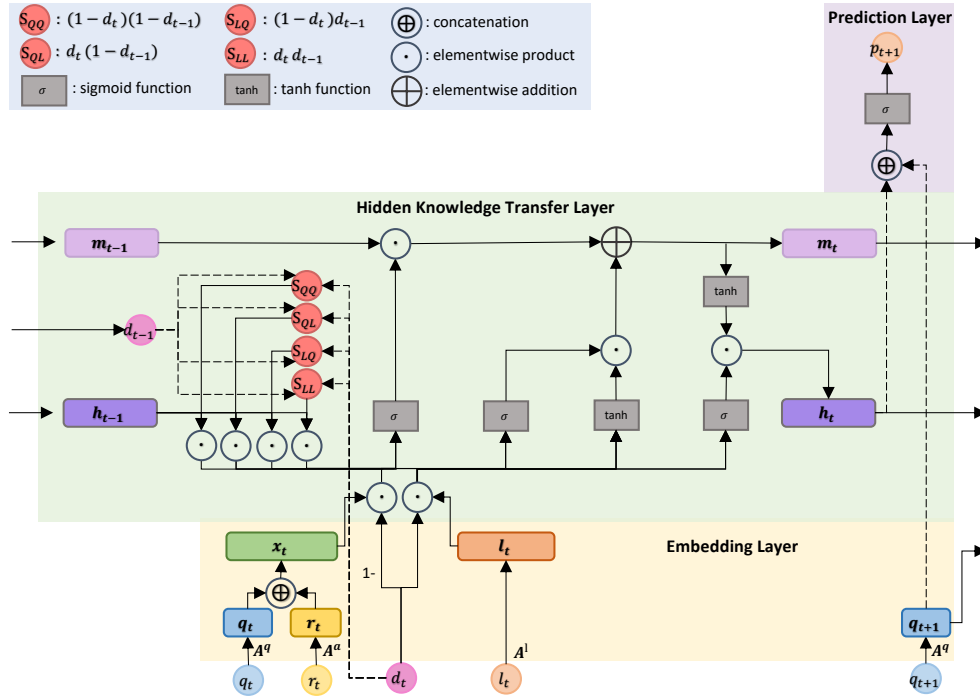


Fig. 1. The TAMKOT model architecture. The solid lines and dashed lines are the same. Different line types are used to keep clarity between the lines that fall over/cross each other.

and formulate the gate updates according to these indicators. For example, assuming video lectures (denoted by “L”) and problems (denoted by “Q”) as two material types, we have four different transitions between problems and video lectures: problems to problems (QQ), problems to video lectures (QL), video lectures to problems (LQ), and video lectures to video lectures (LL). Consequently, we denote the four permutation indication variables at each time step t according to the learning material type indicators d_t and d_{t-1} :

$$s_{QQ} = (1 - d_t)(1 - d_{t-1}) \quad (1)$$

$$s_{QL} = d_t(1 - d_{t-1}) \quad (2)$$

$$s_{LQ} = (1 - d_t)d_{t-1} \quad (3)$$

$$s_{LL} = d_t d_{t-1} \quad (4)$$

where s_{QQ} , s_{QL} , s_{LQ} , and $s_{LL} \in \{0, 1\}$ indicate four transition permutations of learning material type from time step $t-1$ to t . For example, $s_{QL} = 1$ indicates the student has switched from attempting a problem at time $t-1$ to watching a video lecture at time t . As a result of this formulation, at each time step t , only one of four permutation indication variables is equal to 1, with the rest of them being 0.

We use a vector $\mathbf{h}_t \in \mathbb{R}^{d_h}$ to keep track of student knowledge state at time step t , where d_h is the hidden dimension size. At each step, we update $\mathbf{h}_t$ according to the previous state $\mathbf{h}_{t-1}$ and the embedding vector of activity that the student has attempted ($\mathbf{x}_t$ and $\mathbf{l}_t$). Additionally, since we assume that the transition order between activity types are important in how students learn, we update $\mathbf{h}_t$ according to the transition permutations s_{**} defined above. To represent how the knowledge transfers between activity types,

we use transition-specific weight matrices (indicated by $\mathbf{W}$ s) to update the student state. Accordingly, at each time step t , TAMKOT updates $\mathbf{h}_t$ by as follows:

$$\mathbf{i}_t = \sigma \left((1 - d_t) \cdot \mathbf{x}_t \mathbf{V}_{iQ} + d_t \cdot \mathbf{l}_t \mathbf{V}_{iL} + s_{QQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{i_{QQ}} + s_{LL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{i_{LL}} + s_{QL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{i_{QL}} + s_{LQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{i_{LQ}} + \mathbf{b}_i \right) \quad (5)$$

$$\mathbf{g}_t = \tanh \left((1 - d_t) \cdot \mathbf{x}_t \mathbf{V}_{gQ} + d_t \cdot \mathbf{l}_t \mathbf{V}_{gL} + s_{QQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{g_{QQ}} + s_{LL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{g_{LL}} + s_{QL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{g_{QL}} + s_{LQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{g_{LQ}} + \mathbf{b}_g \right) \quad (6)$$

$$\mathbf{f}_t = \sigma \left((1 - d_t) \cdot \mathbf{x}_t \mathbf{V}_{fQ} + d_t \cdot \mathbf{l}_t \mathbf{V}_{fL} + s_{QQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{f_{QQ}} + s_{LL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{f_{LL}} + s_{QL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{f_{QL}} + s_{LQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{f_{LQ}} + \mathbf{b}_f \right) \quad (7)$$

$$\mathbf{o}_t = \sigma \left((1 - d_t) \cdot \mathbf{x}_t \mathbf{V}_{oQ} + d_t \cdot \mathbf{l}_t \mathbf{V}_{oL} + s_{QQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{o_{QQ}} + s_{LL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{o_{LL}} + s_{QL} \cdot \mathbf{h}_{t-1} \mathbf{W}_{o_{QL}} + s_{LQ} \cdot \mathbf{h}_{t-1} \mathbf{W}_{o_{LQ}} + \mathbf{b}_o \right) \quad (8)$$

$$\mathbf{m}_t = \mathbf{f}_t \cdot \mathbf{m}_{t-1} + \mathbf{i}_t \cdot \mathbf{g}_t \quad (9)$$

$$\mathbf{h}_t = \mathbf{o}_t \cdot \tanh(\mathbf{m}_t) \quad (10)$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$ represent the input gate, forget gate, and output gate, and $\mathbf{g}_t$ is the candidate memory cell, and σ is the sigmoid function. In Equations 5 to 8, gates are calculated according to the learning material type transitions in the student sequence. Since knowledge transfer can be different for the four possible transitions, we consider separate transfer weight matrices for them. So, in each gate and cell, $\mathbf{W}_{*_{QQ}}$, $\mathbf{W}_{*_{LL}}$, $\mathbf{W}_{*_{QL}}$, and $\mathbf{W}_{*_{LQ}} \in \mathbb{R}^{d_h \times d_h}$ are knowledge transfer weight matrices that are associated with the four different possible

transition permutations s_{**} . For example, W_{fLQ} captures the knowledge transfer from previous student knowledge state h_{t-1} to the current state in the forget gate, when a student switches from watching video lectures to solving problems. Also, $V_{*Q} \in \mathbb{R}^{(d_q+d_a) \times d_h}$ and $V_{*L} \in \mathbb{R}^{d_l \times d_h}$ are used to map embeddings of problem activity and lecture activity, respectively, to gates and cell in the hidden knowledge layer. V_{*Q} and V_{*L} are activated according to the current learning material type (d_t). $b_* \in \mathbb{R}^{d_h}$ are the bias terms.

Based on our formulation in TAMKOT, at each time step, one learning activity type transition, and consequently one knowledge transfer weight is activated. While we consider two learning material types in this paper, extending this formulation to more than two types would be trivial. Unlike previous attempts at multi-activity knowledge tracing that allowed for a limited number of non-assessed learning materials between every two assessed ones [16], this representation allows us to model unlimited transitions in any order between two learning material types in student sequence. Additionally, unlike sequential multitask learning models that need the same sequence length for all views [29], [30], this representation does not need one-to-one sequence alignment and allows for different sequence lengths for assessed and non-assessed learning activities.

C. Prediction Layer

In this layer, TAMKOT predicts the target student's performance for a given problem q_{t+1} at the next time step $t+1$, according to the student's past learning activities, summarized in the current hidden state h_t . This is achieved by concatenating the hidden state h_t with the next candidate problem's embedding vector q_{t+1} , and passing the concatenation into a fully connected layer with the sigmoid activation function:

$$p_{t+1} = \sigma(W_p^T [h_t \oplus q_{t+1}] + b_p) \quad (11)$$

where the prediction p_{t+1} represents the probability that the student answers the next problem q_{t+1} correctly, $W_p \in \mathbb{R}^{d_q+d_h}$ is the weight matrix, and $b_p \in \mathbb{R}$ is the bias term.

D. Objective Function

We learn the parameters of TAMKOT by minimizing the following regularized binary cross-entropy loss.

$$\mathcal{L} = - \sum_t (r_t \log p_t + (1 - r_t) \log (1 - p_t)) + \lambda_\theta \|\theta\|^2 \quad (12)$$

where r_t represents the actual student performance and θ denotes all the learnable parameters of TAMKOT: the embedding matrices A^q , A^l , and A^r , the weight matrices V_{*Q} , V_{*L} , W_{*QQ} , W_{*LL} , W_{*QL} , and W_{*LQ} , and the bias terms b_* . $\|\theta\|^2$ is the regularization term and λ_θ is the hyperparameter to specify regularization weight.

V. EXPERIMENTS

We evaluate TAMKOT with three sets of experiments. First, we evaluate TAMKOT in the student performance prediction task. Second, to analyze the knowledge transfer between assessed and non-assessed learning material types, we compare

the transition matrices for different transitions. Third, we perform a case study to visualize student knowledge states. Our code and example datasets are available on GitHub ¹.

A. Datasets

We use three real-world datasets in our study. The general statistics of each dataset can be found in Table I.

MORF² [31]: This is a dataset of one online course, available via the MOOC Replication Framework (MORF), from Coursera³. The course subject is 'educational data mining' and is divided into different modules. Each module is associated with a topic, such as 'classification'. We use video lectures (non-assessed), and assignments (assessed) as two learning material types. In each module that is planned for a week, the students need to watch five to seven video lectures and work on one assignment. Each assignment usually contains more than one problem. Only coarse-grained assignment-level data is available, as if the students submit an entire assignment each time rather than submitting a single problem. We treat each student submission of assignment as one assessed activity and the overall score of each submission as the response to this activity.

EdNet⁴ [32]: This is a dataset from a multi-platform AI tutoring service (Santa ⁵) for Korean students to practice while preparing for TOEIC ⁶ English testing. Ednet offers four different levels of data to provide various kinds of actions in a consistent and organized manner. Data from the third level is selected to evaluate our model, which consists of student learning activities in multiple learning material types. During the student practice, the platform recommends questions to students. But, the students can decide to follow the recommendations or not. Each question has a problem explanation that the students can choose to read. We randomly sample 1000 students who interacted with both questions (assessed) and their associated problem explanations (non-assessed) in this dataset.

Junyi⁷ [33]: This dataset is collected from a Chinese e-learning website to teach students math. Students work on studying eight math areas with different difficulty levels. They start from the easiest level and are moved to the more difficult levels as they learn. We use the preprocessed data introduced in [34]. In this dataset, problems (assessed), and hints (non-assessed) are used as two learning material types. During the practice, students have the option to request hints for solving the problems, and each problem may be associated with more than one hint.

B. Baseline Methods

We utilize six state-of-art assessed-only supervised KT models and two multi-activity KT models as original baselines

¹<https://github.com/persai-lab/BigData2022-TAMKOT>

²<https://educational-technology-collective.github.io/morf/>

³<https://www.coursera.org/>

⁴<https://github.com/riid/ednet>

⁵<https://www.aitutorsanta.com/>

⁶<https://www.ets.org/toeic>

⁷<https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1275>

TABLE I
DESCRIPTIVE STATISTICS OF 3 DATASETS.

Dataset	#Users	#Questions	Question Records	Question Responses Mean	Question Responses STD	#Correct Question Responses	#Incorrect Question Responses	#Non-assessed materials	#Non-assessed Records
MORF	686	10	12031	0.7763	0.2507	N/A	N/A	52	41980
EdNet	1000	11249	200931	0.5910	0.2417	118747	82184	8324	150821
Junyi	2063	3760	290754	0.6660	0.2224	193664	97090	1432	69050

to evaluate our proposed method. To provide a fair comparison, we also extend the six assessed-only supervised KT models to be able to consider both assessed and non-assessed learning material types and used them as our baselines. In addition, we also extend the simple multi-layer (MLP) perceptron as another baseline that can incorporate both assessed and non-assessed learning activities as the input. These baselines are identified by a “+M” at the end of their names. In total, we compare our TAMKOT with 14 baselines. Seven of the eight original baselines are based on deep learning, one is a tensor factorization model. The assessed supervised KT baselines are:

- DKT [1] is the first deep learning based KT model that uses recurrent neural networks to model student knowledge gain.
- DKVMN [7] uses memory augmented neural networks to model KT, with one static key matrix for the knowledge concepts and a dynamic value matrix for updating student mastery levels.
- DeepIRT [35] is an extension of DKVMN that integrates the one-parameter logistic item response theory to address overfitting and provide better interpretation.
- SAKT [22] applies the self-attentive model for KT to model the relationship between interactions at different time steps.
- SAINT [25] is the first encoder-decoder model for KT that applies deep self-attentive layers to exercises and responses separately.
- AKT [23] is a context-aware model that uses a monotonic attention mechanism to summarize past student performances that are relevant to the current problem.

In addition to the above, we compare TAMKOT with the following models that support both assessed and non-assessed learning material types:

- MVKM [14] is a multi-view tensor factorization method that explicitly models student knowledge acquisition from multi-type learning activities. It builds separate tensors for students’ activities from each learning material type, and cannot explicitly model knowledge transition between material types.
- DMKT [16] also explicitly models students’ knowledge gain from both assessed and non-assessed activities. It is based on DKVMN and models different read and write operations for assessed and non-assessed learning material types. However, it does not explicitly model knowledge transfer between assessed and non-assessed learning materials. Additionally, it only allows for a fixed number of non-assessed learning activities between every two assessed ones. As a result, it is not flexible to capture the full student sequence and switches between learning material types.
- MLP+M is a simple multi-layer perceptron that considers

a student’s three recent assessed interactions along with three non-assessed interactions as the input and predicts the probability of student mastery level.

•DKT+M [24] and DKVMN+M are extensions of DKT and DKVMN to consider non-assessed learning activities in addition to the assessed ones. They concatenate embedding vectors of all non-assessed learning materials that the student had interacted with between every two assessed activities as an additional feature, with the problem embedding as input for vanilla DKT and DKVMN.

•SAINT+M [25], SAKT+M, and AKT+M are variants of SAINT, SAKT, and AKT. Similar to DKT+M, in these extended models, all the non-assessed learning materials’ embeddings that happen between two assessed activities are summarized as an additional feature. In addition, the position encoding is added to each learning material embedding.

Notably, DKT+M is the closest approach to an ablated version of TAMKOT which does not include the knowledge transfer component and ignores the knowledge transition between non-assessed learning activities.

TABLE II
LEARNED BEST HYPERPARAMETERS OF TAMKOT

Dataset	d_q	d_l	d_a	d_h	L_s	λ_θ
MORF	64	16	64	16	100	0.01
EdNet	64	32	64	16	50	0.01
Junyi	32	32	32	32	100	0.05

C. Experiment Setup

We use 5-fold student stratified cross-validation to split the training set and test set. At each fold, sequences from 80% of the students are used as the training set, and the sequences from the rest 20% of students are used as the testing set. For hyperparameter tuning, we separate another 20% of students from the training set as the validation set. We implement our proposed methods with PyTorch⁸ and use the Adam optimizer to learn the model parameters. All parameters are randomly initialized with the Gaussian distribution with 0 mean and 0.2 standard deviation. We use the norm clipping threshold to avoid gradient exploding. Following the standard KT experiments [1], we truncate or pad all the sequences to the same length. Sequence length (L_s) is treated as another hyperparameter fine-tuned using the validation data. For sequences longer than L_s , we truncate it into multiple sequences. For sequences shorter than L_s , we pad them to length L_s with 0. We use coarse-grained grid search to find the best hyperparameters (reported in table II).

⁸<https://pytorch.org/>

D. Student Performance Prediction

Here, we evaluate TAMKOT on the task of student performance prediction with the baselines introduced in Section V-B. We report average results across the five folds, as well as t-test p-values compared with the proposed model TAMKOT. In EdNet and Junyi datasets, student responses are binary (success or failure). So, we use Area Under Curve (AUC) to evaluate model performances. Higher AUC represents better prediction performance. In the MORF dataset, assignments are graded using a numeric value. We normalize students' assignment scores in range $[0, 1]$ with the maximum possible score for the assignment as student performance. Root Mean Squared Error (RMSE) is used to evaluate the prediction performance of the MORF dataset. Lower RMSE accounts for better prediction performance. Experiments results are presented in table III. Since MVKM is lacking in handling high-dimensional datasets with a high computation time, we only run MVKM on the MORF dataset.

TABLE III
STUDENT PERFORMANCE PREDICTION RESULTS. THE BEST AND SECOND-BEST RESULT ARE IN BOLDFACE AND UNDERLINE, RESPECTIVELY. ** AND * INDICATE PAIRED T-TEST p -value < 0.05 AND p -value < 0.1 , RESPECTIVELY, COMPARED TO TAMKOT.

Methods	MORF	EdNet	Junyi
	RMSE	AUC	AUC
DKT	0.1938*	0.6393**	0.8623**
DKVMN	0.2043**	0.6296**	0.8558**
SAKT	0.2113**	0.6334**	0.8053**
SAINT	0.2019**	0.5205**	0.7951**
AKT	0.2420**	0.6393**	0.8093**
DeepIRT	0.1946**	0.6290**	0.8498**
DKT+M	0.1928	0.6372**	<u>0.8652*</u>
DKVMN+M	0.2251**	0.6343**	0.8513**
SAKT+M	0.2085**	0.6323**	0.7911**
SAINT+M	0.1977**	0.5491**	0.7741**
AKT+M	0.2240**	<u>0.6404**</u>	0.8099**
MLP+M	0.2433**	0.6102**	0.7290**
MVKM	0.1936*	—	—
DMKT	0.1754**	0.6394**	0.8561**
TAMKOT	<u>0.1871</u>	0.6786	0.8745

We see that TAMKOT significantly outperforms all the six supervised assessed KT models in all datasets. This shows that *TAMKOT can successfully model the non-assessed student activities along with the assessed ones, to leverage their added information for improving the performance predictions*. The other two explicit multi-activity KT models MVKM, and DMKT mostly achieve higher prediction performance compared to the assessed-only methods. This shows that explicitly modeling non-assessed learning activities could help improve KT. But, we also see that DKT and AKT perform better than or similar to DMKT in EdNet and Junyi datasets.

One potential reason for this observation could be the difference between MORF and the other two datasets. The learning materials in the MORF dataset are more complex, compared to EdNet and Junyi. While in EdNet and Junyi problems are granular and focused on specific topics, each

TABLE IV
SPEARMAN CORRELATION COEFFICIENTS WITH P-VALUES BETWEEN $\mathbf{W}_{f_{QL}}$ AND $\mathbf{W}_{f_{LQ}}$

	MORF	EdNet	Junyi
Correlation	-0.03686	0.33680	0.38443
p-value	0.55714	3.30e-08	2.09e-37

MORF assignment includes multiple problems, each of which covers multiple concepts. We note that DMKT's structure is also more complex than DKT and AKT. While DKT and AKT use a vector representation $\mathbf{h}_t$ for student state, DMKT uses a complex key-value memory matrix representation for learning material and student knowledge. Accordingly, we hypothesize that DMKT's better performance in MORF dataset can be attributed to a better match of DMKT's complexity with MORF's material complexity. At the same time, this complexity may not be necessary for the EdNet and Junyi datasets with simpler learning material structures.

Comparing TAMKOT with the six multi-activity versions of the assessed KT models (the "+M" methods), we see that TAMKOT significantly outperforms all of them in all datasets. Particularly, one can consider the LSTM-based DKT+M as a simpler version of TAMKOT without explicitly modeling transitions between activity types. TAMKOT significantly outperforms DKT+M in all datasets. This shows that *merely concatenating the assessed activity sequences with the non-assessed ones is not enough*.

In fact, we find that *simply incorporating the non-assessed learning material as additional features can sometimes harm the prediction performance*. For example, the results of DKVMN+M are worse than DKVMN on all the datasets, DKT+M performs worse than DKT in the EdNet dataset, and SAKT+M, SAINT+M, and AKT+M are worse than SAKT, SAINT, and AKT respectively in the Junyi dataset.

Comparing TAMKOT with the two multi-activity baselines, we see that it outperforms both of them in EdNet and Junyi datasets. This shows that, *modeling knowledge transfer and activity transitions, is essential in multi-activity knowledge modeling in these datasets*. In the MORF dataset, TAMKOT is the second-best after DMKT. We hypothesize that this happens because of the MORF learning material complexity reason explained above. Similar to DKT and AKT, TAMKOT uses a simple vector-representation $\mathbf{h}_t$ for student state.

Overall, explicitly modeling both assessed and non-assessed activities, in addition to the transition-aware knowledge transfers between them, is shown to be necessary to accurately represent student knowledge and predict their performance.

E. Knowledge Transfer Analysis

Here we analyze the learned knowledge transfer between assessed and non-assessed learning activity types. Particularly, we study if the knowledge transfer from the assessed learning materials to the non-assessed ones is different from the knowledge transfer from the non-assessed learning materials to the assessed ones. We first inspect the transition weight matrices of the forget gate $\mathbf{W}_{f_{QL}}$ (assessed to non-assessed) and $\mathbf{W}_{f_{LQ}}$

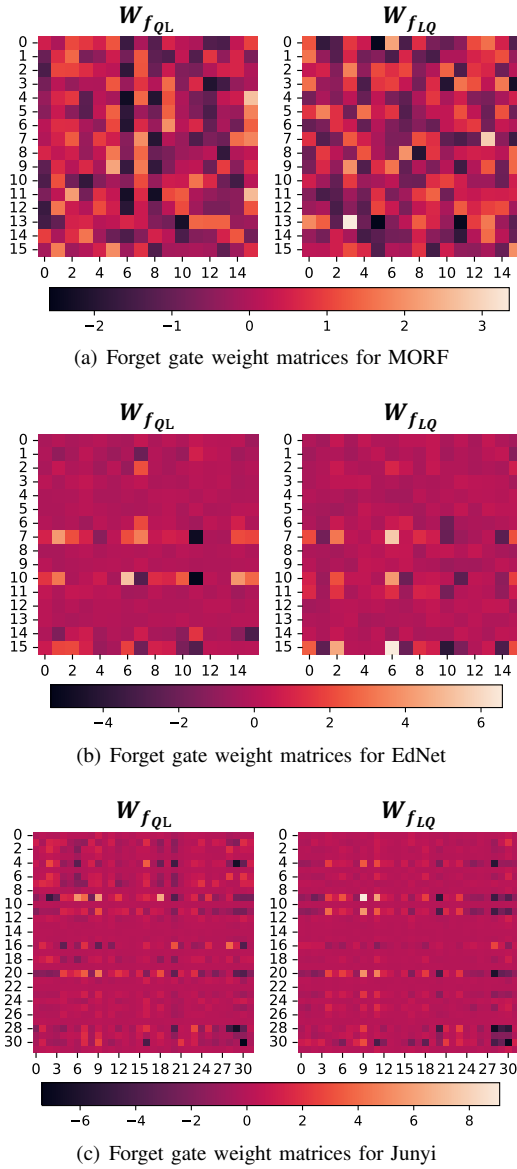


Fig. 2. Heatmaps for comparison of the weight matrices $\mathbf{W}_{f_{QL}}$ (assessed to non-assessed type) and $\mathbf{W}_{f_{LQ}}$ (non-assessed to assessed type) of forget gate for each dataset.

(non-assessed to assessed) in equation 7. Each cell in these matrices represents the knowledge transfer weight between a latent concept to another latent concept when the student transfers from one learning material type to another. Since the weight values can have different ranges in the two weight matrices, we use a rank-based metric for their comparison.

Specifically, to compare $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$, we first flatten them and then run a Wilcoxon signed-rank test [36] on them. This test uses the median pairwise rank difference to identify if the rankings of between-concept transition weights are the same in the two matrices. With a small p-value of 0.02 for the MORF dataset, we can conclude that the median pairwise difference between $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$ is non-zero, and these transition weights are significantly different in the MORF dataset. This means that there are latent-concept pairs

in MORF that can easily transfer to each other when the student transitions from assignments to video lectures (or video lectures to assignments), but they cannot transfer as easily when the students transition in a reverse order. On the other hand, the p-values for EdNet and Junyi datasets are large (> 0.7). Therefore, we cannot reject the Null hypothesis and conclude that transition weights in $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$ are different in EdNet and Junyi.

As a second measure, we calculate the Spearman correlation coefficient [37] between the flattened $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$ (reported in Table IV). As we can see, for EdNet and Junyi, $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$ are positively correlated with a small p-value. So, a higher (lower) transfer weight from questions to problem explanations in a specific latent concept pair in EdNet usually means a higher (lower) transfer weight from problem explanations to questions in the same concept pair. But for MORF, the correlation coefficient is close to 0. This is in accordance with our previous conclusion that the transition weight matrices are different in MORF.

To further investigate, we visualize the weight matrices $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$. For better visualization, we first perform a z-score normalization [38] on each weight matrix so that it has a mean of 0 and a standard deviation of 1. Then, we plot the heatmaps of these normalized transition matrices for all three datasets and show them in figure 2. As we can see, in the MORF dataset, the weight matrices are substantially different from each other, with a few similarities. For example, the values for latent dimensions row 8 to column 7 are close to each other (1.4571 in $\mathbf{W}_{f_{QL}}$ and 2.0010 in $\mathbf{W}_{f_{LQ}}$). This indicates that the between-concept knowledge transfer from assignments to video lectures is different from the knowledge transfer from video lectures to assignments.

In contrast, the transition weights in $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$ are relatively similarly distributed in the Junyi and EdNet datasets. However, there are still some small differences that can be observed. For example, the weights in Junyi are different from latent dimension row 7 to latent dimension row 14: the value is 2.0514 in $\mathbf{W}_{f_{QL}}$, while it is -2.5527 in $\mathbf{W}_{f_{LQ}}$. This shows that in Junyi and EdNet datasets, most concepts transfer similarly between different activity types. But for a few concept pairs, there are different transfer dynamics.

These observations are consistent with the different dataset characteristics. Unlike MORF, in EdNet and Junyi there are close-knit associations between different learning material types. For Junyi, the two types of materials are problems and hints, and for EdNet they are questions and problem explanations. Each hint in Junyi (and each problem explanation in EdNet) is designed to help a single problem (or question) associated with it. As a result, the transitions usually happen between similar learning materials with related concepts. But, in MORF each assignment includes multiple problems and covers many concepts. Similarly, the video lectures are more general than Junyi's hints and EdNet's explanations and introduce a wider range of topics. As a result, the students can transition between diverse and unrelated concepts. This leads to a more complicated association between MORF's

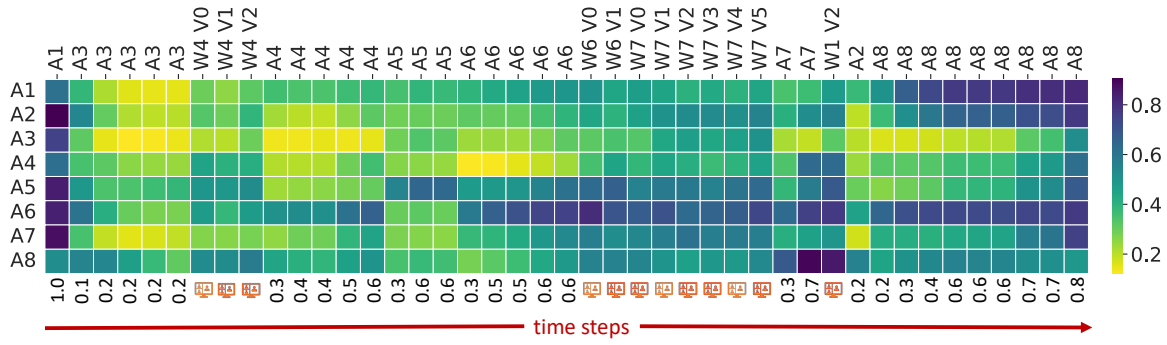


Fig. 3. Visualization of predicted student performance (as a knowledge indicator) for a sample student in the MORF dataset. The top x-axis ticks are learning material titles the student has tried at each time step. The bottom x-axis ticks are real student performance (in assessed activities) or the ‘screen’ icon (in non-assessed ones). The y-axis ticks are the assessed material titles. Each cell shows the student’s predicted performance in an assessed material at each step.

assignments and video lectures, which in turn leads to more complex and dissimilar weight matrices $\mathbf{W}_{f_{QL}}$ and $\mathbf{W}_{f_{LQ}}$.

This analysis shows that knowledge transfer weights could depend on the transition order (permutation) between material types, especially for the datasets in which assessed and non-assessed material types are more complex and are not closely associated with each other. Also, this analysis shows how to interpret knowledge transfer between different learning materials. This could help instructors in arranging course learning materials for the maximum possible knowledge transfer.

F. Student Knowledge State Visualization

To study student knowledge state interpretation, we visualize the student mastery level of each assessed learning material at each time step. In other words, after every activity, we use equation 11 to calculate the student’s predicted performance p_{t+1} in each assessed learning material, as an indicator of student knowledge. For this case study, we base our analysis on the MORF dataset. As Junyi and EdNet have similar patterns as MORF, we don’t show their visualization due to the page limitation. We sample one student’s trajectory from the MORF dataset to visualize (Figure 3). Each row shows the student’s predicted performance in one of the eight course assignments during the learning trajectory. Each column shows the student’s predicted performance in all assignments after attempting a particular learning activity. The title of the attempted learning activity is shown at the top of each column. If the student has attempted a problem, their observed performance is shown at the bottom of each column. If the student has watched a video lecture, a ‘screen’ icon is shown at the bottom of each column. We abbreviate video lecture * of week ** as “W** V*”, and Assignment * as and “A*”. So, “W4 V1” represent the first video of week 4.

We see the student’s initial estimated knowledge is high, after the first interaction. This happens since the student received the full score of 1.0 in the first assignment. But, as we observe 5 low-grade attempts of A3, the student’s mastery level drops. This could be because the student skipped A2 and video lectures for week 2. We then see the student’s knowledge grows by watching video lectures of week 4. However, *watching different videos produce different knowledge improvement*

values for each assignment. For example, after watching week 4’s lectures, although the student’s mastery level of each assignment increased, A4 is the one that has the largest improvement. We also observe an *increase in knowledge as the score of the corresponding assignment increases*. For example, as student scores increases, their knowledge of A5, A6, and A8 also increases. Moreover, when the student watches the multiple lectures between two assignments, the first attempt usually has the largest improvement. It means that *student knowledge does not keep growing while continuously watching multiple lectures one after another*. For example, the student’s knowledge grows significantly after watching W4 V0, but the next two attempts of watching W4 V1 and W4 V2 only improve the student’s knowledge slightly. This conclusion is in line with previous research that shows assessed activities could be more helpful than repeating non-assessed ones [39], [40].

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed Transition-Aware Multi-Activity Knowledge Tracing (TAMKOT), to model student learning from both assessed and non-assessed learning activities and explicitly learn the knowledge transfer between different learning activity types. TAMKOT learns multiple knowledge transfer matrices, one for each transition type between student activities, and allows for unlimited transitions between learning activity types in any order. We performed extensive experiments on three real-world datasets and compared TAMKOT with state-of-the-art baselines in predicting student performance. We also analyzed and interpreted the learned knowledge transfer matrices and student knowledge states. Our experiment results showed that explicitly modeling both assessed and non-assessed activities in TAMKOT, in addition to the transition-aware knowledge transfers between them, is necessary to accurately represent student knowledge and predict their performance. We also concluded that the amount of knowledge transfer between concepts could depend on the transition order (permutation) between activity types, especially for the datasets in which assessed and non-assessed material types are more complex. Finally, we showcased a sample student’s knowledge states and their interpretation that

for that particular student, the assessed activities were more helpful than the non-assessed ones. In the future, we would like to explore TAMKOT's performance in supporting more than two learning activity types and investigate the knowledge transfer among them.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. 2047500.

REFERENCES

- [1] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, and J. Sohl-Dickstein, "Deep knowledge tracing," in *The 28th International Conference on Neural Information Processing Systems - Volume 1*. Cambridge, MA, USA: MIT Press, 2015, p. 505–513.
- [2] F. Drasgow and C. L. Hulin, "Item response theory," *Handbook of Industrial and Organizational Psychology*, pp. 577–636, 1990.
- [3] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon, "Individualized bayesian knowledge tracing models," in *International conference on artificial intelligence in education*. Springer, 2013, pp. 171–180.
- [4] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *Journal of Machine Learning Research*, vol. 15, no. 57, pp. 1959–2008, 2014.
- [5] M. Khajah, Y. Huang, J. González-Brenes, M. Mozer, and P. Brusilovsky, "Integrating knowledge tracing and item response theory: A tale of two frameworks," in *Workshop on Personalization Approaches in Learning Environments*, 2014, pp. 7–12.
- [6] J.-J. Vie and H. Kashima, "Knowledge tracing machines: Factorization machines for knowledge tracing," in *The AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 750–757.
- [7] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *The 26th International Conference on World Wide Web*. New York, NY, USA: ACM, 2017, pp. 765–774.
- [8] S. Sahebi, Y.-R. Lin, and P. Brusilovsky, "Tensor factorization for student modeling and performance prediction in unstructured domain," *The 9th International Conference on Educational Data Mining*, 2016.
- [9] T. Nguyen, "The effectiveness of online learning: Beyond no significant difference and future horizons," *MERLOT Journal of Online Learning and Teaching*, vol. 11, no. 2, pp. 309–319, 2015.
- [10] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.
- [11] A. S. Najar, A. Mitrovic, and B. M. McLaren, "Adaptive support versus alternating worked examples and tutored problems: which leads to better learning?" in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2014, pp. 171–182.
- [12] R. Agrawal, M. Christoforaki, S. Gollapudi, A. Kannan, K. Kenthapadi, and A. Swaminathan, "Mining videos from the web for electronic textbooks," in *International Conference on Formal Concept Analysis*, 2014, pp. 219–234.
- [13] X. Hou, P. F. Carvalho, and K. R. Koedinger, "Drinking our own champagne: Analyzing the impact of learning-by-doing resources in an e-learning course," in *The 11th Int. Conference on Learning Analytics & Knowledge*, 2021.
- [14] S. Zhao, C. Wang, and S. Sahebi, "Modeling knowledge acquisition from multiple learning resource types," in *The 13th International Conference on Educational Data Mining*. The 13th International Conference on Educational Data Mining, 2020, pp. 313–324.
- [15] S. Abdi, H. Khosravi, S. Sadiq, and A. Darvishi, "Open learner models for multi-activity educational systems," in *International Conference on Artificial Intelligence in Education*. Springer, 2021, pp. 11–17.
- [16] C. Wang, S. Zhao, and S. Sahebi, "Learning from non-assessed resources: Deep multi-type knowledge tracing," *The 14th International Conference on Educational Data Mining*, 2021.
- [17] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User modeling and user-adapted interaction*, vol. 4, no. 4, pp. 253–278, 1994.
- [18] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis—a general method for cognitive model evaluation and improvement," in *International conference on intelligent tutoring systems*. Springer, 2006, pp. 164–175.
- [19] —, "Comparing two irt models for conjunctive skills," in *International Conference on Intelligent Tutoring Systems*. Springer, 2008, pp. 796–798.
- [20] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger, "Performance factors analysis—a new alternative to knowledge tracing," *Online Submission*, 2009.
- [21] N. Thai-Nghe, T. Horváth, and L. Schmidt-Thieme, "Factorization models for forecasting student performance," in *Educational Data Mining 2011*. Citeseer, 2010.
- [22] S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," in *The 12th International Conference on Educational Data Mining*, 2019, pp. 384–389.
- [23] A. Ghosh, N. Heffernan, and A. S. Lan, "Context-aware attentive knowledge tracing," in *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2330–2339.
- [24] L. Zhang, X. Xiong, S. Zhao, A. Botelho, and N. T. Heffernan, "Incorporating rich features into deep knowledge tracing," in *The 4th ACM Conference on Learning at Scale*. New York, NY, USA: ACM, 2017, pp. 169–172.
- [25] Y. Choi, Y. Lee, J. Cho, J. Baek, B. Kim, Y. Cha, D. Shin, C. Bae, and J. Heo, "Towards an appropriate query, key, and value computation for knowledge tracing," in *The 7th ACM Conference on Learning at Scale*, 2020, pp. 341–344.
- [26] S. Abdi, "Learner models for learnersourced adaptive educational systems," 2022.
- [27] K. Wauters, P. Desmet, and W. Van Noortgate, "Monitoring learners' proficiency: weight adaptation in the elo rating system," in *Educational Data Mining 2011*, 2011.
- [28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [29] J. Chen, X. Qiu, P. Liu, and X. Huang, "Meta multi-task learning for sequence modeling," in *The AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [30] Q. Liu, S. Wu, D. Wang, Z. Li, and L. Wang, "Context-aware sequential recommendation," in *16th International Conference on Data Mining*, 2016, pp. 1053–1058.
- [31] J. M. L. Andres, R. S. Baker, G. Siemens, D. Gašević, and C. A. Spann, "Replicating 21 findings on student success in online learning," *Technology, Instruction, Cognition, and Learning*, vol. 10, no. 4, pp. 313–333, 2016.
- [32] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim, and J. Heo, "Ednet: A large-scale hierarchical dataset in education," in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 69–73.
- [33] CMU DataShop, "Junyi dataset," <https://pslcdatashop.web.cmu.edu/Project?id=244>, 2015.
- [34] H.-S. Chang, H.-J. Hsu, and K.-T. Chen, "Modeling exercise relationships in e-learning: A unified approach," in *EDM*, 2015, pp. 532–535.
- [35] C. K. Yeung, "Deep-irt: Make deep learning based knowledge tracing explainable using item response theory," in *The 12th Int. Conference on Educational Data Mining*, 2019, pp. 683–686.
- [36] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*. Springer, 1992, pp. 196–202.
- [37] C. Spearman, "The proof and measurement of association between two things," 1961.
- [38] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.
- [39] K. R. Koedinger, J. Kim, J. Z. Jia, E. A. McLaughlin, and N. L. Bier, "Learning is not a spectator sport: Doing is better than watching for learning from a mooc," in *The second ACM conference on learningscale*, 2015, pp. 111–120.
- [40] M. Mirzaei, S. Sahebi, and P. Brusilovsky, "Structure-based discriminative matrix factorization for detecting inefficient learning behaviors," in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2020, pp. 283–290.