Education and Information Technologies (2024) 29:13713–13733
https://doi.org/10.1007/s10639-023-12389-x

Using learners' problem-solving processes in computer-based assessments for enhanced learner modeling: A deep learning approach

Fu Chen[1,2] · **Chang Lu**[3] · **Ying Cui**[4]

Received: 6 June 2023 / Accepted: 26 November 2023 / Published online: 22 December 2023

Abstract

Successful computer-based assessments for learning greatly rely on an effective learner modeling approach to analyze learner data and evaluate learner behaviors. In addition to explicit learning performance (i.e., product data), the process data logged by computer-based assessments provide a treasure trove of information about how learners solve assessment questions. Unfortunately, how to make the best use of both product and process data to sequentially model learning behaviors is still under investigation. This study proposes a novel deep learning-based approach for enhanced learner modeling that can sequentially predict learners' future learning performance (i.e., item responses) based on modeling their history learning behaviors. The evaluation results show that the proposed model outperforms another popular deep learning-based learner model, and process data learning of the model contributes to improved prediction performance. In addition, the model can be used to discover the mapping of items to skills from scratch without prior expert knowledge. Our study showcases how product and process data can be modelled under the same framework for enhanced learner modeling. It offers a novel approach for learning evaluation in the context of computer-based assessments.

Keywords Learner modeling · Collaborative filtering · Deep learning · Process data · Attentive modeling · Computer-based assessment

✉ Fu Chen
fuchen@um.edu.mo

[1] Faculty of Education, University of Macau, Taipa, Macau, China

[2] Institute of Collaborative Innovation, University of Macau, Taipa, Macau, China

[3] School of Education, Shanghai Jiao Tong University, Shanghai, China

[4] Department of Educational Psychology, University of Alberta, Edmonton, Canada

1 Introduction

Analytics of big data in education for enhanced teaching and learning has gained increasing attention over the past years. Advanced by the rapid evolution in information and communication technologies, integrating big data and adaptive learning systems has given rise to a growing personalized learning movement that can tackle conventional educational challenges (Dishon, 2017). Personalized learning benefits learners in multiple ways — learning can be customized, gamified, self-directed, collaborative, and, notably, much more accessible and affordable than traditional learning.

In a personalized learning system, customized learning plans are typically created for learners based on what they know, what they lack, and how they learn best. This requires that data on learning behaviors can be tracked, logged, retrieved, and modelled by digital learning environments. A typical scenario where personalized learning is situated is computer-based assessments (CBAs) for learning, widely used to evaluate and promote learning performance in various learning contexts (Shute & Rahimi, 2017). The popularity and effectiveness of CBAs for personalized learning are attributable to their capacities to evaluate higher-level learner competencies and their flexibility in assessment administration. Moreover, from the data perspective, compared with standardized paper–pencil assessments, CBAs can elicit and collect much more information about how learners perform on and solve each learning task. This enables education practitioners to better evaluate and validate an assessment and to provide learners with finer-grained feedback. Therefore, this study situates the proposed model in the context of CBA for learning.

Making inferences about learners' knowledge states or skill levels based on learners' interactions with learning resources and assessment questions, or learner modeling, is indispensable for an effective CBA. Research on analytics of the two forms of learner data— product and process data—from learners' interactions with CBAs has received increasing attention from communities of educational data mining and educational assessment (Mislevy et al., 2012; Rupp et al., 2012). Product data mainly include final work products interacting with CBAs (e.g., success or failure on assessment tasks and scores of assessment questions). Process data, often represented by log file entries, store the information on learners' problem-solving processes relevant to their final work products (Rupp et al., 2012). Over the past decades, to make the best use of learner product data, tremendous research efforts have been devoted to developing various models and analytic approaches for learner modeling. For example, Bayesian knowledge tracing (BKT; Corbett & Anderson, 1994) is a popular learner model to evaluate and track learners' cognitive states in intelligent tutoring systems (Psotka et al., 1988). In educational measurement, item response theory (IRT; Lord, 1952) models and cognitive diagnosis models (CDM; Tatsuoka, 1990) are two representative families of modern psychometric techniques analyzing learners' item responses to infer their latent skill levels. Despite their popularity, these mainstream approaches are mainly applicable to product data and are limited in

addressing learner process data. Unlike product data, which are often explicit and structured, process data are inherently unstructured with much noise. However, process data is of great potential to reveal a wealth of information on how learners interact with assessments and what contributes to their final working products. As such, in the communities of educational assessment and educational data mining, in recent years, utilizing process data to profile, evaluate and facilitate learning has been an emerging research topic. For example, analytics of process data in the context of CBAs has been used to predict learners' problem-solving outcomes (Chen et al., 2019), probe learners' problem-solving strategies (Greiff et al., 2015), and assess learners' latent skills (Liu et al., 2018). These pioneering studies have shed light on the potential of process data to promote our understanding of how learners approach complex assessment tasks. Nevertheless, the existing approaches are often not generalizable to other CBA settings since they were primarily developed for case studies. There is an urgent need for generic approaches for learner modeling with process data in the context of CBAs.

In recent years, machine learning advances, especially deep learning techniques, have fostered new paradigms of learner data analytics. Machine learning-based approaches for learner modeling are highly scalable and strongly predictive, which greatly benefit large-scale applications of CBAs (e.g., Bergner et al., 2012; Cheng et al., 2019; Lan et al., 2014). Compared with conventional approaches, they are more capable of handling unstructured and incomplete learner data and addressing tremendous amounts of items and learners in large-scale settings. For example, due to their personalized learning nature, most CBAs allow learners access to subsets of assessment items from the item bank. As such, learner data logged by such CBAs are often of large volume and extreme sparseness. To address this, for example, collaborative filtering (CF), a technique widely used for recommender systems, is exceptionally effective for inferring learners' cognitive states or skill levels based on sparse learner data (e.g., Chen et al., 2023). Moreover, to capture a higher degree of complexity of learner data, informed by research from other domains (e.g., He et al., 2017; Zhang et al., 2016), deep learning techniques can be used for enhanced learner modeling through exploiting additional learner and item information or improving the intricacy of model architectures. Unfortunately, to our best knowledge, despite existing deep learning-based approaches (e.g., deep knowledge tracing [DKT]; Piech et al., 2015), theoretical and empirical studies on deep learning approaches for learner modeling applicable to process data in the context of CBAs remain sparse. Considering their modeling flexibility and predictive capacity evidenced by existing studies in other domains, in our study, we attempt to investigate how deep learning-based approaches can be used to address process data for enhanced learner modeling and examine if they are advantageous over conventional approaches.

Specifically, the key objective of the present study is to develop a deep learning-based approach capable of modeling both product and process data for enhanced learner modeling. We attempt to address several specific issues with respect to CBAs for learning in the current study. First, since learners' skill levels improve as they continuously interact with a learning system, the proposed model should address the temporal dependencies between learner-item interactions (i.e., it is a sequential modeling approach). Second, in addition to predicting learners' performance on

unseen items (e.g., item responses to unseen items), the proposed model is expected to be capable of discovering the mapping of items to the targeted latent skills (i.e., item-skill associations). That is, under the assumption that a set of underlying skills affect how learners respond to assessment items, efforts from domain experts in tagging assessment items with skill labels can be less needed if the proposed model can automatically estimate item-skill associations. Finally, given the potential of process data to reflect learners' efforts in attempting assessment questions, the proposed model should be capable of capturing the latent representations of process data to improve the prediction performance. To achieve these, in this study, we proposed a novel deep learning-based approach that can address both product and process data for learner modeling based on deep neural networks, long short-term memory (LSTM) networks, and the attention mechanism. More concretely, the LSTM networks are adopted to capture the temporal dependencies between learner-item interactions and between learners' problem-solving actions; the deep neural networks are adopted to capture the latent representations of learners and items as well as their interactions; and the self-attention mechanism is adopted to estimate the mapping of items to skills from scratch.

In summary, our work makes the following contribution to the literature.

- We investigate the possibility of using deep learning as technical underpinnings for enhanced learner modeling, which has rarely been investigated in previous studies.
- We attempt to develop an approach that can deal with both product and process data for learner modelling in the context of CBAs.
- We attempt to develop an approach with the potential of automatically discovering item-skill associations without expert knowledge. This might benefit large-scale CBA scenarios by reducing human efforts in prespecifying the mapping of items to skills.

2 Literature review

2.1 Existing approaches for learner modeling

In educational measurement, two families of modern psychometric models —IRT and CDMs — are widely used to model the process of learners responding to assessment items measuring one or multiple underlying skills. Psychometric models estimate learners' latent skill levels and item parameters characterizing item features (e.g., difficulty and discrimination). In contrast to computational approaches, psychometric models rely on strong theoretical assumptions regarding the associations of skill mastery with item responses. For example, standard IRT models only allow one latent skill to be measured (i.e., unidimensionality), making them inadequate in addressing multiple skills. Learner modeling by CDMs requires a pre-specified human-labelled mapping of items to latent skills, failing to address CBAs with many assessment items. In addition, unlike computational approaches, constrained by their theoretical assumptions, psychometric models, without sophisticated model

revisions, have limited capacity to discover how assessment items associate with targeted latent skills from scratch. This feature of learner modeling, however, benefits large-scale CBAs since human efforts in defining item-skill associations are less needed. Moreover, psychometric models are mostly used in conventional standardized assessments in which learner data is typically structured, clean, and complete. However, learners may interact with different subsets of CBA items asynchronously. As such, learner data may be of unequal sequence lengths and with much randomness and noise. Therefore, conventional psychometric models are not scalable enough to model large-scale learner data. Finally, since learning occurs as learners continuously interact with CBA items, learner modeling without accounting for how previous learning outcomes affect current and future learning might overlook the dynamic changes in learners' cognitive states. Unfortunately, given that psychometric models typically require the assumption of local independence (i.e., conditional on latent skill levels, item responses are independent of each other), they are limited in modeling the temporal dependencies between item responses.

Bayesian approaches have also been widely used for learner modeling, since they are computationally sound, and highly flexible and expressive (Desmarais et al., 2012). Particularly, Bayesian networks, a type of probabilistic graphic model that graphically represents a joint distribution of random variables (Koller & Friedman, 2009), are of great popularity for learner modeling (de Klerk et al., 2015). To address the dynamic changes in cognitive states across multiple CBA items, variants of Bayesian networks with a temporal dimension — dynamic Bayesian networks (DBNs) and its special case BKT — were developed to estimate and update learners' skill levels as learning progresses. In empirical studies, Bayesian networks and their variants have been used for learner modeling in CBAs assessing high-level skills (e.g., creative problem solving, Shute et al., 2009; 21st-century skills, Shute & Ventura, 2013) and knowledge and skills in science and mathematics (e.g., Cui et al., 2019; Levy, 2014). Despite their popularity, learner modeling with Bayesian approaches in CBAs suffers from the curse of dimensionality — a great number of items and skills may lead to highly complex computations of conditional probabilities. In addition, similar to psychometric models, standard Bayesian approaches typically require the mapping of items to skills to be prespecified so that they cannot be directly used for automatic discovery of item-skill associations.

Another strand of educational data mining research has focused on adapting CF techniques for learner modeling. Initially developed and used for recommender systems, the CF technique has gained increasing popularity in modeling educational data in recent years (e.g., Almutairi et al., 2017; Desmarais & Naceur, 2013; Durand et al., 2015; Lan et al., 2014; Matsuda et al., 2015). For example, matrix factorization, a model-based CF approach, is of great potential for learner modeling because of its effectiveness in recovering unknown user-item interactions given sparse user data. It should be noted that most CF research in educational data mining mainly focused on employing CF to evaluate, discover, or refine the mapping of items to skills (e.g., Desmarais, 2012; Desmarais & Naceur, 2013; Durand et al., 2015; Lan et al., 2014; Matsuda et al., 2015; Sun et al., 2014). For example, the data-driven item-skill associations by CF-based approaches were close to or even outperformed the expert-specified ones (Desmarais & Naceur, 2013; Matsuda et al., 2015; Sun

et al., 2014). In summary, the literature highlights the potential of CF approaches for learner modeling and their capacity to learn item-skill associations from the scratch.

2.2 Deep learning approaches for learner modeling

Recently, deep learning-based approaches have proven exceptionally effective in predicting learners' unknown or future learning outcomes. Learner modeling with deep learning is essentially a supervised learning problem — based on various inputs regarding learners, items, and learning contexts, a deep learning model outputs the predictions of learners' unknown or future item responses (e.g., probabilities of succeeding on unknown or future items). Notably, the variety of deep learning architectures (e.g., deep neural networks and recurrent neural networks [RNNs]) allows the flexibility of deep learning approaches in addressing complex learner data. For example, previous studies have exploited the side information of learners and items (e.g., item context and learner background) to improve the accuracy of learner modeling with convolutional neural networks or RNNs (Chaplot et al., 2018; Cheng et al., 2019; Su et al., 2018). Particularly, DKT (Piech et al., 2015), an RNN-based learner modeling approach, is exceptionally effective in accounting for the temporal dependencies between item responses. The advantages of DKT and its variants in learner modeling over conventional learner models have been well documented in the literature (e.g., Wang et al., 2017; Xiong et al., 2016; Yeung, 2019; Yeung & Yeung, 2018).

More recently, researchers have incorporated deep learning architectures into the CF framework for improved learner modeling (e.g., Chen et al., 2023). Deep learning-based CF approaches can capture a high degree of complexity (e.g., nonlinearity) of learner-item interactions through deep neural networks. For example, multiple neural network layers can be used to learner item and user vectors, resulting in enhanced prediction performance through strong item and learner representations (e.g., He et al., 2017; Nguyen et al., 2018). The inclusion of deep learning architectures largely improves the prediction performance of conventional CF methods because they have a strong capacity to learn finer-grained representations and auxiliary information of users and items. However, the effectiveness of deep learning-based CF approaches for learner modeling in the CBA context remains under-investigated.

2.3 Learner modeling with process data

As mentioned, a few case studies exist showing how process data analytics can inform learning in the CBA settings. For instance, Greiff et al. (2015) analyzed the process data of one question on complex problem-solving in PISA 2012 to identify learners' problem-solving strategies. They extracted a set of frequency-related and time-related features from the process data and examined how these features predicted learners' problem-solving success. Notably, they identified a dominant strategy for solving the question. However, their analyses were conducted in an exploratory fashion with only one item, which is not scalable and extendable in

other settings. With the data of an item from the same assessment, Liu et al., (2018) proposed to use a modified multilevel mixture IRT model to analyze learners' process data, which identified different latent classes of problem-solving strategies and estimated learners' abilities at both the process and item levels. Their approach was also examined with the data of one item and showed limited generalizability. The PISA dataset was also analyzed by the event history analysis model proposed by Chen et al., (2019). Their approach was developed to model the problem-solving process with the aim of predicting both the remaining time a learner needs to complete the item and the final problem-solving outcomes (success or failure). However, their approach suffers the limitation of single-item analysis as well, which cannot be well extended to multiple-item analysis. Similarly, Shu et al., (2017) proposed a Markov-IRT model to extract features from learners' problem-solving processes as evidence for psychometric measurement. However, the Markov property assumed by their approach limits the temporal dependencies in problem-solving between two consecutive actions.

More recently, Tang et al., (2021) proposed a more generalizable approach for extracting informative features from learners' action sequences in solving a problem based on the sequence-to-sequence autoencoder. The learned latent features indicate how learners attempt a problem, which can be used for subsequent statistical or machine-learning analysis. Essentially, their approach is the representation learning of action sequences. However, it is limited in dealing with multiple items simultaneously and modeling the time information. Moreover, in terms of learner modeling or other predictive analyses, a sophisticated model is still needed to connect representation learning of action sequences with different model architectures.

In summary, the existing approaches for learner modeling with process data were mainly developed and examined in specific contexts, and they often fail to deal with multiple items. Moreover, some approaches heavily rely on statistical or psychometric assumptions and require human-specified rules, undermining their scalability and generalizability. Regarding learner modeling, few approaches can model item responses with process data at a large scale across multiple items.

Overall, our review of the existing literature on learner modeling concluded that deep learning-based approaches are of great potential for effective learner modeling in the context of CBAs, but this topic remains under investigated. In addition, existing deep learning-based approaches for learner modeling cannot adequately address process data. Consequently, this study aims to develop a deep learning-based approach to address product and process data for enhanced learner modeling. Specifically, the following three research questions are to be addressed in this study.

- Do the proposed model and its variants show satisfactory prediction accuracy in predicting learning performance in the context of CBAs?
- Does the proposed model outperform another popular deep learning-based learner model (i.e., DKT)?
- Does the proposed model show good prediction performance at different levels of data availability for training?
- Can the proposed model automatically discover interpretable item-skill associations?

3 Method

In the following, we first introduce the proposed model with technical details. Then we describe how to evaluate the effectiveness of the proposed model with a real-world dataset.

3.1 Introduction to the proposed model

The following sections present technical details for the proposed model, starting with the problem formulation, followed by the technical details of the modeling framework.

3.1.1 Problem formulation

Suppose the approach applies to data of m independent learners interacting with an n-item assessment on k latent skills. As such, the learner-item interactions can be represented as $\mathbf{R}_i = \{(\mathbf{m}_i, \mathbf{n}_1^i, R_1^i, L_1^i), (\mathbf{m}_i, \mathbf{n}_2^i, R_2^i, L_2^i), \ldots, (\mathbf{m}_i, \mathbf{n}_T^i, R_T^i, L_T^i)\}$, where $\mathbf{m}_i$ and $\mathbf{n}_t^i$ label learner identifications and item identifications at the t th timestep, respectively. Moreover, R_t^i, taking a value of either one (correct) or zero (incorrect), denotes the learning outcome at the t th timestep, and $L_t^i = \{\boldsymbol{a}_t^i, \boldsymbol{t}_t^i\}$, consisting of an action sequence $\boldsymbol{a}_t^i$ and a time sequence $\boldsymbol{t}_t^i$, denotes the problem-solving process associated with R_t^i at the t th timestep. Given a sequence of a learner's learner-item interactions $\mathbf{R}_i$ over T timesteps, the proposed model aims to learn a model $\mathcal{M}$ that predicts his or her learning outcome $\widehat{R}_{T+1}^i$ on the next item $\mathbf{n}_{T+1}^i$ at the timestep $T+1$. In addition to predictions of future learning outcomes, the proposed model discovers the mapping of items to latent skills from the associations between items during the model training process.

3.1.2 Modeling process of the approach

Figure 1 graphically presents the architecture of the proposed model, which is of two sub-architectures: one architecture for modeling item responses and problem-solving processes and the other for predicting future item responses.

Embeddings of items and learners Given the raw data $\mathbf{R}_i$, the proposed model first learns latent representations of learners and items from the identification vectors $\mathbf{m}_i$ and $\mathbf{n}_t$ through embedding layers. Specifically, the approach converts sparse vectors of learners and items to dense vectors with a pre-specified dimensionality k. As such, learner and item identifications can be represented by a k-dimensional learner representation $\mathbf{U} = \left[\mathbf{u}_1, \ldots, \mathbf{u}_m\right]$ and a k-dimensional item latent representation $\mathbf{V} = \left[\mathbf{v}_1, \ldots, \mathbf{v}_n\right]$ respectively.

Deep learning of problem-solving processes In addition to learner and item embeddings, the process data needs to be processed and learned for sequential modeling

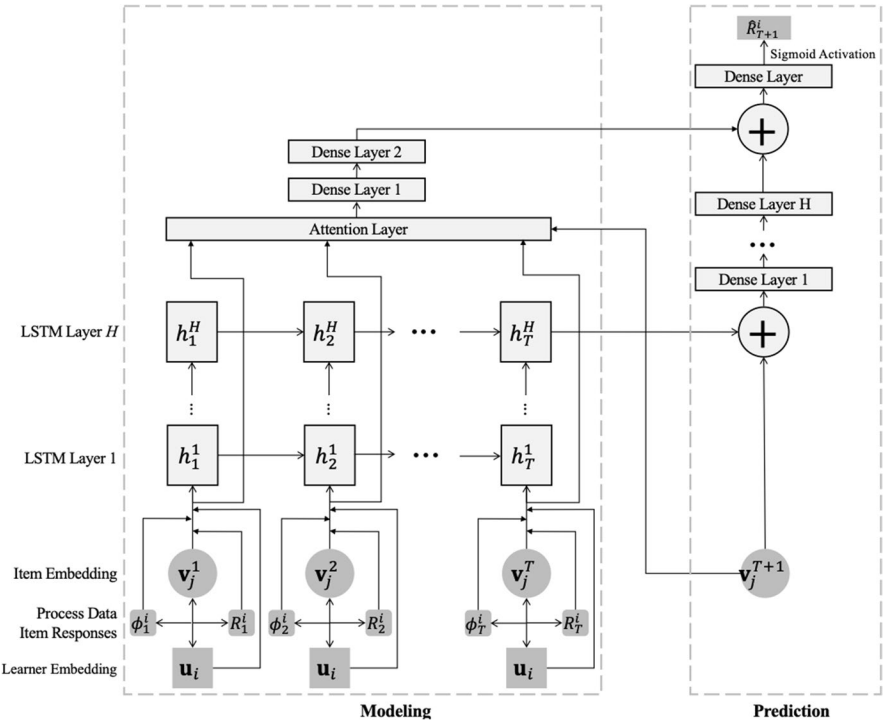

Fig. 1 Graphical representation of the proposed model

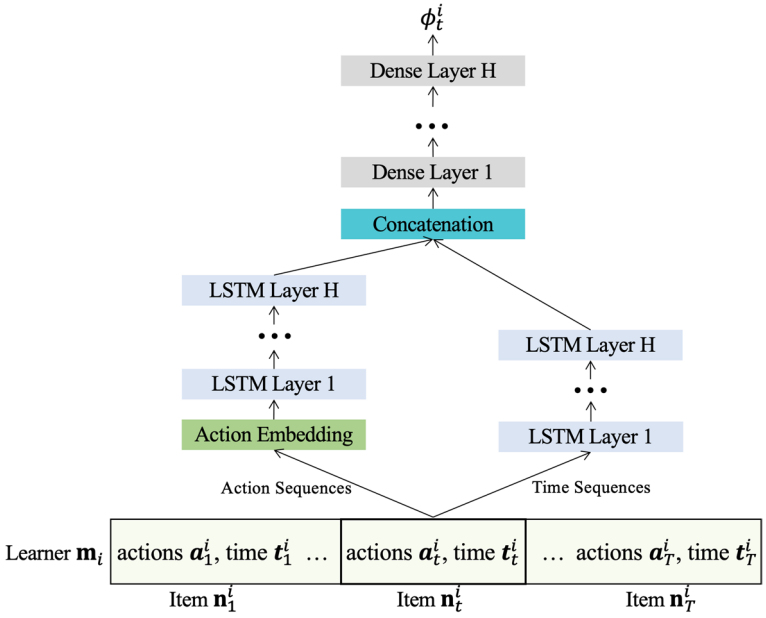

Fig. 2 Architecture for process data learning in the proposed model

(see Fig. 2). At the t th timestep, learner $\mathbf{m}_i$ responding item $\mathbf{n}_t^i$ produces a sequence of problem-solving actions $\boldsymbol{a}_t^i = \{e_1, e_2, \ldots e_Q\}$ and a sequence of action-associated time durations $\boldsymbol{t}_t^i = \{t_1, t_2, \ldots t_Q\}$, where e_q and t_q indicate the q th problem-solving step and associated time duration. Given that $\boldsymbol{a}_t^i$ is a vector with categorical values, the model converts each action e_q to a dense vector of d_0 dimensions through embedding, which is then fed into an LSTM network layer for learning the time-series dependencies between actions. It should be noted that multiple LSTM layers are allowed to better capture the complexity of temporal dependencies across multiple timesteps. The LSTM networks finally produce learned representations of actions and time durations. Subsequently, the approach concatenates learned representations of actions and time durations and feeds them into a deep neural network architecture for learning the interactions between actions and time durations, producing a final learned representation of process data at the t th timestep $\boldsymbol{\phi}_t^i$.

Concatenating leaner-item interactions Next, the proposed model first concatenates the latent representations of learners and items, and the latent representation of process data, resulting in a $(2\,k + d_a)$-dimensional vector, $\mathbf{e}_{ij}$. To concatenate $\mathbf{e}_{ij}$ with the item response R_t^i at timestep t, since R_t^i takes a value of either one or zero, $\mathbf{e}_{ij}$ is extended to a $(2\,k + d_a)$-dimensional vector $0 = (0, 0, \ldots, 0)$, resulting in a final concatenated vector $\mathbf{e}_{ij}^t$ as:

$$\mathbf{e}_{ij}^t = \begin{cases} [\mathbf{e}_{ij} \oplus 0] if R_t^i = 1 \\ [0 \oplus \mathbf{e}_{ij}] if R_t^i = 0 \end{cases} \tag{1}$$

where $\oplus$ indicates concatenation.

Deep learning for sequential learning After concatenations, the model feeds $\mathbf{e}_{ij}^t$ into one or multiple LSTM network layers to learn how item responses temporally associate with each other. Mathematically, an LSTM network layer recurrently updates the hidden state of each $\mathbf{e}_{ij}^t$ at the t th timestep h_t with its previous hidden state h_{t-1}:

$$\begin{aligned} f_t &= \sigma\left(W_f\left[h_{t-1}, \boldsymbol{e}_{ij}^t\right] + b_f\right), \\ i_t &= \sigma\left(W_i\left[h_{t-1}, \boldsymbol{e}_{ij}^t\right] + b_i\right), \\ C_t &= f_t \cdot C_{t-1} + i_t \cdot tanh\left(W_C\left[h_{t-1}, \boldsymbol{e}_{ij}^t\right] + b_C\right), \\ o_t &= \sigma\left(W_o\left[h_{t-1}, \boldsymbol{e}_{ij}^t\right] + b_o\right), \\ h_t &= o_t \cdot tanh\left(C_t\right) \end{aligned} \tag{2}$$

In the above, f_t, i_t and o_t denotes the forget, input, and output gates within an LSTM cell respectively, C_t denotes the cell state at the t $^{\text{th}}$ step, and σ and tanh indicate the *Sigmoid* and the hyperbolic tangent activation functions respectively. In addition, W_f and b_f, W_i and b_i, and W_o and b_o indicate the weights and bias of the forget gate, the input gate, and the output gate respectively. In summary, the three gates of an LSTM cell control what information to be inputted, remembered, forgotten, and outputted through the cell. This feature contributes to the effectiveness of

the LSTM network in learning temporal dependencies. The output sequence of the last LSTM layer $\mathbf{S} = \{s_1^i, s_2^i, \ldots, s_T^i\}$ incorporates the sequential information on how a learner interacts with items over the past T timesteps. Next, the model concatenates s_T^i with the embedding vector of the next item at timestep $T+1$, $\mathbf{v}_j^{T+1}$, and feeds the concatenation into multiple neural network layers, which can be formally stated as:

$$D_{T+1}^i = f_H(\boldsymbol{W}_H^T f_{H-1}(\ldots f_2(\boldsymbol{W}_2^T f_1(\boldsymbol{W}_1^T \begin{bmatrix} \boldsymbol{s}_T^i \\ \boldsymbol{v}_j^{T+1} \end{bmatrix})) \ldots)) \tag{3}$$

In the above, $\mathbf{W}_1$ to $\mathbf{W}_H$, and f_1 to f_H denote the weights and activation functions for the H neural network layers, respectively. The final output of the multiple neural network layers, D_{T+1}^i, combines the information on the current item for prediction and the information regarding all history item responding processes.

Self-attention mechanism To make the model more predictive of item responses, in addition to LSTM networks, the proposed model applies a self-attention layer (Vaswani et al., 2017) to model the relevance of an item for prediction with a learner's history item responding processes. The attention mechanism deals with three types of vector inputs: query, key, and value. Specifically, in the proposed model, the query refers to the item embeddings of an item for prediction, and both keys and values refer to a learner's history item responding processes $\mathbf{e}_{ij}^t$. In the attention mechanism, a compatibility function is used to model the relevance of a query with different keys, represented by attention weights. The output of the attention mechanism is calculated as a weighted sum of value vectors using the attention weights. Specific to the proposed model, the relevance of an item for prediction with previous items can be represented by its attention weights connecting to other items. We used the scaled dot-product attention (Vaswani et al., 2017) in the current study, which is formally calculated as:

$$\text{Attention}(\mathbf{V}, \mathbf{S}, \mathbf{S}) = \text{softmax}\left(\boldsymbol{V}\boldsymbol{S}^T/\sqrt{k}\right)\mathbf{S}, \tag{4}$$

where $\mathbf{S}$, $\mathbf{S}$, and $\mathbf{V}$ denote the query, key, and value matrices of dimension k respectively, and $\text{softmax}(\mathbf{V}\mathbf{S}^T/\sqrt{k})$ generates the attention weights. It is noteworthy that the prediction of an item response at timestep $T+1$ should be solely based on the item responding processes over the past T timesteps, and therefore when computing attention weights, the model omits keys at timesteps later than t for any query at timestep t. In addition, to impose non-linearity on the weighted attention output, according to Vaswani et al., (2017), for each timestep, the output of the attention layer is fed into a feedforward neural network layer and one layer with the ReLU activation. Moreover, through a residual connection (He et al., 2016), the model adds up the input and the output of each layer as the final output so that the importance of lower-layer features can be better captured. Layer normalization (Ba et al., 2016) applies to each layer of the attention mechanism.

Prediction The proposed model makes predictions through feeding the concatenated output of the deep LSTM network architecture D and the attention mechanism F into one neural network layer with Sigmoid activation (see the right-hand part of Fig. 1):

$$\widehat{R}^i_{T+1} = \text{Sigmoid}\left(\boldsymbol{W}^{\boldsymbol{T}}\begin{bmatrix}\boldsymbol{D}\\ \boldsymbol{F}\end{bmatrix}\right) \tag{5}$$

Model learning During the training process, the proposed model updates the following model parameters: the embedding weights for items, learners, and problem-solving actions, the LSTM network weights, and the neural network weights. The binary cross-entropy loss is used as the objective function for model learning:

$$J = -\sum\nolimits_{t=1}^{T} R^i_t \log \widehat{R}^i_t + \left(1 - R^i_t\right)\log\left(1 - \widehat{R}^i_t\right), \tag{6}$$

where $\widehat{R}^i_t$ indicates the model-predicted likelihood of correctly solving items at the t th timestep. The Adaptive Moment Estimation (Adam; Kingma & Ba, 2014) is selected as the optimizer in training.

3.2 Dataset description

To evaluate the effectiveness of the proposed model, we used a real-world dataset accessed from the PSLC DataShop[1] (Koedinger et al., 2010), named "Lab study 2012 (cleanedLogs)." There are 74 learners, 14,959 problem-solving steps, and 37,889 transactions involved in the dataset. Moreover, among the six latent skill models (each corresponds to a different number of latent skills), we selected the one labelled "KC (DefaultFewer_corrected)" for training the model. The data was generated through learners interacting with the web-based tutoring system when solving fraction problems. Notably, learners might take different sets of fraction problems, implying different item sequences for each learner. In this study, since learners might take several problem-solving steps to solve a fraction problem, in this study, one problem-solving step was considered an independent item which involved one or multiple transactions (i.e., specific timestamped problem-solving actions). To preprocess the dataset, we first deleted all system-produced and/or non-timestamped transactions and then treated all problem-solving steps related to hints as intermediate actions for solving a problem. In addition, to make problem-solving actions differentiable, we concatenated the labels of actions and corresponding learner selections, given that actions of the same categories share the same labels. For learners' action and time sequences for solving each item, we fixed the maximum action and time sequence length at six because over 90% of items were attempted with six or fewer actions by learners. Finally, since most sequences of learner-item interactions are of more than 200 timesteps, we split the item sequences of the 74 learners into multiple 20-timestep subsequences to increase the size of item sequences for

[1] https://pslcdatashop.web.cmu.edu/

training. This resulted in a final dataset involving 866 item sequences, 32 unique items, and 15 unique skills.

3.3 Training settings

In this study, the embedding weights of items, learners, and actions were regularized with a finalized regularization weight of 0.001, which was selected from four candidate weights: 0, 0.001, 0.01, and 0.1. In addition, to reduce overfitting, a dropout layer with a dropout rate of 0.5 was applied prior to each neural network layer, which was selected from three candidate rates: 0, 0.2, and 0.5. Regarding the sub-architectures of the proposed model, both the deep LSTM network architecture and the architecture for prediction involve one layer with output dimensions of five and two, respectively. Moreover, we selected a latent dimension of 120 for the embedding layers of items, learners, and actions. The learning rate for Adam was finalized at 0.0001, which was selected from the following four values: 0.0001, 0.001, 0.01, and 0.1. The model was trained for 150 epochs with a finalized batch size of 256, which was selected from the following candidate values: 5, 32, 64, 128, and 256.

3.4 Evaluation settings

In this study, DKT was selected as the baseline for evaluating the effectiveness of the proposed model. DKT is a deep learning-based learner modeling approach that predicts the probabilities of the next learning performance based on modeling history learning performance with an RNN architecture (Piech et al., 2015). DKT was used as a baseline for model evaluation in many learner modeling studies, where it has been found to outperform conventional models such as BKT (e.g., Xiong et al., 2016). In this study, DKT was modelled with a 100-node LSTM layer, and the model was trained at the skill level (i.e., skill IDs were used as inputs) with a learning rate of 0.001. In addition to DKT, the proposed model was compared against its two sub-architectures, the attention and the LSTM variants. Instead of concatenating the attention and the LSTM outputs for final predictions as in the full model, the attention and the LSTM variants make predictions solely based on the outputs of the attention mechanism and the LSTM architecture, respectively. Moreover, to examine if process data learning is effective for improving prediction performance, we compared the proposed model with a variant without the module for process data learning.

In this study, to evaluate the performance of the proposed model, we selected the first 30%, 50%, and 70% of item responses of each learner item response sequence for training. The model was evaluated with both the regression and classification metrics. The classification metrics included Accuracy (ACC) and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC; Ling et al., 2003). ACC scores are computed as the percentage of correctly predicted item responses with a cut-off value of 0.5. Unlike ACC, AUC indicates the area under the plot of sensitivity rates against the false-positive rates and therefore, its calculation does not rely on any specific cut-off values. This feature of AUC makes it insensitive to class

Table 1 Comparison of testing performance between the proposed model and other models

Model	ACC	AUC	MAE	RMSE
Training ratio: 0.7				
DKT	0.7037	0.7157	0.3786	0.4339
Model without process data	0.7143	0.7347	0.3583	0.4298
Full model	**0.7225**	**0.7400**	**0.3580**	**0.4254**
Attention variant	0.7219	0.7395	0.3583	0.4258
LSTM variant	0.6909	0.6928	0.4065	0.4419
Training ratio: 0.5				
DKT	0.6890	0.6974	0.3739	0.4422
Model without process Data	0.7076	0.7266	0.3587	0.4342
Full model	**0.7160**	**0.7323**	**0.3578**	**0.4300**
Attention variant	0.7160	0.7309	0.3589	0.4305
LSTM variant	0.6904	0.6946	0.4028	0.4408
Training ratio: 0.3				
DKT	0.6672	0.6439	0.3764	0.4748
Model without process data	0.7065	0.7182	0.3595	0.4382
Full model	**0.7126**	0.7259	**0.3616**	**0.4330**
Attention variant	0.7118	**0.7261**	0.3617	0.4335
LSTM variant	0.6847	0.6882	0.4093	0.4453

ACC = Accuracy, *AUC* = Area under the ROC Curve, *MAE* = Mean Absolute Error, *RMSE* = Root Mean Square Error. Values in bold represent the metric of the optimum model of the ones compared

imbalance (e.g., the majority of item responses are correct, and few are incorrect). The regression evaluation metrics (Willmott & Matsuura, 2005) included the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE).

4 Results

4.1 Prediction performance

The evaluation performance of each model on the test datasets across different training/test partition ratios is presented in Table 1. Generally, disregarding the training/test partition ratios, the proposed model demonstrates higher ACC and AUC rates and lower MAE and RMSE rates than DKT and the variant without process data learning. Moreover, using more history items for training slightly improves the prediction accuracy of the proposed model, shown by slightly higher ACC and AUC rates and slightly lower MAE and RMSE rates.

Regarding the comparison between the proposed model and its two sub-architecture variants, it is evident that the proposed model has a similar or higher prediction performance than its two sub-architecture variants. However, the attention variant slightly outperforms the LSTM variant.

4.2 Mapping of items to skills

In this study, the proposed model adopted the approach by Pandey & Karypis, (2019) to discover the mapping of items to skills. In the model, through the attention mechanism, the attention weights of each item for prediction (i.e., the query) can be used to indicate the connection strength of an item with its previous items. The attention weights for each possible item pair (i.e., [query item, key item]) can be summed over all learners to derive their relevance weights, which are then normalized for each query item so that weights of each item sum to one. According to the relevance weights of each item, the items measuring the same skill can be indicated by the clusters of items with the strongest connections to each other.

According to the heatmap of item relevance weights (see Fig. 3), even though the clustering of items is not fully clear-cut, it can be found that a major item cluster includes items 10 to 21, shown by their stronger connections to each other than others, indicating that items 10 to 21 might measure the same skill. To validate this, we compared the discovered item-skill associations with the original skill model. According to the skill model, items 10 to 21 were designed to evaluate

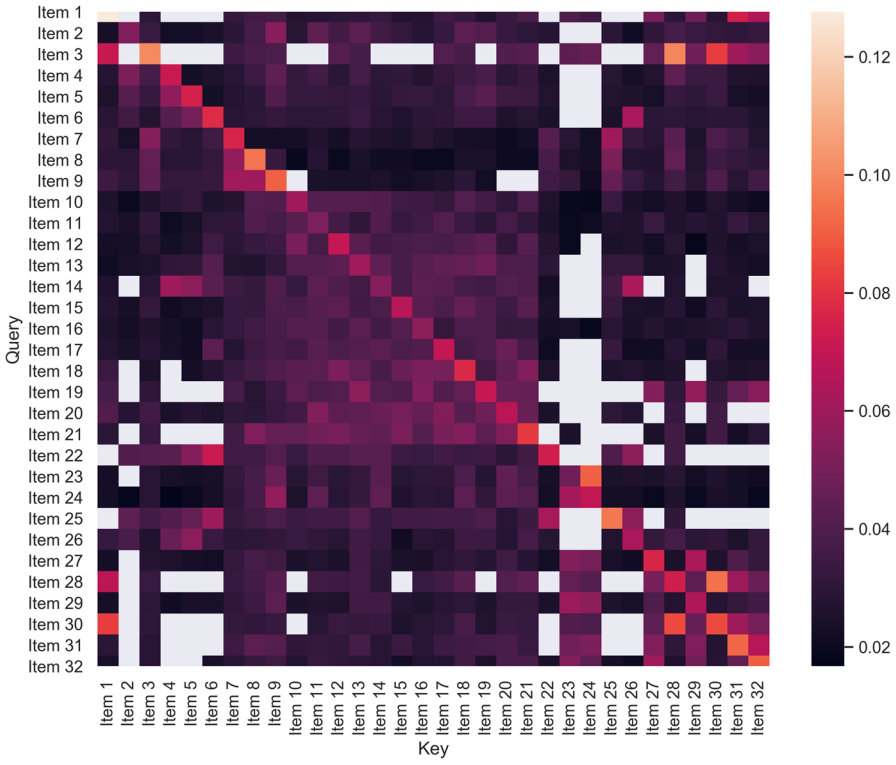

Fig. 3 Heatmap indicating item relevance by the proposed model. Note. The item and skill names are presented in Table 2

Table 2 Item and skill labels

ID	Item label	Skill label
1	combo1 UpdateComboBox	equivFractEquivalent
2	combo1_3 UpdateComboBox	compFract
3	combo2 UpdateComboBox	relationEquivMultiplySameNumber
4	combo2_1 UpdateComboBox	compSectSize
5	combo2_2 UpdateComboBox	compNumSect
6	combo2_3 UpdateComboBox	compFract
7	combo3 UpdateComboBox	relationEquivConserveAmount
8	combo4 UpdateComboBox	relationEquivSameAmount
9	combo5 UpdateComboBox	relationEquivDiffNumbers
10	dragTarget1 WasJustHitByA Circle	equivDragFract
11	dragTarget1 WasJustHitByA NL	equivDragFract
12	dragTarget1 WasJustHitByA Rect	equivDragFract
13	dragTarget2 WasJustHitByA Circle	equivDragFract
14	dragTarget2 WasJustHitByA NL	equivDragFract
15	dragTarget2 WasJustHitByA Rect	equivDragFract
16	dragTarget3 WasJustHitByA Circle	equivDragFract
17	dragTarget3 WasJustHitByA NL	equivDragFract
18	dragTarget3 WasJustHitByA Rect	equivDragFract
19	dragTarget4 WasJustHitByA Circle	equivDragFract
20	dragTarget4 WasJustHitByA NL	equivDragFract
21	dragTarget4 WasJustHitByA Rect	equivDragFract
22	fract1_denom1 UpdateTextArea	relationCompTotalSectNumber
23	fract1_denomMultiply1 UpdateTextArea	equivMultiplyDenom
24	fract1_numMultiply1 UpdateTextArea	equivMultiplyNum
25	fract2_denom1 UpdateTextArea	relationCompTotalSectNumber
26	fract2_num1 UpdateTextArea	numSectZeroDot
27	fract3_denom UpdateTextArea	equivNameDenomFract
28	fract3_denomMultiply1 UpdateTextArea	equivMultiplyDenom
29	fract3_num UpdateTextArea	equivNameNumFract
30	fract3_numMultiply1 UpdateTextArea	equivMultiplyNum
31	fract4_denom UpdateTextArea	equivNameDenomFract
32	fract4_num UpdateTextArea	equivNameNumFract

the same skill of "equivDragFract" (see Table 2), which suggests the potential of the proposed model to identify item-skill associations from the scratch automatically. However, unfortunately, the heatmap shows that the proposed model might be less capable of discovering item-skill associations in case only one or two items are developed for measuring a skill. Unsurprisingly, given few items developed for a skill, the skill might not be fully represented and measured by the items, and learners might not have adequate opportunities to exercise the skill. As such, the relevance weights might be calculated with much randomness, resulting

in a less clear-cut clustering of items. Moreover, it should be noted that despite multiple skill models for the dataset, the ground truths regarding the connections between items and skills are never known. Therefore, we cannot fully validate the estimated item-skill associations by the proposed model. To sum up, the proposed model successfully identified the mapping of the major skill to most assessment items, supporting its potential to discover item-skill associations from scratch.

5 Discussion and future work

This work proposed a novel deep learning-based model to sequentially model learning outcomes using product and process data. According to the evaluation results, we conclude that the proposed model can predict learning outcomes with high accuracy and automatically identify the mapping of items to skills without prior expert knowledge. Compared with the model without process data learning, the proposed model accounts for additional information from learner problem-solving processes to improve prediction accuracy.

Notably, our approach aligns with the multiple purposes of learning outcome modeling in the context proposed by Pelánek (2017). Specifically, learners' future interactions with a system are affected by the outputs (e.g., predictions of future item responses) of a model analyzing learner data extracted from the system through three loops. During the process of learners interacting with an assessment item, the process data modelling module of the proposed model has the potential to process learners' problem-solving steps and produce estimated probabilities of item successes, which is characterized as affecting learners' short-term behaviors within the "inner loop" of learning outcome modelling. Regarding the predictions of future item responses, they can be used to inform the instructional policies for improved learning effects. For example, suppose a system predicts that a learner will correctly solve the next item with a 95% chance. In that case, the system will stop presenting other similar items for measuring the same skill since the learner is very likely to have mastered the skill. Moreover, if a system predicts that an item is too hard or too easy for a learner, then the system will skip or delay presenting the item to the learner to maximize his or her learning effects. These exemplify how the proposed model can affect learners' future interactions through the "outer loop" of learning outcome modeling by Pelánek (2017). Regarding the third loop with human involvement, the item-skill associations discovered by the proposed model can be used to provide actional insights. For example, if a system discovers that an item is of low quality or not related to most other items, a human expert might consider dropping this item from the item bank to improve the validity of the assessment system. In summary, the proposed model, the proposed model, bears great potential in promoting personalized learning through its three major features of process data learning, learner modeling, and domain modeling, which correspond to the three loops of learning outcome modeling proposed by Pelánek (2017).

Pedagogically, our study posits that the proposed model substantively contributes to personalized learning applications, particularly in the context of CBAs, in

several key dimensions. First, our study offers a novel, scalable learning outcome modelling approach, affording education practitioners a valuable tool to adeptly leverage both learner product and process data. As suggested by our findings, incorporating student process data significantly enhances the predictive capacity of personalized learning systems. This implies that education practitioners should not only prioritize learners' explicit performance but also examine their problem-solving processes for a comprehensive understanding of how learners achieve learning objectives. Second, the high predictive capability of the proposed model facilitates a more efficient personalized learning system, proficiently tailoring recommendations for learning materials and assessments to individual students. Finally, the functionality of the proposed model to discover the mapping of items to skills benefits the development of large-scale assessments. Since items can be automatically mapped onto their targeted skills, the development of a large-scale CBA can be more expeditious and cost-effective. In terms of practical implications, educators are encouraged to implement the proposed model to create more reliable and predictive personalized learning experiences for learners, with insights into both the "where" and "why" of learners' performance. In addition, the proposed model streamlines the development of large-scale CBAs. More importantly, the proposed model holds promise for adaptation to other digital learning environments where learners' product and process data are available (e.g., massive open online courses) to inform the optimization of learning outcomes and the learning context.

Inevitably, several limitations exist for the current work. First, since the model discovers item-skill associations through identifying major item clusters based on the estimated relevance weights between items instead of parameterizing latent skills, the mapping of items to skills might be discovered with randomness, especially when skills are measured by a limited number of assessment items. Second, despite its satisfactory performance in addressing dichotomous item responses, the model needs to be adapted to deal with polytomous item responses in future work since non-binary scoring is prevailing in most educational settings. Third, the better demonstrate and understand how learners acquire new knowledge, the interpretability of the proposed model can be enhanced considering the black box nature of deep learning architectures.

Funding This paper was supported by the University of Macau Start-up Research Grant (Grant No. SRG2021-00023-FED).

Data availability The datasets generated during and/or analysed during the current study are available in the PISA official database, https://www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm

Declarations

Ethics approval Not applicable.

Conflict of interest The authors declare that they have no competing interests.

Informed consent Not applicable.

References

Almutairi, F. M., Sidiropoulos, N. D., & Karypis, G. (2017). Context-aware recommendation-based learning analytics using tensor and coupled matrix factorization. *IEEE Journal of Selected Topics in Signal Processing, 11*(5), 729–741. https://doi.org/10.1109/JSTSP.2017.2705581

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. arXiv preprint. https://doi.org/10.48550/arXiv.1607.06450

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 95–102). International Educational Data Mining Society.

Chaplot, D. S., MacLellan, C., Salakhutdinov, R., & Koedinger, K. (2018). Learning cognitive models using neural networks. In *International Conference on Artificial Intelligence in Education* (pp. 43–56). Springer. https://doi.org/10.1007/978-3-319-93843-1_4

Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.00486

Chen, F., Lu, C., Cui, Y., & Gao, Y. (2023). Learning outcome modeling in computer-based assessments for learning: A sequential deep collaborative filtering approach. *IEEE Transactions on Learning Technologies, 16*(2), 243–255. https://doi.org/10.1109/TLT.2022.3224075

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., ... & Hu, G. (2019). Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 2397–2400). Association for Computing Machinery. https://doi.org/10.1145/3357384.3358070

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

Cui, Y., Chu, M. W., & Chen, F. (2019). Analyzing student process data in game-based assessments with Bayesian knowledge tracing and dynamic Bayesian networks. *Journal of Educational Data Mining, 11*(1), 80–100. https://doi.org/10.5281/zenodo.3554751

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter, 13*(2), 30–36. https://doi.org/10.1145/2207243.2207248

Desmarais, M. C., Baker, R. S., & d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1–2), 9–38. https://doi.org/10.1007/s11257-011-9106-8

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In *International Conference on Artificial Intelligence in Education* (pp. 441–450). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39112-5_45

Dishon, G. (2017). New data, old tensions: Big data, personalized learning, and the challenges of progressive education. *Theory and Research in Education, 15*(3), 272–289. https://doi.org/10.1177/1477878517735233

Durand, G., Belacel, N., & Goutte, C. (2015). Evaluation of expert-based Q-matrices predictive quality in matrix factorization models. In *Design for teaching and learning in a networked world* (pp. 56–69). Springer, Cham. https://doi.org/10.1007/978-3-319-24258-3_5

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web* (pp. 173–182). International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3038912.3052569

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint. https://doi.org/10.48550/arXiv.1412.6980

Koedinger, K. R., Baker, R. S. J., & d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC dataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. D. Baker (Eds.), *Handbook of educational data mining* (pp. 43–55). CRC Press.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. The MIT Press.

Kong, S. C., & Song, Y. (2015). An experience of personalized learning hub initiative embedding BYOD for reflective engagement in higher education. *Computers & Education, 88*, 227–240. https://doi.org/10.1016/j.compedu.2015.06.003

Lan, A. S., Waters, A. E., Studer, C., & Baraniuk, R. G. (2014). Sparse factor analysis for learning and content analytics. *The Journal of Machine Learning Research, 15*(1), 1959–2008.

Levy, R. (2014). Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments. CRESST Report 837. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A better measure than accuracy in comparing learning algorithms. In *Conference of the canadian society for computational studies of intelligence* (pp. 329–341). Springer. https://doi.org/10.1007/3-540-44886-1_25

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2018.01372

Lord, F. M. (1952). A *theory of test scores (Psychometric Monograph, No. 7)*. Psychometric Corporation.

Matsuda, N., Furukawa, T., Bier, N., & Faloutsos, C. (2015). Machine beats experts: Automatic discovery of skill models for data-driven online course refinement. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 101–108). International Educational Data Mining Society.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48. https://doi.org/10.5281/zenodo.3554641

Nguyen, D. M., Tsiligianni, E., & Deligiannis, N. (2018). Extendable neural matrix completion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6328–6332). IEEE. https://doi.org/10.1109/ICASSP.2018.8462164

Pandey, S., & Karypis, G. (2019). A self-attentive model for knowledge tracing. In *2th International Conference on Educational Data Mining, EDM 2019* (pp. 384–389). International Educational Data Mining Society.

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

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015, December). Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Vol 1, pp. 505–513).

Psotka, J., Massey, L. D., & Mutter, S. A. (1988). *Intelligent tutoring systems: Lessons learned*. Psychology Press.

Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1), 1–10. https://doi.org/10.5281/zenodo.3554639

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285–295). Association for Computing Machinery. https://doi.org/10.1145/371920.372071

Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling, 59*(1), 109–131.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning, 33*(1), 1–19. https://doi.org/10.1111/jcal.12172

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 2009*, 421425. https://doi.org/10.1155/2009/421425

Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., et al. (2018). Exercise-enhanced sequential modeling for student performance prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 2435–2443).

Sun, Y., Ye, S., Inoue, S., & Sun, Y. (2014). Alternating recursive method for Q-matrix learning. In *Proceedings of the 7th international conference on educational data mining* (pp. 14–20). International Educational Data Mining Society.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1–33. https://doi.org/10.1111/bmsp.12203

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Lawrence Erlbaum Associates Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 5998–6008).

Wang, H., Wang, N., & Yeung, D. Y. (2015). Collaborative deep learning for recommender systems. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1235–1244). Association for Computing Machinery. https://doi.org/10.1145/2783258.2783273

Wang, L., Sy, A., Liu, L., & Piech, C. (2017). Deep knowledge tracing on programming exercises. In *Proceedings of the fourth annual ACM conference on learning at scale* (pp. 201–204). https://doi.org/10.1145/3051457.3053985

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79–82. https://doi.org/10.3354/cr030079

Xiong, X., Zhao, S., Van Inwegen, E. G., & Beck, J. E. (2016). Going deeper with deep knowledge tracing. In *Proceedings of the 9th international conference on educational data mining* (pp. 545–550). International Educational Data Mining Society.

Yeung, C. K., & Yeung, D. Y. (2018). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the fifth annual ACM conference on learning at scale* (pp. 1–10). https://doi.org/10.1145/3231644.3231647

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

Zhang, F., Yuan, N. J., Lian, D., Xie, X., & Ma, W. Y. (2016). Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 353–362). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939673.