Federated Deep Knowledge Tracing

Jinze Wu[1], Zhenya Huang[1], Qi Liu[1,*], Defu Lian[1], Hao Wang[1], Enhong Chen[1]
Haiping Ma[2], Shijin Wang[3]
[1]Anhui Province Key Laboratory of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China
[2]Anhui University
[3]iFLYTEK Research & State Key Laboratory of Cognitive Intelligence, iFLYTEK Co., Ltd,
{hxwjz,wanghao3}@mail.ustc.edu.cn,{huangzhy,qiliuql,liandefu,cheneh}@ustc.edu.cn,
hpma2020@163.com,sjwang3@iflytek.com

ABSTRACT

Knowledge tracing is a fundamental task in intelligent education for tracking the knowledge states of students on necessary concepts. In recent years, Deep Knowledge Tracing (DKT) utilizes recurrent neural networks to model student learning sequences. This approach has achieved significant success and has been widely used in many educational applications. However, in practical scenarios, it tends to suffer from the following critical problems due to data isolation: 1) *Data scarcity*. Educational data, which is usually distributed across different silos (e.g., schools), is difficult to gather. 2) *Different data quality*. Students in different silos have different learning schedules, which results in unbalanced learning records, meaning that it is necessary to evaluate the learning data quality independently for different silos. 3) *Data incomparability*. It is difficult to compare the knowledge states of students with different learning processes from different silos. Inspired by federated learning, in this paper, we propose a novel Federated Deep Knowledge Tracing (FDKT) framework to collectively train high-quality DKT models for multiple silos. In this framework, each client takes charge of training a distributed DKT model and evaluating data quality by leveraging its own local data, while a center server is responsible for aggregating models and updating the parameters for all the clients. In particular, in the client part, we evaluate data quality incorporating different education measurement theories, and we construct two quality-oriented implementations based on FDKT, i.e., FDKTCTT and FDKTIRT-where the means of data quality evaluation follow Classical Test Theory and Item Response Theory, respectively. Moreover, in the server part, we adopt hierarchical model interpolation to uptake local effects for model personalization. Extensive experiments on real-world datasets demonstrate the effectiveness and superiority of the FDKT framework.

CCS CONCEPTS

• **Applied computing** → **Computer-assisted instruction**.

*Corresponding Author.

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441747

KEYWORDS

Federated learning; Knowledge tracing; Data isolation; Data quality evaluation; Intelligent education

ACM Reference Format:
Jinze Wu, Zhenya Huang, Qi Liu, Defu Lian, Hao Wang, Enhong Chen, Haiping Ma, Shijin Wang. 2021. Federated Deep Knowledge Tracing. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8-12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441747

1 INTRODUCTION

Knowledge Tracing (KT) is a fundamental task in intelligent education. It aims to trace knowledge states of students based on their historical learning trajectories. The success of knowledge tracing can benefit both personalized and adaptive learning so that has attracted significant attention over the past decades [1, 6, 21].

In the literature, many efforts have been made towards knowledge tracing. Among them, Bayesian Knowledge Tracing (BKT) is one of the representative early works [6]. Recently, considering students' learning on interrelated multiple concepts, Deep Knowledge Tracing (DKT)-based models have been proposed [35]. As the recent DKT-based models have stronger representational ability, they have been widely used in various educational applications, including in-class assessment and online diagnosis [31, 41, 46].

In order to learn high-quality DKT models, it inevitably requires a substantial amount of comprehensive data for guaranteeing the stability of neural networks during training [32]. However, practical educational scenarios usually suffer from critical data isolation problem [18, 19], which means that the learning data of students is usually collected and stored separately in the case of isolated silos (e.g., different schools). As a result, conventional DKT-based models could become inapplicable due to the following unique characteristics in intelligent education: 1) *Data scarcity*. Typically, learning data tends to be distributed across different schools and highly proprietary so that it is difficult to gather the data for training [17, 25]. To be more specific, as shown in Figure 1, each school only stores its own local data, because students reject to make the data public. Accordingly, it is necessary to find an appropriate solution for training DKT independently while alleviating data scarcity. 2) *Different data quality*. As [42] suggested, the success of knowledge tracing relies heavily on the quality of the learning data. However, different schools usually have different learning schedules, thus, students isolated may practice inconsistent and unbalanced

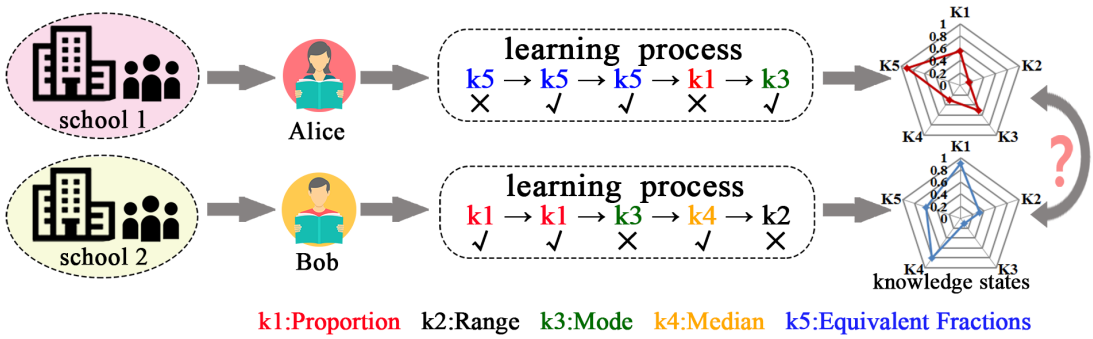


Figure 1: Example: Left part shows that two isolated schools hold their private learning data. Middle part shows the learning processes of two students from two schools, where the processes are inconsistent on concepts. The radars on the right illustrate their knowledge states on five concepts after completing the exercises.

learning processes [10]. For example, as shown in Figure 1, we can see that Alice (in school 1) mainly learns the concept "Equivalent Fractions" while Bob (in school 2) even ignores this concept. In this case, the learning data reveals inconsistent properties and settings (e.g., difficulty) among schools which results in bias of data quality [13, 36, 39]. It may affect the performances of knowledge tracing models [4]. Therefore, an effective way to evaluate data quality is highly required. Meanwhile, such quality-oriented issue also leads to a non-independent identically distributed (Non-IID) scenario in practice. That is, the distributions of data in local silos depend on data owners and are significantly different from the global one [23, 37, 49]. In this case, we should enhance the model personalization. In other words, the local model espects to aptly fit the local data for corresponding client. 3) *Data incomparability*. It is difficult to compare the knowledge states of students from different silos [5]. In other words, we are curious about which of Alice or Bob in Figure 1 learns better on the same concept eventually. Hence, the solution should also consider this demand in a flexible way.

Inspired by federated learning, in this paper, we address above problems in a principled way. We propose a novel client-server architecture framework named Federated Deep Knowledge Tracing (**FDKT**). Specifically, the client part is designed for training DKT models independently and evaluating the local data quality. Correspondingly, the server part takes charge of aggregating and updating all local models for clients. In particular, to evaluate the data quality, we propose two implementations of FDKT that incorporate quality-oriented aggregation strategies following two educational measurement theories. The first one, named FDKTCTT, incorporates Classical Test Theory (CTT) [9], which considers the statistical confidence to evaluate learning data quality. The second one, named FDKTIRT, follows Item Response Theory (IRT) [28], which investigates the information confidence. Therefore, the server aggregates models based on data quality so that we can pay more attention to models with high-quality data rather than large-scale data, and reasonably aggregate the more positive training effects. Moreover, to enhance model personalization, we modify the update strategy via hierarchical model interpolation with measuring the disparity between models. Therefore, the server updates new local models adopting local effects so that we can make local models more suitable for local data to avoid limitations from Non-IID scenario [15]. Furthermore, through our framework, on one hand, we can expand available data to mitigate data scarcity; on the other hand, students naturally become comparable by synthesizing effects of decentralized models. Extensive experiments on real-world datasets clearly demonstrate that our framework outperforms the baselines in terms of knowledge tracing effectiveness, communication cost and comparability. To the best of our knowledge, FDKT is the first framework that is specifically designed for federated deep knowledge tracing while considering both quality-oriented aggregation and personalized update strategy.

2 RELATED WORK

In this section, we briefly review some related works.

2.1 Knowledge Tracing

Knowledge tracing (KT) is a fundamental task in intelligent education dating back to the 1990s, which aims to trace the knowledge states of students based on their historical learning performances [6]. Regarding tasks of this kind, Bayesian Knowledge Tracing (BKT)-based models are one kind of the representative models [6, 34, 47]. This approach utilizes the Hidden Markov Model (HMM) to separately represent and update student knowledge states as a set of binary variables. Recently, researchers proposed Deep Knowledge Tracing (DKT) by leveraging recurrent neural networks to update students' knowledge states [35]. Subsequently, many extensions have been proposed with considering extra information, such as the knowledge-exercise relationships [14, 48], the contents of exercises [26] and graph structures [33, 40]. Experimental results show that DKTs have stronger representational abilities. However, practical educational scenarios usually suffer from data isolation problem, which restricts the application of the conventional DKT-based models since training a high-quality DKT is usually data-hungry [32].

2.2 Federated Learning

Federated learning (FL) is one of the most promising techniques in recent years, and has achieved great success in various domains including personal devices [11, 44] and banking [45]. The main idea of FL is to build and aggregate machine learning models based on data that are localized on multiple mobile devices [30, 43]. Specifically, in terms of model aggregation, researchers have proposed various strategies, such as FedSGD, FedAVG [30], FedATT [16] and LoAdaBoost [12]. However, existing FL studies tend to focus on basic statistics of data (e.g., data scale) for aggregating and Non-IID

data will impair FL performances simultaneously [42]. In this paper, we pay attention to the unique characteristics of data quality in intelligent education. For educational applications, we propose quality-oriented aggregation strategies and regard data quality as the importances of models. Moreover, we modify a hierarchical model interpolation-based update strategy for personalized federated learning to fit clients' own partial data.

2.3 Educational Measurement Theory

Educational Measurement theory provides the foundation via item quality analysis for many educational scenarios, including examination arrangement [38] and adaptive testing [20]. Generally, there are two widely-used theories, i.e., Classical Test theory (CTT) [9] and Item Response Theory (IRT) [28]. In more detail, CTT is a statistical theory that evaluates the learning data quality from item perspectives, including difficulty, discrimination, and reliability [8]. Correspondingly, IRT directly assesses the item information as learning data quality by designing an information function called Item Characteristic Curve (ICC), which considers both the student and the item characteristics [3]. In this paper, we make full use of the above measurement theories in order to improve the aggregation performances in our proposed framework from a data quality perspective, where the solutions can be naturally applied in intelligent education.

3 PRELIMINARIES

In this section, we introduce the concepts for item quality analysis and provide the formal definition of the Federated Deep Knowledge Tracing problem.

3.1 Educational Measurement Theory

In intelligent education, item analysis is an important field which aims to evaluate item quality [9]. The educational measurement theories, i.e., CTT and IRT, have been applied in it. In this paper, we integrate some important item analysis concepts and technologies into the learning data quality evaluation as follows.

3.1.1 Classical Test Theory (CTT). CTT is one of the educational measurement theories which focuses on separating errors between response results and real values. It has been widely applied to item analysis. CTT regards the statistical confidence as the data quality, and researchers have established several indicators to measure quality from different aspects, including difficulty, discrimination, and reliability [8, 9]. Specifically, difficulty reflects how difficult an item is for students; discrimination is the ability of an item to distinguish the mastery of knowledge concepts of students; reliability reflects the consistency of all the items. However, CTT is also affected by some limitations. For example, it simply relies on a weak linear hypothesis, and indicators in CTT are overly dependent on the samples of data which exhibits some biases for measurements [8]. Therefore, a new theory that can fix these disadvantages, Item Response Theory, has attracted more attentions [2].

3.1.2 Item Response Theory (IRT). IRT aims to the latent characteristics of students in the learning processes to overcome the above-mentioned critical limitations of CTT. It focuses on information confidence of learning data and researchers design Item Characteristic Curve [3] to measure the item information. It is a logistic function to fit the connection between the student and the item characteristics. The model of IRT is denoted as follows:

$$P(\theta) = c + \frac{1}{1 + e^{-\mathrm{D} \times a(\theta - b)}}, \tag{1}$$

where θ represents the latent trait, that is, knowledge states of a students [27]. Moreover, a denotes the discrimination of an item; while b denotes the difficulty; and parameter c is generally referred to the "guess parameter", as it indicates the accuracy of the response when a student totally guesses the item.

3.2 Problem Definition

In this section, we formally introduce the issue of Federated Deep Knowledge Tracing. In our focused educational scenarios, there are $|S|$ schools isolated. In a specific school s, there are $|N_s|$ students who process not exactly the same $|Q_s|$ exercises as other schools. Specifically, we define the learning records of a student as $r = \{(q_0, g_0), (q_1, g_1), \cdots, (q_l, g_l)\}$, where $q_l \in Q_s$ represents the exercise practiced by the student at time l, and g_l denotes the corresponding score. Generally, if the student correctly answers exercise q_l, $g_l = 1$; otherwise, $g_l = 0$. All exercises are derived from K concepts (e.g., "*Mode*"), which are confessed and consistent for all schools in practical educational scenarios. In our problem, we aim to train $|S|$ local DKT models, i.e., $\{\Theta_1, \Theta_2, \cdots, \Theta_s\}$ for each school, where the s-th DKT model Θ_s can trace the students in school s of knowledge states (represent the students' mastery of concepts). Please note that in practical scenarios, most learning data in schools is distributed and proprietary so that it is difficult for schools to gather or share data with others in this case, which results in data isolation.

4 FEDERATED DEEP KNOWLEDGE TRACING FRAMEWORK

In this section, we first illustrate the pipeline of the proposed Federated Deep Knowledge (FDKT) framework. Then we further introduce each part of our proposed models in following sections.

4.1 Model Overview

To solve the problem mentioned, we propose a novel Federated Deep Knowledge Tracing (FDKT) framework, which is a client-server architecture as illustrated in Figure 2. It is an iteration process. Specifically, at each round, each client is responsible for two tasks: (1) training an independent DKT with the private data; (2) evaluating the data quality with confidence measurements; and the center server also has two steps: (1) receiving and aggregating local DKTs delivered; (2) composing and updating the models for local clients. We will introduce the technical details of the client part and server part in FDKT, respectively.

4.2 Client Design

Each client completes two processes, i.e., local DKT training and data quality evaluation with local data.

4.2.1 Local DKT modeling. Deep Knowledge Tracing (DKT) is one of the state-of-the-art models used to trace student's knowledge states with recurrent neural networks (RNN) [35]. In our client

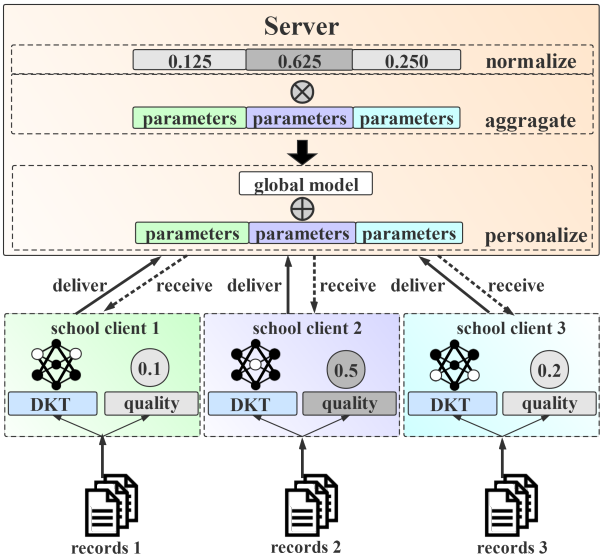


Figure 2: Federated Deep Knowledge Tracing Framework

part, we totally need to train $|S|$ DKT models for $|S|$ schools independently. Specifically, given the learning records of a certain student, DKT uses a RNN to model her knowledge presentations $\{h_1, h_2, \cdots, h_l\}$ and output her knowledge states (mastery levels) $\{y_1, y_2, \cdots, y_l\}$ on multiple concepts, which can be denoted as:

$$h_l = \tanh(\mathbf{W_{hx}} x_l + \mathbf{W_{hh}} h_{l-1} + \mathbf{b_h}),$$
$$y_l = \text{sigmoid}(\mathbf{W_{yh}} h_l + \mathbf{b_y}), \tag{2}$$

where the input of DKT, $x_l \in \{0,1\}^{2K}$ is the one-hot encoding of tuple (q_l, g_l), which represents the combination of which concept of item is answered and whether the item is answered correctly. Moreover, we define the model parameters as $\mathbf{\Theta} = \{\mathbf{W_{hx}}, \mathbf{W_{hh}}, \mathbf{b_h}, \mathbf{W_{yh}}, \mathbf{b_y}\}$. For training DKT models, we typically treat student performance prediction as the objective [26].

4.2.2 Data quality evaluation. In practice, training a high-quality DKT usually requires abundant data [32]. However, our scenario suffers from the data isolation problem, such that the DKT model of a certain client only expects to use the local learning data of the corresponding silo (e.g., school). Therefore, the centralized training strategy of traditional DKT is infeasible for our problem.

We address the problem with federated learning to collectively improve performances of all local DKTs. Traditionally, existing federated learning methods such as FedSGD and FedAvg [30] focus primarily on aggregating local models together while referring to data scales. However, as illustrated in Figure 1, the inconsistency of learning schedule among schools leads to a bias of data quality, which causes problems when aggregating bad models from low-quality data. Then, we propose two data quality evaluation methods with confidence estimation. They follow educational measurement theories that are commonly used in item quality analysis, i.e., Classical Test Theory and Item Response Theory, respectively.

(1) CTT confidence. We implement an aggregation strategy with Classical Test Theory (CTT), which evaluates the data quality from item statistics perspective as CTT confidence. Generally, we define the CTT confidence α_{CTT} for local data on school s as:

$$\alpha_{CTT} = F(P(Q_s), D(Q_s), CR(Q_s)), \tag{3}$$

where $F(\cdot)$ can be any workable function. Moreover, $P(Q_s)$, $D(Q_s)$, $CR(Q_s)$ are the difficulty, discrimination and reliability of all the Q_s items, respectively. The estimations of them are as following:

- *Difficulty*: Difficulty reflects how difficult an item is [8]. Specifically, *extreme group* is a more effective method to estimate difficulty. With extreme group method, the difficulty P_i of item i is denoted as: $P_i = (P_i^H + P_i^L)/2$, where P_i^H (P_i^L) is the average score on item i of students with higher (lower) scores. Therefore, the difficulty of data $P(Q_s)$ in school s with the Q_s items can be calculated as:
$$P(Q_s) = -\log \sum_{i=1}^{|Q_s|} \beta_i \times |P_i - P_0|, \tag{4}$$
where β_i indicates the ratio of item i occurring in the data of school s, and P_0 is a reference value. Under most circumstances, researchers usually ensure that item difficulty is close to a specific value in order to control the test effect.
- *Discrimination*: Discrimination is an indicator of how much an item can distinguish the mastery on concepts of students [8]. Specifically, the discrimination of data, $D(Q_s)$, in school s with the Q_s items can be calculated in a similar way to the difficulty factor with extreme group method as:
$$D(Q_s) = \sum_{i=1}^{|Q_s|} \beta_i \times D_i, \tag{5}$$
where the discrimination D_i of item i can be calculated as: $D_i = P_i^H - P_i^L$.
- *Reliability*: Reliability reflects the consistency of items and the Cronbach α reliability coefficient is currently one of the most commonly used reliability coefficients [7]. Here, we calculate the reliability of data $CR(Q_s)$ in school s following the commonly used Cronbach coefficient as:
$$CR(Q_s) = \frac{|Q_s|}{|Q_s| - 1} \times (1 - \frac{\sum_{i=1}^{|Q_s|} \beta_i \times S_i^2}{S_T^2}), \tag{6}$$
where S_i^2 and S_T^2 are the variance of the average scores of item i and the variance of the average total scores.

Following the estimation of difficulty, discrimination and reliability, we simply implement a general workable function $F(\cdot)$ by multiplying all three statistical factors as (please note that comparing different $F(\cdot)$ is not the main focus of this work):

$$\alpha_{CTT} = P(Q_s) \times D(Q_s) \times CR(Q_s). \tag{7}$$

CTT confidence presents a straightforward way of estimating data quality from the statistical perspective. However, there are some limitations mentioned. In the following, we propose another sophisticated data quality evaluation using IRT confidence.

(2) IRT confidence. We also implement an aggregation strategy with Item response theory (IRT), which evaluates data quality based on the item information as IRT confidence. Specifically, we define the IRT confidence α_{IRT} for local data on school s as:

$$\alpha_{IRT} = \max(\sum_{i=1}^{|Q_s|} \beta_i \times I_i(\theta)), \tag{8}$$

Algorithm 1 FDKT. The $|S|$ schools are indexed by s. B is the local mini-batch size, E is the number of local rounds, and η is the learning rate.

Server executes:
initialize $\mathbf{\Theta}^0$.
for each round t = 1, 2,... **do**
  for each client index $s \in S$ **in parallel do**
    $\mathbf{\Theta}_s^{t+1}, \alpha_s^{t+1} \leftarrow$ **ClientUpdate**$(s, \mathbf{\Theta}^t)$
    $\hat{\alpha}_s^{t+1} = \frac{\alpha_s^{t+1}}{\sum_{i=1}^{S} \alpha_i^{t+1}}$
  $\mathbf{\Theta}^{t+1} \leftarrow \sum_{s=1}^{S} \hat{\alpha}_s^{t+1} \times \mathbf{\Theta}_s^{t+1}$ by Eq. (11)
  for each client index $s \in S$ **in parallel do**
    for each layer $l = 1, 2, \ldots$ **do**
      $\lambda^l = \frac{\Theta_s^{t+1,l} \cdot \Theta^{t+1,l}}{\|\Theta_s^{t+1,l}\| \times \|\Theta^{t+1,l}\|}$ by Eq. (13)
    $\mathbf{\Theta}_s^{t+1} = \boldsymbol{\lambda} \cdot \mathbf{\Theta}_s^{t+1} + (1 - \boldsymbol{\lambda}) \cdot \mathbf{\Theta}^{t+1}$ by Eq. (12)

ClientUpdate(s,$\mathbf{\Theta}$):
estimate α_{CTT} by Eq. (7) or α_{IRT} by Eq. (8)
$\mathbf{B} \leftarrow$ (split dataset into batches of size B)
for each local round i from 1 to E **do**
  for batch $b \in \mathbf{B}$ **do**
    $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \eta \nabla l(\mathbf{\Theta}; b)$
return parameters $\mathbf{\Theta}$ and confidence α to server

where β_i indicates the ratio of occurrence as mentioned and $I_i(\theta)$ is the information function of item i, which can be calculated as:

$$I_i(\theta) = \frac{(P_i'(\theta))^2}{P_i(\theta)(1 - P_i(\theta))}. \tag{9}$$

Here, to obtain $I_i(\theta)$, we learn an IRT model with an Item Characteristic Curve, $P_i(\theta)$, that combines both the student's parameter θ and item i's parameters a_i, b_i and c_i. Specifically, $P_i(\theta)$ following Eq. (1) is denoted as:

$$P_i(\theta) = c_i + \frac{1}{1 + e^{-\mathrm{D} \times a_i(\theta - b_i)}}, \tag{10}$$

where θ refers to the latent trait of a certain student, while a_i, b_i, c_i are factors in the estimated model based on the so-called discrimination, difficulty and guess factors mentioned above.

4.3 Server Design

In FDKT, we design the server to be responsible for two stage tasks, i.e., appropriately aggregating the local DKTs and adaptively updating models.

4.3.1 Model Aggregation. In round t, the center server first receives two parts of information from all the clients: (1) all the local confidences: $\{\alpha_1^t, \alpha_2^t, ..., \alpha_s^t\}$ (α_s^t can be either CTT confidence α_{CTT} by Eq. (7) or IRT confidence α_{IRT} by Eq. (8) in school s); (2) all the local DKT models: $\{\mathbf{\Theta}_1^t, \mathbf{\Theta}_2^t, \cdots, \mathbf{\Theta}_s^t\}$.

Then the server integrates all local DKT models to a global one $\mathbf{\Theta}^t$. We follow the general setting of naive parameter averaging [30] but innovatively adopt data quality instead of data scales for aggregation. Then we perform the average of model parameters on the current communication round t as follows:

$$\mathbf{\Theta}^t = \sum_{s=1}^{S} \hat{\alpha_s^t} \times \mathbf{\Theta}_s^t, \tag{11}$$

Table 1: The statistics of two datasets: MATH and ASSIST.

Statistics	MATH	ASSIST
# of schools	7	38
# of records	204,293	801,645
# of students	3,830	7,395
# of exercises	4,145	27,288
# of knowledge concepts	112	200
Avg. records per student	53.34	108.40
Avg. exercises per concept	37.01	136.44

where $\hat{\alpha_s^t}$ is the normalized confidence of school s at round t: $\hat{\alpha_s^t} = \alpha_s^t / \sum_{i=1}^{S} \alpha_i^t$. Moreover, through the model aggregation process, we will integrate models and perform comparable results among all clients, which will make the global model meaningful.

4.3.2 Model Update. The second process of the center server is to update the models for clients before next round. The traditional global model in federated learning is a general model that is expected to fit the overall data distribution from all schools. However, as mentioned above, in educational scenarios, inconsistent properties and settings cause Non-IID characteristics so that it is difficult to fit all items with a uniform global model [22]. In FDKT, to better suit Non-IID local data for each client, at round t, after aggregating the global model $\mathbf{\Theta}^t$, we adopt hierarchical model interpolation to fine-tune the global model. In particular, at round t, we obtain the personalized model from the global model $\mathbf{\Theta}^t$ and the local model $\mathbf{\Theta}_s^t$ from school s as:

$$\mathbf{\Theta}_s^t = \boldsymbol{\lambda} \cdot \mathbf{\Theta}_s^t + (1 - \boldsymbol{\lambda}) \cdot \mathbf{\Theta}^t, \tag{12}$$

where $\boldsymbol{\lambda}$ is the interpolated weight. In our work, we calculate the interpolated weight to a vector instead of a simple scalar quantity [29]. Following this approach, we measure the differences between the global model and the local model by layers with cosine similarity. We here primarily focus on the global model, since it has more comprehensive information from clients. If the global model is more similar to the local model, we adopt some effects of local model by integrating local model parameters. The computation of interpolated weight $\boldsymbol{\lambda}$ with the neural layers l, can be denoted as:

$$\lambda^l = \frac{\Theta_s^{t,l} \cdot \Theta^{t,l}}{\|\Theta_s^{t,l}\| \times \|\Theta^{t,l}\|}. \tag{13}$$

With the hierarchical model interpolation, we design a personalized model update strategy. In this way, we effectively retain the personalized information of the models and make the local model fit better with the private data.

As mentioned above, the workflow of FDKT is presented in Algorithm 1. In summary, our proposed FDKT framework does not gather or share the data across schools. Instead, it simply delivers the model parameters of local DKTs. Therefore, it effectively alleviates data scarcity while protecting the data privacy in training. To our best knowledge, FDKT is the first attempt to leverage federated learning framework for knowledge tracing in intelligent education.

5 EXPERIMENTS

In this section, we first introduce our experimental datasets and setups. Then, we report our experimental results from the following

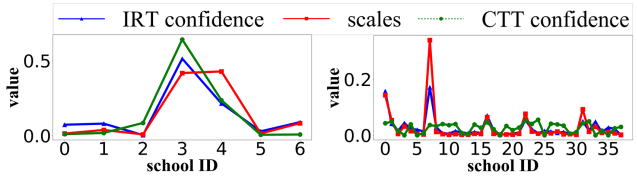
Table 2: Results of student performance prediction under four metrics.

(a) Results of student performance prediction on MATH

method \ dataset	MATH			
	epoch	RMSE	AUC	ACC
BKT	-	0.463	0.692	0.701
DKT	-	0.453	0.705	0.712
FedSGD	-	0.455	0.696	0.694
FedAvg	13	0.449	0.721	0.713
FedAtt	14	0.453	0.718	0.708
LoAdaboost	8	0.450	0.726	0.708
FedInter	8	0.449	0.733	0.719
FDKTCTT	**4**	0.448	0.735	0.717
FDKTIRT	**4**	**0.446**	**0.739**	**0.721**

(b) Results of student performance prediction on ASSIST

method \ dataset	ASSIST			
	epoch	RMSE	AUC	ACC
BKT	-	0.452	0.743	0.681
DKT	-	0.413	0.814	0.75
FedSGD	-	0.425	0.798	0.746
FedAvg	20	0.387	0.861	0.791
FedAtt	22	0.386	0.862	0.792
LoAdaboost	28	0.384	0.863	0.792
FedInter	**11**	0.376	0.875	0.796
FDKTCTT	17	0.379	0.872	0.795
FDKTIRT	**11**	**0.375**	**0.877**	**0.802**



(a) MATH (b) ASSIST

Figure 3: Distributions of confidence values and data scales of two datasets: MATH (left), ASSIST (right).

three aspects: (1) the overall performances of knowledge tracing models; (2) the effectiveness of data quality on deep knowledge tracing; (3) the performances on comparability among schools.

5.1 Experimental Datasets

In our experiments, we use two real-world datasets, namely MATH and ASSIST. MATH is a private dataset collected from daily exercise records of senior high school students on mathematics problems from 2016 to 2017. ASSIST (short for Assistments) is a public dataset, i.e., 2009-2010 "Non-skill builder"[1], which records the mathematics learning logs from an online tutoring program.

MATH consists of about 200,000 records of 3,830 students in 7 schools. All items in it belong to 112 concepts, such as "Set" and "Vector". For ASSIST, we divide the data by school id and filter out the schools whose records are fewer than 1000. After preprocessing, we obtain about 800,000 learning records of 7,395 students in 38 schools. The items in it belong to 200 concepts, such as " Range" and "Proportion". More statistics of our dataset are presented in Table 1. In our scenario, each school only holds the data subset belong to it, which causes data isolation. If we can train DKT models independently with distributed datasets and aggregate the effects of all the DKT models, it will benefit all the clients equally by avoiding the data isolation problem. It is equivalent to expanding the available training sets.

Furthermore, we deeply analyze both datasets in Figure 3. Here, we present the distributions of data scales and data quality (reflected by CTT confidence and IRT confidence) in different schools. From the figure, we can see that there is an inconsistency between data scales and data quality, which means that larger datasets do not necessarily have higher data quality. Therefore, it is necessary to consider data quality for learning DKT in educational scenarios.

[1]https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/non_skill-builder-data-2009-2010

5.2 Experimental Settings

5.2.1 Data partition. In both MATH and ASSIST, we randomly partition 90% of students learning records in every school for training, while the remainders are for testing. It is worth noting that we do not put the data in schools together, but leave them isolated.

5.2.2 FDKT Setting. We specify the framework setups in FDKT, including the DKT settings, data quality evaluation settings and federated learning settings. For DKT part of FDKT, it is designed to a general structure with one hidden layer of 50 dimensions. The dimension of output is equal to the total number of concepts, and the dimension of input is double. For data quality evaluation settings, we select 30% higher (lower) grade students as the high (low) group and set P_0 as 0.5 in CTT confidence estimation. Moreover, we set parameter D to 1.7 in IRT confidence estimation. For federated learning settings, the models are all trained under the same settings, with batch size of 64, local rounds of 5 (except FedSGD) and initial learning rate of 0.001. To facilitate further research in FDKT, we have published our code[2].

5.2.3 Baseline Approaches. To demonstrate the effectiveness of the FDKT framework, we first compare two typical KT methods without federated settings, i.e., **BKT**, **DKT**, both of which are trained independently only with private data of each school.

- **BKT** [6] is a kind of Hidden Markov Model (HMM) that models students' knowledge states as a set of binary values.
- **DKT** [35] is a deep-learning knowledge tracing model that integrates recurrent neural networks to model knowledge states with sequential learning data.

Then, we compare some state-of-the-art federated learning methods, which mainly focus on statistics, to verify the effectiveness of our strategies. All local models in federated methods are DKT.

- **FedSGD** [30] is a method based on iterative optimization. Each client model takes one step of gradient descent on the current local model, after which the server takes a scale weighted average of all models.
- **FedAvg** [30] also aggregates models by data scales. However, FedAvg assigns more computation to clients by iterating the local update with E rounds and B batch size.
- **FedAtt** [16] is a method incorporating soft-attention. The main idea is to consider the importance of models through

[2]https://github.com/bigdata-ustc/federated-deep-knowledge-tracing

aggregating by layers with the weights of distances between the global model and local models.
- **LoAdaboost** [12] is an adaptive boosting training method designed to produce additional training for clients that still have losses higher than those with median loss, until their loss is lower than the median loss.

Besides, we also introduce a method only with hierarchical model interpolation for personalization without data quality evaluation to highlight the effectiveness of data quality, which can be viewed as a variant of our FDKT model, denoted as **FedInter**.

Generally, all baselines only focus on data scales for model aggregation. We define two methods based on FDKT with CTT confidence and IRT confidence for aggregation, respectively, and both methods consider local model personalization, which are named **FDKTCTT** and **FDKTIRT**. For fairness, all methods are implemented by Python, and all experiments are run on a Linux server with two NVIDIA Tesla K80 GPUs and 256G memory to achieve the best performance in the following experiments.

5.2.4 Evaluation metrics. To observe the effectiveness of the FDKT framework, we use the widely-used ROC Curve (AUC), Prediction Accuracy (ACC), and Root Mean Square Error (RMSE) metrics on both regression and classification perspectives to measure how approximative the DKT prediction is to the ground truth [26]. Among them, AUC and ACC are commonly used for classification tasks with the range of $[0, 1]$, the larger the values are, the better the results. Moreover, RMSE is commonly used for regression tasks whose range is $[0, 1]$, the lower the value is, the better the result.

Following existing federated learning works [30], we evaluate convergence efficiency with efficiency metric "epoch". That is, we hope to evaluate how many epochs are required when the model reaches the high-level targeted AUC performance. The less the number is, the faster the model reaches the target performance, which means the better efficiency the model has. Note that in our experiments, we set the target AUC value as 0.70 (0.85) AUC in MATH (ASSIST).

5.3 Experimental Results

5.3.1 Overall performance. To evaluate the performances of all the above methods in isolated schools, we conduct the typical student performance prediction task [26], which asks us to train knowledge tracing models and predict the future performance of each student in different schools. We repeat the experiments 5 times and summarize the average of results. Table 2 reports the overall results on both datasets with all evaluation metrics mentioned.

Some key observations as follows: (1) Methods with federated learning settings perform better than those training with clients' private data independently. It shows federated learning settings that can harness more data usually result in better DKT models. Notably, FDKTIRT has the best performances on both datasets while FDKTCTT shows comparable results. This means that our methods can more effectively alleviate the data isolation problem for DKT with federated learning settings. (2) Compared with data scales-based methods, FedInter achieves relatively good performances, which demonstrates the effectiveness of model personalization. While FDKTIRT has the best performances, it shows quality-oriented aggregation with personalization update is beneficial for training DKTs and necessary in educational scenarios. (3) FDKTCTT performs no most outstanding result, which demonstrates the limitations of CTT confidence. It cannot achieve comprehensive quality evaluation, making it less suitable for our purposes than IRT confidence.

As mentioned before, communication cost is important in federated learning. Indeed, our method has the most significant performances on the communication costs among all federated learning methods with the metric, epoch. As shown in Table 2, FDKTIRT reaches target AUCs faster in both datasets, while FedInter may reach slower. It demonstrates FDKTIRT can produce a dramatic decrease in communication costs, while ignoring data quality causes palliation in convergence of FedInter.

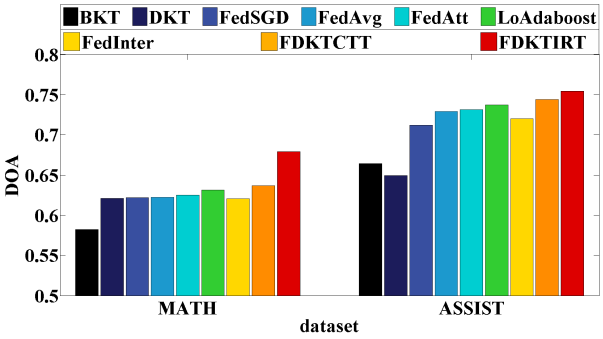
5.3.2 Effectiveness of data quality. We analyze the effectiveness of data quality in depth by CTT confidence and IRT confidence. We report the performances of different methods on isolated schools on ASSIST with the AUC metric in Table 3. For better illustration, we choose 6 representative schools by stratification with different scales of data in this experiment, as two larger-scaled ones (i.e., school 1 and 2), two medium-sized ones (i.e., school 3 and 4) and two smaller ones (i.e., school 5 and 6). Then we list the statistics of data scales, and two data quality values, i.e., CTT confidence and IRT confidence values. We choose DKT, LoAdaboost and FedInter as baselines. Here, DKT is an independently trained method on each school and LoAdaboost is the most solid baseline.

From the table, we can derive the following observations among samples: (1) When comparing DKT with federated learning methods, DKT performs significantly poorly in schools with extremely small amounts of data (i.e., school 5 and school 6), which means federated learning settings can expand the available data and improve DKTs performances. (2) FedInter achieves competitive results, which further proves the effectiveness of model personalization. (3) In particular, FDKTIRT and FDKTCTT perform better than the most solid baseline, LoAdaboost. It proves that compared with the amount of data, data quality is a more important factor on behalf of the importance of model effect for DKTs aggregation. Specially, FDKTIRT performs better than FDKTCTT meaning IRT confidence is the more accurate and consistent evaluation to data quality. We can also find that federated learning settings may result in a reduction on large-scale data (i.e., school 1), while our methods can better alleviate it. In summary, we can conclude that FDKT is both effective and efficient in training DKTs while considering quality-oriented aggregation with personalization update strategies. Meanwhile, according to our analysis, data scales and data quality in the data are inconsistent. Our phenomenon shows that data quality is more in line with educational scenario.

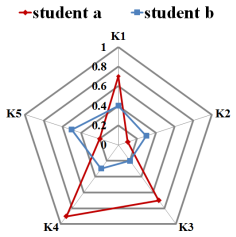
5.3.3 Performance on comparability. As we argued earlier, the comparability of students in different schools is important in educational scenarios. We convert this measurement into a ranking problem following [24]. Intuitively, if one KT model diagnoses that student a in school s_1 masters better than student b in school s_2 on knowledge concept k, she may have a higher probability of responding correctly to exercises related to concept k than student b. We adopt the Degree of Agreement (DOA) [5] metric to evaluate the ranking performance of each knowledge tracing model. Specifically, a DOA result on a specific knowledge k is defined as:

Table 3: Statistics of confidence, scale and results of student performance prediction with AUC of partial datasets.

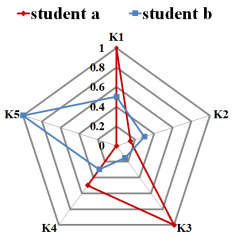
category	name	school 1	school 2	school 3	school 4	school 5	school 6
statistic	scales	114,627	42,899	10554	9,743	3103	1,163
	CTT confidence	0.043	0.048	0.041	0.049	0.034	0.012
	IRT confidence	0.157	0.041	0.022	0.016	0.011	0.010
method	DKT	0.871	0.830	0.838	0.809	0.798	0.535
	LoAdaboost	0.858(-1.3%)	0.816(-1.4%)	0.811(-2.7%)	0.796(-1.3%)	0.865(+6.7%)	0.737(+20.2%)
	FedInter	0.876(+0.5%)	0.843(+1.3%)	0.846(**+0.8%**)	0.867(+5.8%)	0.927(+12.9%)	0.801(+26.6%)
	FDKTCTT	0.871(+0.0%)	0.844(+1.4%)	0.842(+0.4%)	0.871(+6.2%)	0.934(+13.6%)	0.801(+26.6%)
	FDKTIRT	0.879(**+0.8%**)	0.848(**+1.8%**)	0.846(**+0.8%**)	0.875(**+6.6%**)	0.937(**+13.9%**)	0.805(**+27.0%**)



(a) Overall performances on comparability with metric DOA.



(b) Example: performances on states.



(c) Example: performances on scores.

Figure 4: Left bar chart is DOA results of methods. Right radars are comparable examples of two students' knowledge states and true scores from isolated schools. (K1: Scatter Plot; K2: Proportion; K3: Point Plotting; K4: Graph shape; K5: Congruence;)

$$DOA(k) = \sum_{a=1}^{|N_{s_1}|} \sum_{b=1}^{|N_{s_2}|} I_{abk} \frac{\delta(y_{ak}, y_{bk}) \cap \delta(\bar{g}_{ak}, \bar{g}_{bk})}{\delta(y_{ak}, y_{bk})} \quad (14)$$

Here, $|N_{s_1}|$ and $|N_{s_2}|$ denote the numbers of students in school s_1 and s_2, while y_{ak} indicates the knowledge state of student a on knowledge concept k obtained by the output of DKT models (Eq. 2), and $\bar{g}_{ak}$ is the average score of student a on concept k. $\delta(x, y)$ is an indicator function, where $\delta(x, y)$=1, if $x > y$; otherwise, $\delta(x, y)$=0. I_{abk} is another indicator function, where I_{abk}=1 if both students have learned the concept k before. Furthermore, we average the DOA(k) of all concepts as DOA to measure the overall results, which is denoted as $DOA = \sum_{k=1}^{K} DOA(k)/K$, $DOA \in [0.0, 1.0]$, the larger the DOA, the better the performance.

Figure 4(a) illustrates the overall performances on DOA. We can conclude the following from the results: (1) BKT and DKT perform worst on both datasets, meaning that independent training on isolated schools is incomparable with students in different schools. (2) All federated learning methods perform better than BKT and DKT, demonstrates that it is effective with federated learning strategies for knowledge tracing. (3) FDKTIRT performs best, followed by FDKTCTT, which demonstrates that the proposed different quality-oriented aggregation strategies with personalization update strategy are more effective at achieving comparable results among all clients, which can be adapted to practical educational scenarios. Moreover, FedInter does not perform very well on both datasets, which demonstrates that ignoring quality-oriented aggregation will damage comparability among DKT models.

Moreover, we take an example in Figure 4. For better illustration, we visualize the state outputs of the local DKT model of two example students on 5 concepts in Figure 4(b) (two examples of scores are shown in Figure 4(c)). We can observe that student a masters better in K1, K3 and K4. Correspondingly, she performs better on scores of these concepts. This shows the comparability of students from different schools.

6 CONCLUSION

In this paper, we designed a novel client-server architecture framework, called Federated Deep Knowledge Tracing (FDKT), to solve the critical problem faced by current DKT tasks: data isolation problem. Specifically, we combined federated learning to train DKT models while alleviating data scarcity. Subsequently, in the client part, two implementations of quality-oriented aggregation strategies under the framework were provided; in the server part, hierarchical model interpolation was explored for personalized model update. Finally, our quantitative experiments showed that FDKT has achieved significant improvements, which demonstrated that high-quality DKT models can be trained in federated settings and obtain great comparable results on real-world data.

In the future, we will explore more applications of federated learning in the educational field, extend FDKT to many other KT methods and develop a general platform. Moreover, we hope to explore ways to model item and user characteristics appropriately under federated settings.

ACKNOWLEDGMENTS

This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), the National Natural Science Foundation of China (Grants No. 61922073, 61976198 and 62022077), the Foundation of State Key Laboratory of Cognitive Intelligence (Grant No. iED2020-M004), and the Iflytek joint research program. Haiping Ma gratefully acknowledges the support of the CCF-Tencent Open Research Fund.

REFERENCES

[1] Ghodai Abdelrahman and Qing Wang. 2019. Knowledge Tracing with Sequential Key-Value Memory Networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 175–184.

[2] Nabeel Abedalaziz and Chin Hai Leng. 2018. The relationship between CTT and IRT approaches in Analyzing Item Characteristics. *MOJES: Malaysian Online Journal of Educational Sciences* 1, 1 (2018), 64–70.

[3] Allan Birnbaum. 1969. Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology* 6, 2 (1969), 258–276.

[4] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, 177–186.

[5] Yuying Chen, Qi Liu, Zhenya Huang, Le Wu, Enhong Chen, Runze Wu, Yu Su, and Guoping Hu. 2017. Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 989–998.

[6] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[7] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[8] Barbara B Ellis and Alan D Mead. 2002. Item analysis: Theory and practice using classical and modern test theory. (2002).

[9] Harold Gulliksen. 1950. Theory of mental tests. (1950).

[10] Nina Guyon, Eric Maurin, and Sandra McNally. 2012. The effect of tracking students by ability into different schools a natural experiment. *Journal of Human Resources* 47, 3 (2012), 684–721.

[11] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. 2018. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604* (2018).

[12] Li Huang, Yifeng Yin, Zeng Fu, Shifa Zhang, Hao Deng, and Dianbo Liu. 2018. LoAdaBoost: Loss-Based AdaBoost Federated Machine Learning on medical Data. *arXiv preprint arXiv:1811.12629* (2018).

[13] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests.. In *AAAI*. 1352–1359.

[14] Zhenya Huang, Qi Liu, Yuying Chen, Le Wu, Keli Xiao, Enhong Chen, Haiping Ma, and Guoping Hu. 2020. Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students. *ACM Transactions on Information Systems (TOIS)* 38, 2 (2020), 1–33.

[15] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. 2018. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479* (2018).

[16] Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang. 2019. Learning private neural language modeling with attentive aggregation. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[17] Di Jiang, Yuanfeng Song, Yongxin Tong, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. 2019. Federated Topic Modeling. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1071–1080.

[18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977* (2019).

[19] Eugene Kharitonov. 2019. Federated online learning to rank with evolution strategies. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 249–257.

[20] G Gage Kingsbury and David J Weiss. 1983. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In *New horizons in testing*. Elsevier, 257–283.

[21] George D Kuh, Jillian Kinzie, Jennifer A Buckley, Brian K Bridges, and John C Hayek. 2011. *Piecing together the student success puzzle: research, propositions, and recommendations: ASHE Higher Education Report*. Vol. 116. John Wiley & Sons.

[22] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2018. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127* (2018).

[23] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).

[24] Qi Liu, Enhong Chen, Hui Xiong, Chris HQ Ding, and Jian Chen. 2011. Enhancing collaborative filtering by user interest expansion via personalized ranking. *T SYST MAN* 42, 1 (2011), 218–233.

[25] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. 2011. Personalized travel package recommendation. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 407–416.

[26] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering* 33, 1 (2019), 100–115.

[27] Qi Liu, Runze Wu, Enhong Chen, Guandong Xu, Yu Su, Zhigang Chen, and Guoping Hu. 2018. Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST)* 9, 4 (2018), 1–26.

[28] FM Lord, MR Novick, and Allan Birnbaum. 1968. Statistical theories of mental test scores. (1968).

[29] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. 2020. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619* (2020).

[30] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).

[31] Kritphong Mongkhonvanit, Klint Kanopka, and David Lang. 2019. Deep Knowledge Tracing and Engagement with MOOCs. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, 340–342.

[32] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data* 2, 1 (2015), 1.

[33] Hiromi Nakagawa, Yusuke Iwasawa, and Yutaka Matsuo. 2019. Graph-based Knowledge Tracing: Modeling Student Proficiency Using Graph Neural Network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 156–163.

[34] Zachary A Pardos and Neil T Heffernan. 2011. KT-IDEM: introducing item difficulty to the knowledge tracing model. In *International conference on user modeling, adaptation, and personalization*. Springer, 243–254.

[35] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in neural information processing systems*. 505–513.

[36] James J Ryan. 1968. Teacher judgments of test item properties. *Journal of Educational Measurement* 5, 4 (1968), 301–306.

[37] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2019. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems* (2019).

[38] Jessica Sharkness and Linda DeAngelo. 2011. Measuring student involvement: A comparison of classical test theory and item response theory in the construction of scales from student surveys. *Research in Higher Education* 52, 5 (2011), 480–507.

[39] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6153–6161.

[40] Hao Wang, Tong Xu, Qi Liu, Defu Lian, Enhong Chen, Dongfang Du, Han Wu, and Wen Su. 2019. MCNE: An end-to-end framework for learning multiple conditional network representations of social network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1064–1072.

[41] Lisa Wang, Angela Sy, Larry Liu, and Chris Piech. 2017. Deep knowledge tracing on programming exercises. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 201–204.

[42] Xiaolu Xiong, Siyuan Zhao, Eric G Van Inwegen, and Joseph E Beck. 2016. Going deeper with deep knowledge tracing. *International Educational Data Mining Society* (2016).

[43] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 12.

[44] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903* (2018).

[45] Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. 2019. FFD: A Federated Learning Based Method for Credit Card Fraud Detection. In *International Conference on Big Data*. Springer, 18–32.

[46] Chun-Kit Yeung and Dit-Yan Yeung. 2018. Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10.

[47] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. 2013. Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education*. Springer, 171–180.

[48] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. 2017. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 765–774.

[49] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582* (2018).