Group-Level Cognitive Diagnosis: A Multi-Task Learning Perspective

Jie Huang[1], Qi Liu[1,*], Fei Wang[1], Zhenya Huang[1], Songtao Fang[1], Runze Wu[2], Enhong Chen[1], Yu Su[1,3], 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
{jiehuang, wf314159, songtao}@mail.ustc.edu.cn, {qiliuql, huangzhy, cheneh}@ustc.edu.cn
[2]Fuxi AI Lab, NetEase Inc., Hangzhou, China, wurunze1@corp.netease.com
[3]IFLYTEK Research, {yusu, sjwang3}@iflytek.com

2021 IEEE International Conference on Data Mining (ICDM) | 978-1-6654-2398-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDM51629.2021.00031

***Abstract*—Most cognitive diagnosis research in education has been concentrated on individual assessment, aiming at discovering the latent characteristics of students. However, in many real-world scenarios, group-level assessment is an important and meaningful task, e.g., class assessment in different regions can discover the difference of teaching level in different contexts. In this work, we consider assessing cognitive ability for a group of students, which aims to mine groups' proficiency on specific knowledge concepts. The significant challenge in this task is the sparsity of group-exercise response data, which seriously affects the assessment performance. Existing works either do not make effective use of additional student-exercise response data or fail to reasonably model the relationship between group ability and individual ability in different learning contexts, resulting in sub-optimal diagnosis results. To this end, we propose a general Multi-Task based Group-Level Cognitive Diagnosis (MGCD) framework, which is featured with three special designs: 1) We jointly model student-exercise responses and group-exercise responses in a multi-task manner to alleviate the sparsity of group-exercise responses; 2) We design a context-aware attention network to model the relationship between student knowledge state and group knowledge state in different contexts; 3) We model an interpretable cognitive layer to obtain student ability, group ability and exercise factors (e.g., difficulty), and then we leverage neural networks to learn complex interaction functions among them. Extensive experiments on real-world datasets demonstrate the generality of MGCD and the effectiveness of our attention design and multi-task learning.**

***Index Terms*—Group-Level Cognitive Diagnosis, Multi-Task Learning, Attention Mechanism, Data Sparsity**

I. Introduction

Cognitive diagnosis (CD) has long been a crucial and fundamental task to explore and analyze the learning status of students in intelligent education systems, which is beneficial in improving students' learning proficiency. To date, most studies focused on modeling the cognitive state of each individual student, e.g., the student's proficiency on knowledge concepts [1], [2]. Nevertheless, in real-world situations, students often accomplish their academic goals through learning in groups (e.g., schools, classes, and study groups), which has been proven to bring more benefits for students [3]. Moreover, group-level teaching is still the most important and irreplaceable teaching method at present and will continue to play a critical role in the future. Therefore, it is of great value to analyze groups' learning states so as to assess and improve the teaching achievements. Moving beyond the traditional task of cognitive modeling for individuals, in this work we concentrate on modeling cognitive ability for a group of students, known as the group-level cognitive diagnosis (GCD).

Besides education, modeling the cognitive level of a group is a basic task in many research fields, such as games and medical evaluation [4], [5]. Specifically, in the educational context, GCD aims to mine groups' actual knowledge states. Figure 1 shows a toy process of GCD. Generally, a group of students first choose or are assigned to practice a set of exercises (e.g., taking a class test) and leave responses (e.g., true or false). Based on the Q-matrix (an exercise-knowledge correlation matrix labeled by educational experts) [6] and the response logs, our goal is to mine the group's proficiency on relevant knowledge concepts (e.g., 'Function'). In practice, these diagnostic reports could be further applied to many real-world applications, such as the teaching quality assessment [7] and the teaching plan improvement.

In the literature, there are many efforts on designing cognitive diagnosis models (CDMs), such as item response theory (IRT) [8], Multidimensional IRT (MIRT) [9], [10], Matrix Factorization (MF) [11] and Neural Cognitive Diagnosis (NeuralCD) framework [12], and most of them focus on modeling cognitive abilities for individual students. To perform cognitive diagnosis on the group level, a common solution is to treat the group as a single virtual unit and then select the traditional individual CDMs for assessment. Although technically feasible, such a straightforward solution cannot achieve good performance due to data sparsity of group-exercise responses (common exercises that all students in the group have practiced). Students will do different exercises according to their own abilities in daily learning and group-exercise response logs often come from group test evaluation. Therefore, the group-exercise responses are usually more sparse than student-exercise responses.

To alleviate the sparsity of group-exercise responses, an intuitive solution is to integrate the data of student-exercise responses which provide additional information about individual students' cognitive abilities. Existing related works are devoted

*Corresponding Author.

978-1-6654-2398-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDM51629.2021.00031

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

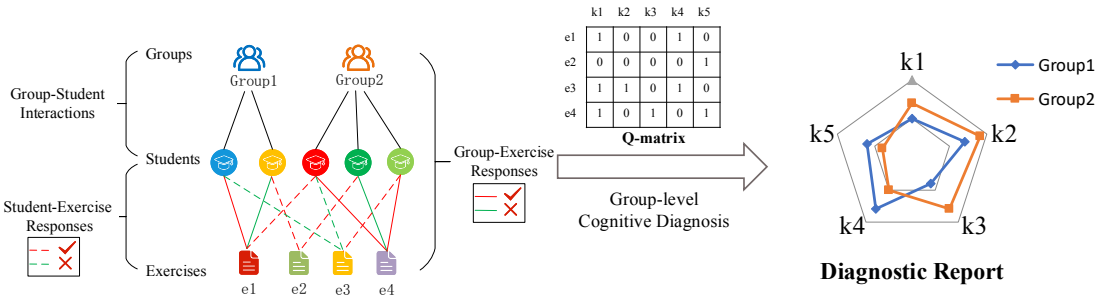


Fig. 1. A toy process of group-level cognitive diagnosis. The interaction data can be divided into two non-overlapping records: student-exercise responses (dotted line) and group-exercise responses (solid line). The Q-matrix is the correlation matrix between exercises and knowledge concepts, e.g., exercise e1 contains knowledge concepts k1 and k4. The diagnostic report is visualized as a radar chart, and each point represents the mastery level of the certain knowledge concept.

to modeling the ability of individuals in a group through their whole response data, and then regard the group's ability as the average ability of all these students [13], [14]. However, the validity of the obtained diagnosis results is greatly influenced by the distribution of exercises practiced by each student in a group. For instance, as shown in Figure 1, consider the knowledge concept $k1$, which is contained by $e1$, $e3$, $e4$ according to the Q-matrix. From the group-student-exercise interactions, we can observe that the number of exercises practiced by each student in group 2 is unbalanced. In practice, most traditional methods require a balance [15] in the exercise responses of students on each knowledge concept to reduce the overall assessment bias, but it is not practicable for complex educational scenarios. In addition, this predefined aggregation scheme is not suitable for adaptively modeling the correlation between groups and individuals in different learning contexts, where the influence weight of each student may be different. According to the research of educational behavior, one of the factors determining the influence of students in a group is their relative abilities [16]. For example, consider students $s1$ and $s2$ with similar abilities (knowledge states) in different groups (named as $g1$ and $g2$ respectively), if the overall ability level of $g1$ is much lower than that of $g2$, $s1$ may own a higher influence than $s2$. As such, to make more effective use of student-exercise response data, another challenge is how to model the relationship between the learning states of individual students and the group they belong to.

In this work, we propose a general multi-task based group-level cognitive diagnosis (MGCD) framework to handle these challenges. Specifically, we simultaneously model the student performance and group performance. The information from student-exercise logs is transferred to group representations through shared student representations, between which the relationship is modeled with attention mechanism. Moreover, inspired by the Neural Cognitive Diagnosis Model (Neural-CDM) [12], we leverage neural networks with monotonicity assumption to model the complex interactions on both student-exercise responses and group-exercise responses. Particularly, our MGCD is a general framework since it can be flexibly combined with different interaction functions from individual cognitive diagnosis models (e.g., MF, IRT, MIRT, Neural-CDM). Finally, extensive experiments are performed on real-world datasets to prove the effectiveness and interpretability of our method.

In summary, our key contributions are listed as follows:

- To the best of our knowledge, this is the first comprehensive attempt to apply deep learning to model the cognitive ability for a group of students. Specifically, we propose a novel context-aware attention network to adaptively model the relationship between student knowledge state and group knowledge state in different learning contexts.
- We propose a novel solution for GCD from a perspective of multi-task learning to effectively leverage the student-exercise response data to alleviate the sparsity of the group-exercise response data.
- We conduct extensive experiments on real-world datasets to validate the effectiveness of MGCD with both accuracy and interpretability guarantee.

II. Related Work

A. Cognitive Diagnosis

In educational psychology, a wide range of cognitive diagnostic models (CDMs) has been developed to provide fine-grained information about students' cognitive ability [17], [18]. The existing research mainly focuses on individual assessment, among which DINA [1], [19] and IRT [8] were two of the most typical works, which characterize students by latent traits. Specifically, in DINA, the latent trait was a binary vector, which denotes whether a student masters the knowledge concepts required by the problem. IRT regarded students' abilities as unidimensional and continuous latent traits and used a logistic function to model the probability that a student correctly solves a problem. MIRT [10] is an extension of IRT, which can characterize students' cognitive abilities through multidimensional latent traits. In addition, some works leveraged matrix factorization (MF) to obtain the latent trait vectors of students and exercises by decomposing the score matrix [20], [21]. Different from these traditional methods which try to model the interactions between students and exercises with linear functions, the Neural Cognitive Diagnosis Framework (NeuralCD) [12] is a pioneer work that incorporates neural networks to learn the complex high-order student-exercise interactions.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

B. Group-Level Cognitive Diagnosis

In recent years, group-level cognitive diagnosis (GCD) has received a lot of attentions and has been widely applied in various domains. Especially in international large-scale assessment projects (e.g. PISA and TIMSS) [22], [23], GCD is widely used to explore the differences of group-level teaching between different countries and areas. To our knowledge, there are two solutions for traditional GCD tasks. The first solution is to extend the traditional CDMS and then apply them to assess the group ability. Traditional group-level cognitive diagnosis models (GCDMs) were mainly developed based on GIRT [4], [24], [25], which was an extended study of IRT on the group level. Specifically, following an IRT-like logistic model, the correct rate of group i answering exercise j is $P\left(r_{ij}=1 \mid \theta_i\right)=\operatorname{sigmoid}\left(a_j\left(\theta_i-\beta_j\right)\right)$, which is a typical 2-parameter logistic (2PL) model [26]. Different from IRT, the GIRT combines matrix sampling [15], a student-exercise sampling method, to collect group response data, so the performance of the model is greatly affected by the sampling design. [25] utilized GIRT to explain group differences in mathematics achievement from an international perspective. In [4], GIRT was applied in medical evaluation, which evaluated the quality of health plan for different consumer groups. We should note that the traditional GIRT method describes a group with latent variables, which cannot provide intuitive and interpretative results for each group. The other solution for GCD is to first focus on modeling individual abilities, then assume that the collective ability of a group is the average ability level of its members [13], [14]. This solution generally assumes that the response records of students to each knowledge concept are evenly distributed in a group, so as to ensure the balance of ability aggregation. However, this assumption may not hold true in real and complex learning environments. Besides, this method lacks the ability to adaptively model the influence weights of students in different educational contexts, resulting in sub-optimal assessment results.

C. Multi-Task Learning

As a promising area in machine learning, multi-task learning (MTL) aims to leverage useful information contained in multiple learning tasks to help learn a more accurate learner for each task [27], [28]. The superior performance of MTL has been demonstrated in many fields, such as natural language process, computer vision [29], and recommendation system [30], [31]. Some previous work has shown that MTL is helpful to improve the performance of the main task by alleviating the data sparsity problem. [32] demonstrated that learning representations to predict the position and shape of facial landmarks could improve expression recognition. To improve the performance of bundle recommendation task, [31] proposed a multi-task neural network to share the information between two tasks (user-item modeling and user-bundle modeling). To alleviate the sparsity of group-item interactions, the AGREE model [30] is proposed to effectively leverage the user-item interactions by a MTL method.

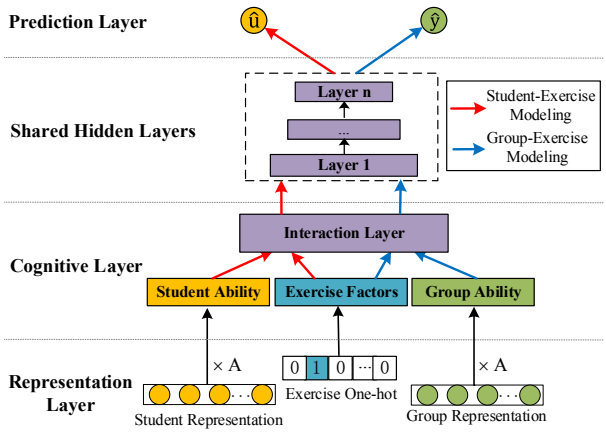


Fig. 2. Illustration of joint modeling on student-exercise response and group-exercise response.

Our work is orthogonal to the above-mentioned works, as we exploit the deep neural network to tackle the group-level cognitive diagnosis task under the multi-task learning framework. Moreover, we employ the attention mechanism to learn group representation as well. Besides, compared with the traditional GCDMs, our method has better interpretability, which can obtain the groups' mastery on the knowledge concepts for group assessment.

III. Problem Formulation

Following the convention, we use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to represent matrices and vectors, respectively. We employ non-bold letters (e.g., x) to denote scalars and uppercase calligraphic symbols (e.g., $\mathcal{X}$)to denote sets. All vectors are in column forms if not clarified.

Suppose there are n Students $\mathcal{S}=\{s_1, s_2, \ldots, s_n\}$, h Groups $\mathcal{G}=\{g_1, g_2, \ldots, g_h\}$, m Exercises $\mathcal{E}=\{e_1, e_2, \ldots, e_m\}$, and t Knowledge concepts $\mathcal{K}=\{k_1, k_2, \ldots, k_t\}$. The l-th group $g_l \in \mathcal{G}$ is consisted of a set of students, i.e., group members with student indexes $\mathcal{K}_l=\{k_{l,1}, k_{l,2}, \ldots, k_{l,|g_l|}\}$, where $s_{k_{l,*}} \in \mathcal{S}$ and $|g_l|$ is the size of the group.

There are two kinds of collected responses data among $\mathcal{S}$, $\mathcal{G}$, and $\mathcal{E}$, namely, group-exercise responses and student-exercise responses. A student will choose some exercises for practice in daily learning, and the student-exercise response logs F are denoted as a set of triplet (s, e, u) where $s \in \mathcal{S}$, $e \in \mathcal{E}$ and u is a binary variable — 1 means that student s has a correct response on exercise e and 0 otherwise. The group-exercise response logs H often come from group evaluation(e.g., class test), and we use a set of triplet (g, e, y) to denote it, where $g \in \mathcal{G}$, $e \in \mathcal{E}$ and y is the correct rate that group g got on exercise e. We should note that there is no intersection between H and F, that is, F does not include the common response data of group students. In addition, we have Q-matrix $\mathbf{Q}=\{Q_{ij}\}_{m \times t}$ as the prior knowledge from education experts to guarantee interpretability($Q_{ij}=1$ if exercise e_i requires skill k_j and 0 otherwise).

Problem Definition *Given group-exercise response logs H, student-exercise response logs F and the Q-matrix* $\mathbf{Q}$*, the goal*

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

of our group-level cognitive diagnosis task is to mine groups' proficiency on knowledge concepts through jointly modeling the tasks of group performance prediction and student performance prediction.

IV. Methodology

There are three key designs in MGCD: 1) multi-task learning that jointly models student-exercise responses and group-exercise responses; 2) group representation learning that aggregates its student representations with attention mechanism with group-exercise responses; and 3) cognitive layer modeling that projects student, exercise, group representations to interpretable factor vectors and learn their interactions with neural networks. This section is organized to elaborate the three parts.

A. Multi-Task Learning

In order to make advantage of the information from student-exercise responses to overcome the sparsity of the group-exercise responses, we propose to jointly model these two kinds of responses in a multi-task manner.

Student Performance Prediction. As the ground-truth of students' knowledge states are inaccessible, we adopt the method used in traditional cognitive diagnosis works, i.e., training the students' ability vectors through predicting their performances [2]. In this work, the goal of student performance prediction task is to predict whether a student can respond correctly to a given exercise.

Group Performance Prediction. Similarly, the goal of group performance prediction task is to predict the correct rate of a group for a given exercise, and then we use the performance of this task to indirectly evaluate the effectiveness of GCD.

The general flow of these two prediction tasks is shown in Figure 2, we project student representation and group representation to obtain student ability and group ability respectively, and then learn interaction function (colored as purple) to predict student-exercise response and group-exercise response. Specifically, the information transfer between these two tasks are achieved by sharing student representations and exercise factors, of which the former is then used to form the group representation (details in Section IV-B).

B. Group Representation Learning

Due to the fact that a group is composed of its students, we naturally obtain group representation from the representations of its containing students. This would allow information learned from responses transferred between the two tasks, and therefore get better diagnostic results.

We associate student $\mathbf{s}_j$ with an embedding $\mathbf{r}_{s_j}$, directly projecting student one-hot vector $\mathbf{x}_{s_j}$ to the latent space.

$$\mathbf{r}_{s_j} = \mathbf{x}_{s_j} \times \mathbf{R}, \tag{1}$$

where $\mathbf{R}$ is a trainable matrix. Next, we aggregate the student embeddings in a group to obtain the group's embedding. Before introducing our method, we first recapitulate some common aggregation strategies.

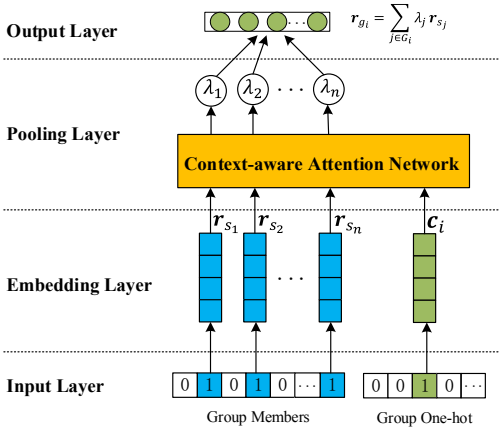


Fig. 3. Illustration of the student embedding aggregation component based on neural attention network.

There are several predefined strategies in neural networks to aggregate embeddings, such as max pooling, average pooling, and min pooling [31]. In general, these aggregation strategies are also known as heuristic strategies, where they first predict the students' proficiency scores on specific knowledge concepts, and then aggregate those predicted scores of each member in a group via the strategies to obtain the group's proficiency. These aggregation strategies can be explained from the cognitive ability level of the group. For example, the max pooling strategy tries to maximize the greatest ability of a group by choosing the highest proficiency score among its members on each knowledge concept.

We argue that these heuristic aggregation strategies are not sufficient to model the relationship between students and groups due to the inflexibility in adjusting the weights of members in the group. It should be noticed that the role of a student in a group is highly related to the student's characteristic (e.g., knowledge state), and the importance might be different according to trait of the group context (e.g., cooperative) [33]. For example, two students with similar abilities in different groups may have different influences due to the diversity of learning contexts. Toward this end, we design an adaptive weighted sum operation which is inspired by the attention mechanism in neural networks [30], [34]. Let $\mathbf{r}_{g_i}$ be the representation for group g_i, we obtain it by:

$$\mathbf{r}_{g_i} = \sum_{j \in \mathcal{G}_i} \lambda_j \mathbf{r}_{s_j}, \tag{2}$$

where λ_j denotes the influence weight of student $\mathbf{s}_j$. To dynamically model the influence weight of students in different contexts, we design a novel context-aware attention mechanism to learn the weights from the historical data of group exercise responses:

$$\begin{aligned} o_j &= \mathbf{h}^T \tanh\left(\mathbf{W}_K \mathbf{r}_{s_j} + \mathbf{W}_Q \mathbf{c}_i\right), \\ \lambda_j &= \operatorname{softmax}(o_j) = \frac{\exp(o_j)}{\sum_{j' \in G_i} \exp(o_{j'})}, \end{aligned} \tag{3}$$

where $\mathbf{c}_i$ is the group-level context vector of g_i and can be obtained by multiplying the one-hot vector of g_i with a trainable matrix $\mathbf{W}_C$, i.e., $\mathbf{c}_i = \mathbf{x}_{g_i} \times \mathbf{W}_C$. $\mathbf{W}_K$ and $\mathbf{W}_Q$ are the key matrix and query matrix of the attention network

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

respectively that convert student embedding and group-level context vector to hidden layers, respectively. We use tanh as the activation function, and then project it to a score o_j with a weight vector $\mathbf{h}$. Lastly, we normalize the scores with a softmax function, which makes the attention network a probabilistic interpretation.

Figure 3 illustrates our design of the student embedding aggregation component. With such a soft attention mechanism, we allow individually modeling the influence of students, where the weights depend on students' latent characteristics and the group's context property, which are learned from the data of group-exercise responses and student-exercise responses (to be discussed in Section IV-C).

C. Cognitive Layer Modeling

The goal of cognitive layer modeling is to obtain explainable student abilities and group abilities, and model the complicated interactions among students, groups and exercises. Details are introduced as below.

Group Ability. After obtaining a group's representation, the next target is to model the cognitive ability of the group, which can characterize the group's traits and affect the group's response to exercises. Specifically, We use a cognitive ability vector $\mathbf{h}_g$ to characterize a group:

$$\mathbf{h}_g = \text{sigmoid}\left(\mathbf{r}_g \times \mathbf{A}\right), \tag{4}$$

where $\mathbf{h}_g \in (0,1)^{1\times K}$ and $\mathbf{A}$ is a trainable matrix.

Student Ability. our proposed framework need to co-train student performance prediction task and group performance prediction task, so we model students' cognitive ability to predict student-exercise responses. Similarly, We use a vector $\mathbf{h}_s$ to represent the student's cognitive ability:

$$\mathbf{h}_s = \text{sigmoid}\left(\mathbf{r}_s \times \mathbf{A}\right), \tag{5}$$

where the matrix $\mathbf{A}$ is shared in MGCD framework.

Exercise Factors. For a group-level cognitive diagnosis system, exercise factors is another important element to be considered [17], which characterize the traits of exercises. In this work, the exercise factor we first consider is exercise-related knowledge concepts $\mathbf{Q}_e$ to ensure the interpretability of our model. In cognitive diagnosis tasks, Q-matrix is given as the prior knowledge from education experts for denoting which knowledge concepts are needed for each exercise, so the $\mathbf{Q}_e$ can be obtained by:

$$\mathbf{Q}_e = \mathbf{x}_e \times \mathbf{Q}, \tag{6}$$

where $\mathbf{Q}_e \in \{0,1\}^{1\times K}$, $\mathbf{x}_e$ is the one-hot vector of exercise *e*. In addition, we also consider other two exercise factors: the knowledge difficulty $\mathbf{h}_{diff}$ and the exercise discrimination $\mathbf{h}_{disc}$, which are widely used in CDMs for more effective diagnosis. $\mathbf{h}_{diff}$ indicates the difficulty of each knowledge concept related to a given exercise, given by:

$$\mathbf{h}_{diff} = \text{sigmoid}\left(\mathbf{x}_e \times \mathbf{B}\right), \tag{7}$$

where $\mathbf{B}$ is a trainable matrix. $\mathbf{h}_{disc}$ refers to the ability to distinguish groups or students with different knowledge proficiency, which can be obtained by:

$$\mathbf{h}_{disc} = \text{sigmoid}\left(\mathbf{x}_e \times \mathbf{D}\right), \tag{8}$$

where $\mathbf{D}$ is a trainable matrix.

Interaction Function. Many choices of interaction function explored in traditional cognitive diagnosis models can be applied here, such as inner product used in MF methods and item response functions used in IRT and MIRT methods. In this work, We opt for the Neural Cognitive Diagnosis Model (NeuralCDM) [12] for two reasons: 1) NeuralCDM is more flexible in designing multiple nonlinear layers to learn complex interaction functions, which allows us to seamlessly incorporate student-exercise responses modeling into the group-exercise model; 2) NeuralCDM is a neural network architecture, being suitable to perform end-to-end learning on both embeddings (that represent students, exercises, and groups) and interaction functions (that predict student-exercise and group-exercise responses). The first layer of the interaction layers is formulated as:

$$\begin{cases} \mathbf{x}_g = \mathbf{Q}_e \circ \left(\mathbf{h}_g - \mathbf{h}_{diff}\right) \times \mathbf{h}_{disc} \\ \mathbf{x}_s = \mathbf{Q}_e \circ \left(\mathbf{h}_s - \mathbf{h}_{diff}\right) \times \mathbf{h}_{disc} \end{cases}, \tag{9}$$

where $\circ$ is element-wise product. Then shared hidden layers are used to capture the nonlinear and higher-order correlations among students, groups, and exercises.

$$\begin{cases} \mathbf{z}_1 = \phi\left(\mathbf{W}_1\mathbf{x} + b_1\right) \\ \mathbf{z}_2 = \phi\left(\mathbf{W}_2\mathbf{z}_1 + b_2\right) \\ \cdots \\ \mathbf{z}_h = \phi\left(\mathbf{W}_h\mathbf{z}_{h-1} + b_h\right) \end{cases}, \tag{10}$$

where ϕ is the activation function. Finally, the output of the last hidden layer $\mathbf{z}_h$ is transformed to a prediction score via:

$$\begin{cases} \hat{y}_{lj} = \phi\left(\mathbf{W}_{h+1}\mathbf{z}_h + b_{h+1}\right), \text{if } \mathbf{x} = \mathbf{x}_{\text{g}} \\ \hat{u}_{ij} = \phi\left(\mathbf{W}_{h+1}\mathbf{z}_h + b_{h+1}\right), \text{if } \mathbf{x} = \mathbf{x}_{\text{s}} \end{cases}, \tag{11}$$

where $\hat{y}_{lj}$ and $\hat{u}_{ij}$ represent the prediction for a group-exercise response pair (g_l, e_j) and a student-exercise response pair (s_i, e_j), respectively. It is worth mentioning that we have purposefully designed the prediction of the two tasks share the same hidden layers and matrix $\mathbf{A}$. This is because that the group embedding is aggregated from student embeddings, which makes them in the same semantic space by nature.

Here, in order to ensure that the diagnostic results of student ability vector $\mathbf{h}_s$ and group ability vector $\mathbf{h}_g$ are reasonable, we utilize the monotonicity assumption to guarantee the interpretability of the cognitive layer, which is used in some IRT and MIRT models [10], defined as follows:

Monotonicity Assumption. *When the proficiency of any knowledge concept increases, the probability of correct response to the exercise also increases.*

Specifically for GCD, the assumption means that the group's correct rate for an exercise increases as any dimension of the cognitive ability vector increases, which places limits on the mathematical forms considered for the interaction function. For example, in traditional methods, the logistic is widely used as the interaction function, which can easily be proved to satisfy the monotonicity assumption due to the inherent monotonic characteristics.

In this work, we simply restrict each matrix of the interaction function to be positive to satisfy the monotonicity assumption (details in Section V-B). Moreover, during model training,

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

the Q-matrix is used to control the change of each dimension of the ability vector. Finally, the obtained ability vector will be interpretable, which can represent the proficiency on each knowledge concept.

Joint Training. We use the cross-entropy loss function for this student performance prediction task as follows:

$$L_{student} = -\sum_{i=1}^{n} \sum_{j \in \mathcal{F}_i} \left(\hat{u}_{ij} \log u_{ij} + (1 - \hat{u}_{ij}) \log (1 - u_{ij})\right), \tag{12}$$

where $\mathcal{F}_i$ denotes all exercises responded by student s_i. The group performance prediction task aims to predict groups' correct rate for a given exercise, so we choose the mean square error loss function for this prediction task as follows:

$$L_{group} = \frac{1}{T} \sum_{i=1}^{h} \sum_{j \in \mathcal{H}_i} \left(\hat{y}_{ij} - y_{ij}\right)^2, \tag{13}$$

where T represents the total number of student-exercise responses in the training data and $\mathcal{H}_i$ denotes all exercises responded by group g_i. The final objective function for the joint model is the sum of the two tasks' objective functions:

$$L = L_{student} + L_{group}. \tag{14}$$

After training, the value of h_g is what we get as diagnosis result, which denotes the group's proficiency on each knowledge concept.

Generality of Cognitive Layer.

As mentioned before, many traditional methods can be applied to modeling the interaction among group ability, student ability, and exercise factors. Specifically, in order to keep MGCD as a general framework, the cognitive layer we deliberately designed can cover many traditional interaction functions. The extendibility over representative models are illustrated as follows:

IRT. Take the typical formation: $y = \sigma\left((h_g - h_{diff}) \times h_{disc}\right)$ as the example of IRT, where h_g, h_{diff} and h_{disc} denote group ability, exercise difficult and discrimination respectively. To extend from IRT, in the cognitive layer, we project group embedding r_g and exercise one-hot x_e to unidimensional h_g and h_{diff} respectively, and set $Q_e \equiv 1$ and h_{disc} as additional trainable parameters. As for the shared hidden layers, we directly replace them with the sigmoid function.

MIRT. MIRT is a multidimensional extension of IRT [35]: $y = \sigma\left(Q_e \cdot (h_g - h_{diff})\right)$, where Q_e is the one-hot index vector of related concepts for exercise e. To extend from MIRT, we just compute the sum of elements in x_g (Eq. (9)) and choose sigmoid as the activation function to get the model prediction result without additional shared hidden layers.

MF. According to [12], MF can be treated as a special case of the NeuralCDM model where $h_{diff} \equiv 0$ and $h_{disc} \equiv 1$. In particular, we can implement MF by factoring score matrix to get h_g and Q_e, i.e. $y = Q_e \cdot h_g$. Therefore, extending from MF is straightforward, all the shared hidden layers need to sum up the values of each entry in x_g.

It is noteworthy that interaction functions are shared in our multi-task learning framework. Therefore, these traditional methods introduced above are also suitable for modeling student-exercise response.

TABLE I
DATASET SUMMARY.

Dataset	Math	NIPS_Edu	ASSIST2012
# Students	12,853	4,430	1,495
# Groups	1,716	634	97
# Exercises	7,997	18,701	13,342
# Knowledge concepts	1,566	356	162
AVG. group size	7.49	6.98	15.41
AVG. #responses for a student	45.58	81.25	66.24
AVG. #responses for a group	37.12	10.24	9.02

V. EXPERIMENTS

In this section, we conduct extensive experiments on three real-world datasets aiming to answer following research questions and validate our technical contributions.

RQ1 How is the generality of our proposed framework? Can it be applied to different cognitive diagnosis models and obtain better performance?

RQ2 How is the effectiveness of our designed attention network? Can it perform better than other predefined aggregation strategies?

RQ3 Can multi-task learning framework improve the performance of group-level cognitive diagnosis? Can it alleviate the sparsity of group-exercise responses?

RQ4 How about the interpretation of MGCD on mining group knowledge states for group-level cognitive diagnosis?

A. Dataset Description.

We experimented with three real-world datasets, i.e., Math, NIPS_Edu, and ASSIST2012. the Math dataset is collected from a widely-used online learning system[1], which contains the response logs of high school students to mathematical exercises. The NIPS_Edu dataset is from a diagnostic questions competition: The NeurIPS 2020 Education Challenge[2] [36], which provides students' response logs to mathematics questions in two school years (2018-2020). The ASSIST2012 dataset is provided by ASSISTment online tutoring platform[3] and is widely used for cognitive diagnosis tasks.

All these datasets contain group labels, and students from the same group belong to the same class. Each dataset contains two records that do not overlap: student-exercise responses and group-exercise responses. Specifically, for each group-exercise response, we calculate the correct rate of this group of students on the exercise as the response result. We filter out groups with less than 5 group-exercise response logs in each dataset to guarantee that each group has enough response data for diagnosis. The statistics of datasets are shown in Table 1.

B. Experimental Setup

1) Evaluation metric: For performance evaluation, each group-exercise response dataset is randomly split into two parts: 80% as the training set and 20% as the test set. In this

[1]https://www.zhixue.com

[2]https://competitions.codalab.org/competitions/25449

[3]https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXPERIMENTAL RESULTS ON VERIFYING THE GENERALITY OF MGCD.

Model	Math		NIPS_Edu		ASSIST2012	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
IRT	0.2400	0.1999	0.2977	0.2582	0.3055	0.2730
IRT-AVG	0.2347	0.1898	0.2683	0.2401	0.2805	0.2637
MGCD-IRT	**0.2215**	**0.1744**	**0.2430**	**0.2211**	**0.2654**	**0.2305**
MIRT	0.2175	0.1645	0.2693	0.2202	0.2695	0.2564
MIRT-AVG	0.1912	0.1501	0.2349	0.2197	0.2731	0.2637
MGCD-MIRT	**0.1866**	**0.1442**	**0.2311**	**0.2011**	**0.2238**	**0.2231**
PMF	0.2835	0.2276	0.2963	0.2524	0.2945	0.2681
PMF-AVG	0.2843	0.2216	0.2835	0.2431	0.2967	0.2637
MGCD-MF	**0.2411**	**0.2085**	**0.2564**	**0.2241**	**0.2563**	**0.2102**
NeuralCDM	0.2059	0.1558	0.2675	0.2102	0.2610	0.1945
NeuralCDM-AVG	0.1823	0.1421	0.2427	0.1984	0.2517	0.2113
MGCD	**0.1787**	**0.1293**	**0.2236**	**0.1761**	**0.2106**	**0.1561**

work, the effectiveness of group-level cognitive diagnosis is indirectly validated by the group performance prediction task. Therefore, considering the task is to predict the correct rate of a group for a given exercise, we adopt two metrics for regression task to evaluate the performance, i.e., root mean square error (RMSE), mean absolute error (MAE).

2) Framework Setting: The dimensions of the hidden layers (Eq. (10)) are 128, 64, 1 respectively, and the activation function of the last layer is sigmoid and that of the other layers is tanh. Indeed, we empirically set the number of the shared hidden layers to 3, which has achieved good results in the experiments. Too many layers may increase noise irrelevant to the group-level cognitive diagnosis task, and too few layers may not be conducive to modeling the complex interaction among students, groups, and exercises. Moreover, we use the tower structure for hidden layers and leave the further tuning on the structure as future work. In order to satisfy the monotonicity assumption to regularize the model, we restrict the parameters of each matrix (in Equation 10) to be positive. A simple implementation scheme is to map the parameters of each matrix to a non-negative number through the relu function during forward propagation, and then perform the corresponding matrix multiplication operation.

3) Training Details: All models are implemented by Tensorflow using Python, and all experiments are run on a Linux server with four 2.0GHz Intel Xeon E5-2620 CPUs and a Tesla K20m GPU.

To set up training process, we initialize all network parameters with Xavier initialization strategy [37]. Each parameter is sampled from $U\left(-\sqrt{2/\left(n_{in}+n_{\text{out}}\right)}, \sqrt{2/\left(n_{\text{in}}+n_{\text{out}}\right)}\right)$, where n_{in} and n_{out} denote the numbers of neurons feeding in and feeding out, respectively. We used the Adam optimizer for all gradient-based methods, where the mini-batch size and learning rate were searched in [32, 64, 128] and [0.001, 0.005, 0.01, 0.05, 0.1], respectively. All models are evaluated with 5-fold cross validation.

C. Baseline Approaches.

To show the effectiveness of our method, we compared it with the following models.

- **IRT** [8]: IRT is a cognitive diagnosis method modelling students' latent trait and the parameters of exercises like difficulty and discrimination.
- **MIRT** [10], [38]: MIRT is an extension of the unidimensional IRT models that seek to explain an item (exercise) response according to a student's standing across multiple latent dimensions.
- **PMF** [39]: probabilistic matrix factorization is a latent factor model projecting students and exercises into a low-dimensional space.
- **NeuralCDM** [12]: NeuralCDM is a novel neural cognitive diagnosis model, which can leverage multi neural layers for modeling complex interactions between students and exercises.

Among all the above-mentioned baselines, only the NeuralCDM is interpretable for the diagnostic result. As for all the other models, there are no clear correspondence between their latent features and knowledge concepts. In order to facilitate subsequent interpretability experiment, inspired by [40], [41], we extend MIRT and PMF by integrating Q-matrix to improve the explanatory power. Moreover, all the above-mentioned baselines are single-task models, which regard each group as a virtual unit, and then mine the group's cognitive ability by group-exercise response data.

D. Performance Comparison (RQ1)

In order to verify the generality of MGCD framework, we use different interaction functions based on our framework, including IRT, MIRT, PMF, and NeuralCDM, and then compare the performance with the corresponding baseline respectively. In addition, to avoid data unfairness between single-task learning and multi-task learning, we further choose the second solution (introduced in section II-B) as the control experiment. Specifically, we first leverage the baseline methods to diagnose the knowledge state of individual students through all the response data and then regard the group ability as the average ability of the students, and the effect is measured by the group performance prediction task. To distinguish from those baselines, we add the suffix '-AVG' to their names.

As shown in Table 2, we can see that co-training the group-exercise and student-exercise responses perform better

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

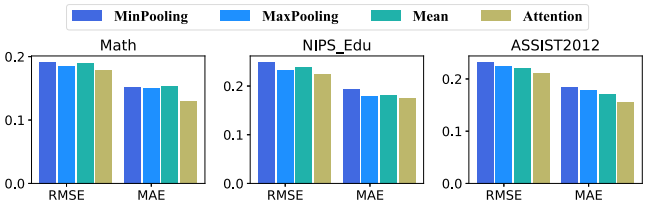


Fig. 4. The performance comparison of different fusing strategies.

in different interaction functions, which verifies the generality of MGCD. Moreover, we can observe that even if the student-exercise response data is additionally used, the performance of the second solution is still suboptimal, which could be caused by the uneven distribution of ability and response logs in the group. This further proves that multi-task learning can make better use of student-exercise response data. Besides, the performance of neural network-based interaction is superior to other methods, demonstrating the superiority of neural networks, especially their great ability in modeling the high-order interactions among students, groups, and exercises.

E. Effect of Attention (RQ2)

In order to investigate the effectiveness of the attention network, we compare it with other predefined aggregation strategies, including min pooling, max pooling, and average pooling. As shown in Figure 4, compared with other aggregation strategies, the attention network achieves a relative performance improvement on all datasets with respect to both metrics. Experimental results show that the attention aggregation module has strong representation power in complex learning contexts, which provides evidence on the effectiveness of the attention network.

In addition, in order to prove that the context-aware attention network can model the influence weights of students in different learning contexts, we randomly select two groups (group a and group b) for visualization. We first calculate the average of the ability vectors of a and b in all dimensions respectively. Then we reduce the representation vector dimension of each student in the two groups to two-dimension space using the T-SNE method [42] and draw a scatter figure. From Figure 5 (a) we can see that although the ability distribution of students in a group is extensive, the similarity between students in the same group is still higher than that among different groups, which indicates the commonness among group members. We select two students ($a1$ and $b1$) in Figure 5 (a) with similar abilities who belong to a and b respectively. We then randomly select another 4 students from a and normalize the attention weights for visualization, and the same method for group b. As shown in Figure 5 (b), we can observe that different students (e.g., $a1$ and $a2$) in the same group have different weights. Although the abilities of $a1$ and $b1$ are similar, since the overall proficiency of a is lower than b, $a1$ may have a higher influence in a, which causes a different weight in Figure 5 (b).

F. Effect of Multi-task Learning (RQ3)

To investigate whether multi-task learning improves the model performance and alleviates the sparsity of group-exercise responses, we vary the proportion of the group-exercise responses training set and compare the performance of our model with baselines. The experimental results are shown in Figure 6. We have the following observations: 1) The multi-task learning model we proposed can perform better than these single-task learning models, which verifies that the insufficient response data between group and exercise limits the model performance. More importantly, with the increasing of the sparsity of the training data (training data ratio declines from 80% to 20%), the superiority of our method becomes more and more significant. 2) Compared with the Math dataset, as shown in Table 1, the group-exercise responses of the other two datasets are sparser, and our model has more obvious performance improvements.

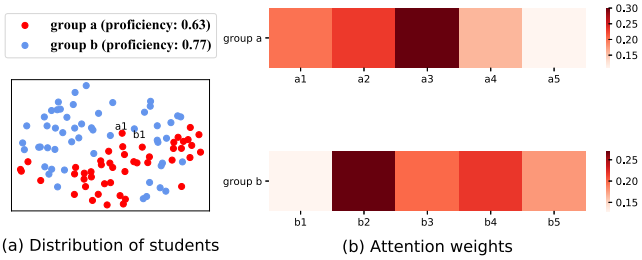


Fig. 5. Visualization for attention weights.

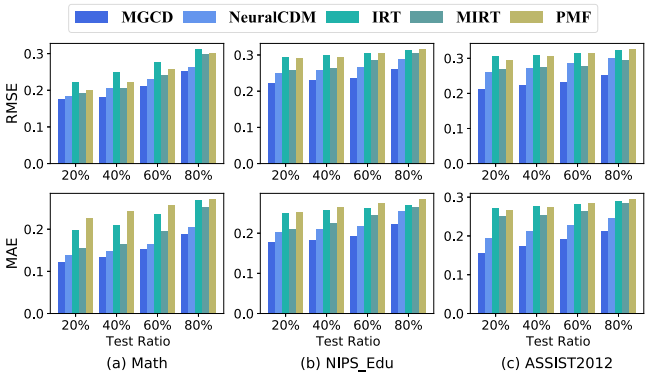


Fig. 6. Experimental results with different test ratios.

G. Interpretation of the Diagnosis (RQ4)

To assess the interpretability of MGCD framework (i.e., whether the diagnostic result is reasonable), we further conduct the following experiments.

1) Ranking Performance Evaluation: Intuitively, if group a has a better mastery than group b on knowledge concept k, then a is more likely to get better performance to answer exercises related to k. We adopt Degree of Agreement (DOA) [40] metric to evaluate this ranking performance.

Particularly, for a specific knowledge k, the DOA result on k is defined as:

$$DOA(k) = \frac{1}{Z} \sum_{j=1}^{m} I_{jk} \sum_{a=1}^{h} \sum_{b=1}^{h} \frac{\delta\left(F_{ak}, F_{bk}\right) \wedge \delta\left(y_{aj}, y_{bj}\right)}{\delta\left(F_{ak}, F_{bk}\right)} \tag{15}$$

where $Z = \sum_{a=1}^{h} \sum_{b=1}^{h} \delta\left(F_{ak}, F_{bk}\right)$. F_{ak} is knowledge proficiency of group a on knowledge concept k. y_{aj} (denoted

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

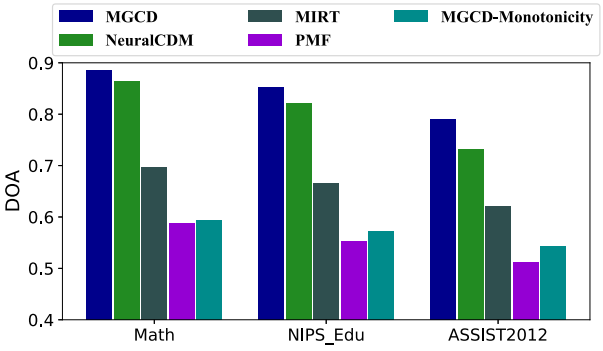


Fig. 7. DOA results of models.

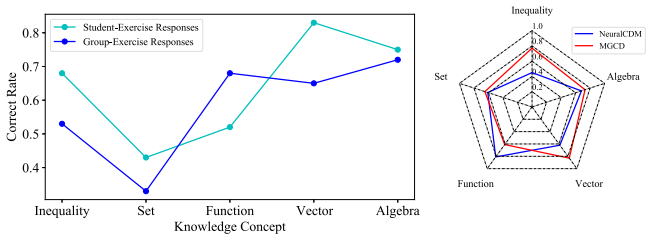
in Eq. (13)) is group a's response on exercise j. $\delta(x, y)$ is an indicator function, where $\delta(x, y) = 1$ if $x > y$. I_{jk} is an another indicator function, where $I_{jk} = 1$ if exercise e contains knowledge concept k. The DOA value ranges from 0 to 1 and the larger the better. Furthermore, we average $DOA(k)$ on all knowledge concepts for measuring the overall quality of diagnostic result.

As mentioned before, we have integrated the Q-matrix to expand MIRT and PMF to ensure the interpretability of the diagnosis results (Section V-C). Besides, to prove the importance of monotonicity assumption, we design a simplified model (denoted as MGCD-Monotonocity), in which monotonicity assumption is removed by eliminating the positive restriction on the shared hidden layers.

As shown in Figure 7, we can obtain that MGCD has better performance than other baseline methods in the metric of DOA, which further demonstrates that the groups' knowledge state obtained by our method is reasonable. When the monotonicity assumption is removed, the DOA result of MGCD-Monotonocity drops significantly. To satisfy the interpretability of MF, the latent trait of exercise here only considers the exercise-related knowledge concept $\mathbf{Q}_e$ (Eq. (6)), and the interaction between group ability and exercise feature is modeled through the inner product form, which is not enough to mine the complex high-order group-exercise relationship, thus obtaining a lower DOA result.

2) Case Study: We present an example of the diagnosis results of a group of students on each knowledge concept in Dataset Math using NeuralCDM and MGCD. Here, we evaluate the interpretability based on whether the diagnostic proficiency of groups is reasonable with the additional student-exercise response data. In order to show the knowledge state of groups more intuitively, we visualize the diagnosis results, which are shown in Figure 8.

The lines in Figure 8 (a) show the correct rate of the exercises related to each knowledge concept and we consider the group-exercise responses and the student-exercise responses respectively. Figure 8 (b) shows the diagnostic result and each point on the radar diagram represents the mastery level of the certain knowledge concept. We can observe that both methods can obtain interpretatively meaningful diagnosis results. However, we can notice that there are some differences



(a) Distribution of Correct rate (b) Radar Digram

Fig. 8. Diagnosis results for a group of students

in the results of these two methods, e.g., the proficiency of 'Inequality' diagnosed by MGCD is higher than that obtained by NeuralCDM. From the line chart, we can observe that the correct rate of 'Inequality' obtained by the student-exercise responses is higher than that of the group-exercise responses. Therefore, the single-task model is likely to get suboptimal diagnostic results due to the lack of sufficient response data. From the results, we can see that owing to the effective use of student-exercise response data, which provides additional information about students' abilities in a group, MGCD is able to provide a better interpretable insight on diagnosing group knowledge states for GCD.

VI. Conclusion and Future Work

In this paper, we focused on the problem of cognitive diagnosis for a group of students from the perspective of multi-task learning. Specifically, to mitigate the sparsity of group-exercise response data, we jointly modeled student-exercise response and group-exercise response to share the information of two tasks. Then, we designed an attention network to aggregate the student embeddings in a group to obtain the group's representation. Moreover, in order to mine the latent characteristics of students, exercises, and groups, we modeled student ability, group ability, and exercise factors respectively, and then we leveraged neural networks to learn the complicated interactions among them. Extensive experimental results on three real-world datasets clearly demonstrated the effectiveness, generality and interpretability of MGCD framework. We hope this work could lead to further studies.

This paper provides a novel solution for group-level cognitive diagnosis. The research area is still in its infancy, and we anticipate that more techniques will be developed in the future. Specifically, we plan to find more reasonable metrics on group assessment. Besides, if an exercise can only be considered as a group's response when all members of the group have practiced, it will limit the practicality of the diagnostic model, so we will try to design an efficient sampling algorithm in the future. Meanwhile, we'd like to apply our method to other fields, such as team ability assessment in the game field [5].

Acknowledgement. This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 61922073, U20A20229 and 62106244), the Foundation of State Key Laboratory of Cognitive Intelligence (No. iED2020-M004) and the Iflytek joint research program.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] J. De La Torre, "The generalized dina model framework," *Psychometrika*, vol. 76, no. 2, pp. 179–199, 2011.

[2] R. Wu, Q. Liu, Y. Liu, E. Chen, Y. Su, Z. Chen, and G. Hu, "Cognitive modelling for predicting examinee performance," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[3] E. Hammar Chiriac, "Group work as an incentive for learning–students' experiences of group work," *Frontiers in psychology*, vol. 5, p. 558, 2014.

[4] S. P. Reise, R. R. Meijer, A. T. Ainsworth, L. S. Morales, and R. D. Hays, "Application of group-level item response models in the evaluation of consumer reports about health plan quality," *Multivariate behavioral research*, vol. 41, no. 1, pp. 85–102, 2006.

[5] Y. Gu, Q. Liu, K. Zhang, Z. Huang, R. Wu, and J. Tao, "Neuralac: Learning cooperation and competition effects for match outcome prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 5, 2021, pp. 4072–4080.

[6] L. T. DeCarlo, "On the analysis of fraction subtraction data: The dina model, classification, latent class sizes, and the q-matrix," *Applied Psychological Measurement*, vol. 35, no. 1, pp. 8–26, 2011.

[7] J. Douglas and A. Douglas, "Evaluating teaching quality," *Quality in Higher Education*, vol. 12, no. 1, pp. 3–13, 2006.

[8] S. E. Embretson and S. P. Reise, *Item response theory*. Psychology Press, 2013.

[9] T. A. Ackerman, M. J. Gierl, and C. M. Walker, "Using multidimensional item response theory to evaluate educational and psychological tests," *Educational Measurement: Issues and Practice*, vol. 22, no. 3, pp. 37–51, 2003.

[10] M. D. Reckase, "Multidimensional item response theory models," in *Multidimensional item response theory*. Springer, 2009, pp. 79–112.

[11] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[12] F. Wang, Q. Liu, E. Chen, Z. Huang, Y. Chen, Y. Yin, Z. Huang, and S. Wang, "Neural cognitive diagnosis for intelligent education systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6153–6161.

[13] R. Agrawal, B. Golshan, and E. Terzi, "Grouping students in educational settings," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1017–1026.

[14] Y. Liu, Q. Liu, R. Wu, E. Chen, Y. Su, Z. Chen, and G. Hu, "Collaborative learning team formation: a cognitive modeling perspective," in *International Conference on Database Systems for Advanced Applications*. Springer, 2016, pp. 383–400.

[15] J. M. Gonzalez and J. L. Eltinge, "Multiple matrix sampling: A review," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*. American Statistical Association, 2007, pp. 3069–3075.

[16] D. J. Veldman and J. P. Sanford, "The influence of class ability level on student achievement and classroom behavior," *American Educational Research Journal*, vol. 21, no. 3, pp. 629–644, 1984.

[17] L. V. DiBello, L. A. Roussos, and W. Stout, "31a review of cognitively diagnostic assessment and a summary of psychometric models," *Handbook of statistics*, vol. 26, pp. 979–1030, 2006.

[18] X. Li, W.-C. Wang, and Q. Xie, "Cognitive diagnostic models for rater effects," *Frontiers in Psychology*, vol. 11, 2020.

[19] B. W. Junker and K. Sijtsma, "Cognitive assessment models with few assumptions, and connections with nonparametric item response theory," *Applied Psychological Measurement*, vol. 25, no. 3, pp. 258–272, 2001.

[20] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," *KDD cup*, 2010.

[21] N. Thai-Nghe and L. Schmidt-Thieme, "Multi-relational factorization models for student modeling in intelligent tutoring systems," in *2015 Seventh international conference on knowledge and systems engineering (KSE)*. IEEE, 2015, pp. 61–66.

[22] J. Mazzeo and M. von Davier, "Linking scales in international large-scale assessments," *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, pp. 229–258, 2014.

[23] S. Sellar and B. Lingard, "International large-scale assessments, affective worlds and policy impacts in education," *International Journal of Qualitative Studies in Education*, vol. 31, no. 5, pp. 367–381, 2018.

[24] R. J. Mislevy, "Item response models for grouped data," *Journal of Educational Statistics*, vol. 8, no. 4, pp. 271–288, 1983.

[25] M. Birenbaum, C. Tatsuoka, and T. Yamada, "Diagnostic assessment in timss-r: Between-countries and within-country comparisons of eighth graders' mathematics performance," *Studies in Educational Evaluation*, vol. 30, no. 2, pp. 151–173, 2004.

[26] R. Tate and M. Heidorn, "School-level irt scaling of writing assessment data," *Applied Measurement in Education*, vol. 11, no. 4, pp. 371–383, 1998.

[27] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[28] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 109–117.

[29] Y. Fang, Z. Ma, Z. Zhang, X.-Y. Zhang, X. Bai *et al.*, "Dynamic multi-task learning with convolutional neural network." in *IJCAI*, 2017, pp. 1668–1674.

[30] D. Cao, X. He, L. Miao, Y. An, C. Yang, and R. Hong, "Attentive group recommendation," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 645–654.

[31] L. Chen, Y. Liu, X. He, L. Gao, and Z. Zheng, "Matching user with item set: Collaborative bundle recommendation with deep attention network." in *IJCAI*, 2019, pp. 2095–2101.

[32] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *2014 Canadian Conference on Computer and Robot Vision*. IEEE, 2014, pp. 98–103.

[33] E. Rosenqvist, "Two functions of peer influence on upper-secondary education application behavior," *Sociology of Education*, vol. 91, no. 1, pp. 72–89, 2018.

[34] H. Yin, Q. Wang, K. Zheng, Z. Li, and X. Zhou, "Overcoming data sparsity in group recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[35] R. J. Adams, M. Wilson, and W.-c. Wang, "The multidimensional random coefficients multinomial logit model," *Applied psychological measurement*, vol. 21, no. 1, pp. 1–23, 1997.

[36] Z. Wang, A. Lamb, E. Saveliev, P. Cameron, Y. Zaykov, J. M. Hernández-Lobato, R. E. Turner, R. G. Baraniuk, C. Barton, S. P. Jones *et al.*, "Diagnostic questions: The neurips 2020 education challenge," *arXiv preprint arXiv:2007.12061*, 2020.

[37] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[38] D. L. Knol and M. P. Berger, "Empirical comparison between factor analysis and multidimensional item response models," *Multivariate behavioral research*, vol. 26, no. 3, pp. 457–477, 1991.

[39] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257–1264.

[40] Y. Chen, Q. Liu, Z. Huang, L. Wu, E. Chen, R. Wu, Y. Su, and G. Hu, "Tracking knowledge proficiency of students with educational priors," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 989–998.

[41] Z. Huang, Q. Liu, Y. Chen, L. Wu, K. Xiao, E. Chen, H. Ma, and G. Hu, "Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 2, pp. 1–33, 2020.

[42] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 12,2022 at 08:21:20 UTC from IEEE Xplore. Restrictions apply.