Front. Comput. Sci., 2025, 0(0): 1–17
https://doi.org/10.1007/sxxxxx-yyy-zzzz-1

RESEARCH ARTICLE

# Breaking Student-Concept Sparsity Barrier for Cognitive Diagnosis

**Pengyang Shao[1], Kun Zhang(✉)[1], Chen Gao(✉)[2], Lei Chen[2], Miaomiao Cai[1], Le Wu[1], Yong Li[2], Meng Wang[1]**

1 Key Laboratory of Knowledge Engineering with Big Data, Hefei University of Technology, Anhui 230601, China

2 Department of Electronic Engineering, Tsinghua University, Beijing 100084, China



**Abstract** Educational Cognitive Diagnosis (CD) aims to provide students' mastery levels on different concepts. One common observation is that students often conduct many exercises but engage with a small subset of concepts, leading to a sparsity barrier. Current CD models mostly adopt mastery levels on all concepts as student modeling, overlooking the sparsity barrier. If a student does not interact with all concepts, we can not ensure that each dimension of mastery levels on concepts can be well-trained. In this paper, we propose a novel *E*nhancing *S*tudent *R*epresentations in *C*ognitive *D*iagnosis (*ESR-CD*), which combines application abilities and comprehension degrees for mastery levels on concepts. To model application ability, we propose a sparsity-based mask module that solely depends on the dense student-concept entries. Simultaneously, to further enhance comprehension degrees, we propose two layers: a matrix factorization layer and a relation refinement layer. Extensive experiments on two real-world datasets demonstrate the effectiveness of ESR-CD.

**Keywords** Cognitive Diagnosis, Student Modeling, Educational Data Mining

## 1 Introduction

Educational data mining has experienced rapid developments due to easy-to-collect data in online education platforms and superior performance, e.g., cognitive diagnosis [1–4], knowledge tracing [5,6]. Cognitive Diagnosis (CD) has received much attention due to its practical significance and relatively high performance. As shown in Figure 1, CD takes students' response logs on exercises and relations between exercises and concepts (exercise-concept relations) as input, and outputs students' mastery levels on all concepts [7, 8].

Received month dd, yyyy; accepted month dd, yyyy

E-mail: zhang1028kun@gmail.com, chgao96@gmail.com

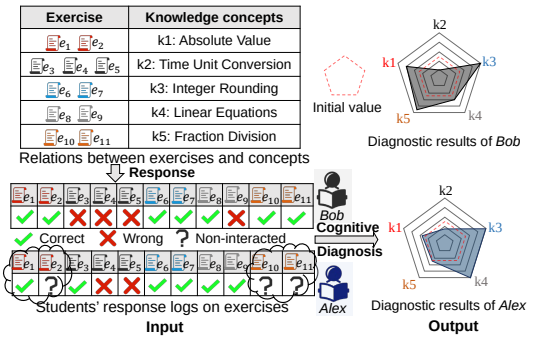



**Fig. 1** A tiny example of cognitive diagnosis.

Recently, neural network-based CD models have achieved impressive performance [7, 9, 10]. These models output students' mastery levels on all concepts. Each dimension of mastery levels is independent of each other, leading to the sparsity barrier between students and concepts (student-concept sparsity barrier). Taking Figure 1 as an example, Bob provides sufficient interactions with all five concepts while Alex does not provide sufficient interactions with concept $k1\&k5$. These CD models can provide accurate diagnostic results for Bob, but they cannot provide satisfactory results for Alex on concepts $k1\&k5$. To further analyze the existence and effects of sparsity barriers, we conduct experiments, and the experimental results are presented in Section 3.3. From the results, we observe sparsity barriers on real-world datasets (i.e., most students conduct many exercises involved with a limited number of concepts), and find that existing neural network-based CD models are heavily constrained by the sparsity barrier.

One promising direction is to use expert-annotated relationships among different concepts [11,12]. However, this requires additional costs, and there are mistakes with these annotations, e.g., the presence of cyclic concept dependencies in Junyi [11, 12]. Hence, relying on concept relationships to enhance the quality of mastery levels is not a satisfactory option. Further, previous studies have already identified incomplete annotations in exercise-concept relations [8, 13], which also leads to difficulties in improving the quality of mastery levels. We consider: *how to improve quality of mastery levels in CD with the presence of the sparsity barrier?*

To this end, we propose *E*nhancing *S*tudent *R*epresentations in *C*ognitive *D*iagnosis (*ESR-CD*), an accurate and robust CD model. Our intuition is from a conclusion from *Kay et al.* [14]: a correct answer require both knowledge and foundational abilities. We take an exercise example to show why foundational abilities are necessary: *"A road needs to be repaired with daily progress of 48 meters, and it will take 25 days. If we want to finish the task 5 days* ***earlier****, how many meters should be repaired each day?"* To correctly answer this exercise, a student should not only grasp the arithmetic concept, but also notice the essential word "earlier". Noticing the essential word can be seen as a part of reflection on this student's foundational abilities. To this end, we decompose the overall student modeling into two parts in this paper: 1) *comprehension degree* denotes how much level students master the concepts. 2) *application ability* denotes the ability to apply learned knowledge to real exercises. In the presence of the student-concept barrier, certain dimensions of comprehension degrees are not adequately trained. However, for application ability, sufficient interactions with a subset of knowledge concepts are enough. A student's application ability maintains consistency across interactions with all concepts, therefore, we can assess a student's application ability by analyzing her performance on a small subset of concepts. Specifically, to obtain application ability, we propose a sparsity-based mask module that solely depends on the dense student-concept entries. To enhance com-

prehension degrees, we propose two layers: a matrix factorization layer and an exercise-concept relation refinement layer. Extensive experiments on two real-world datasets have demonstrated the effectiveness and robustness of ESR-CD. In summary, our contributions are as follows:

- We find that the student-concept sparsity barrier limits CD developments, making it impossible to assess students' mastery levels on concepts they have not encountered. This sparsity phenomenon is widely prevalent in real-world datasets.
- To break the barrier, we argue that student modeling should consider both application ability and comprehension degrees on concepts, and propose ESR-CD, an accurate and robust CD model.
- Extensive experiments on two datasets demonstrate the effectiveness of ESR-CD. For random split, ESR-CD achieves AUC improvements of average 0.5%. For concept weak coverage split, compared to KaNCD, ESR-CD has AUC improvements of over 1.5% on the ASSIST dataset and over 6% on the MOOC-Radar dataset.

## 2 Related Work

### 2.1 Cognitive Diagnosis

CD takes students' response logs on exercises and exercise-concept relationships as input and outputs students' mastery levels on all concepts [15–19]. Classic methods include item response theory [15], meta-knowledge dictionary learning [20], and the Deterministic Inputs, Noisy And gate model [21]. Recently, researchers have focused on leveraging neural networks in CD [8, 22, 23]. Neural Cognitive Diagnosis Model (NCDM) first uses high-dimensional representations to model students' mastery level on each knowledge concept, then proposes diagnostic functions for parameter incorporation, and finally identifies students' non-mastery on their unanswered exercises [7]. The NCDM framework not only offers student performance predictions but also provides students' mastery levels on concepts. Relation map driven Cognitive Diagnosis (RCD) integrates different pre-defined graphs among students, items, and concepts to model connections [11]. For long-term development, researchers propose beginner-friendly frameworks [24], and emphasize aspects such as efficiency [25], interpretability [26, 27], and fairness [28–31].

### 2.2 Alleviating Sparsity in User Modeling

Data sparsity is a prevalent challenge in user modeling, e.g., recommender systems [32–35]. It occurs when there is limited interaction data, resulting in a scarcity of ratings or records. To address the issue, researchers have proposed effective methods including side information (e.g., social neighbors [32, 36], review text [37], and multimedia information [38]), or mining high-order user-item bipartite graph structure information [39].

**Our Distinction:** The student-concept sparsity barrier in CD and the classical definition differ in granularity. Classical user modeling tasks (e.g., recommender systems) do not pursue the interpretability of each dimension in the user representation. If a user interacts with any items, regardless of the categories, we consider this user's data is not sparse. In contrast, CD emphasizes students' mastery levels on all concepts. If student $s$ does not provide interactions with concept $k$, there is a sparsity barrier between student $s$ and concept $k$.

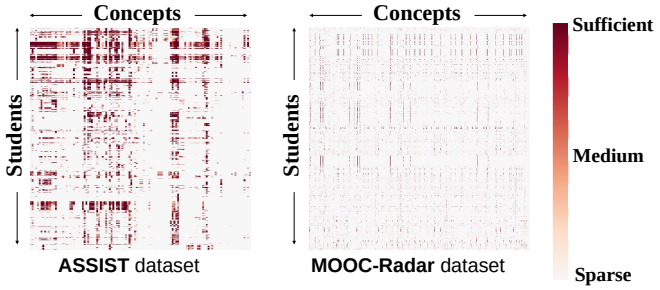
## 3 Preliminary

### 3.1 Task Overview

Suppose that there are student entities $S(|S| = M)$, exercise entities $E(|E| = N)$, and knowledge con-

| Model | $\eta$ (%) | ACC ↑ | AUC ↑ | DOA ↑ |
|---|---|---|---|---|
| NCDM | 0 | 0.7158 | 0.7687 | 0.6514 |
| | 20 | 0.7066 | 0.7480 | 0.6423 |
| | 40 | 0.7009 | 0.7389 | 0.6408 |
| | 60 | 0.6902 | 0.7220 | 0.6214 |
| | 80 | 0.6733 | 0.6976 | 0.6067 |
| | 90 | 0.6384 | 0.6451 | 0.5345 |
| | 100 | 0.6280 | 0.6075 | 0.4953 |
| KaNCD | 0 | 0.7304 | 0.7785 | 0.6752 |
| | 20 | 0.7205 | 0.7624 | 0.6545 |
| | 40 | 0.7111 | 0.7528 | 0.6489 |
| | 60 | 0.7042 | 0.7332 | 0.6285 |
| | 80 | 0.6857 | 0.7057 | 0.6167 |
| | 90 | 0.6496 | 0.6642 | 0.5675 |
| | 100 | 0.6350 | 0.6083 | 0.5185 |

(a) Performance of two representative CD models on the **ASSIST** dataset




(b) Sparsity visualization on two datasets

**Fig. 2** Analyses about sparsity between students and concepts. Students are not directly related to concepts. In CD, if a student never or rarely conducts exercises related to a concept, we think that interactions between this student and the concept are sparse. **(a)** Performance of two representative neural network-based CD models on ASSIST dataset. We manually simulate different ratios ($\eta$) of sparsity between students and selected concepts. The simulation method can be found at **Varying Sparsity Levels** in Section 5.4. **(b)** Among all student-concept entries, over 92% entries correspond to sparse interactions on the ASSIST dataset, and over 98% entries correspond to sparse interactions on the MOOC-Radar dataset.

cepts $K(|K| = T)$ in an intelligent education system. There are two types of relationships among these entities. First, students will practice some exercises, forming response logs $\{(s, e, r_{se})\}$. If student $s$ answers exercise $e$ correctly, $r_{se} = 1$. Otherwise if student $s$ answers exercise $e$ wrongly, $r_{se} = 0$. $R_{train}$ and $R_{test}$ respectively denote the training and testing response logs. Second, the relations between exercises and concepts are denoted by $\mathbf{Q} = \{q_{ek}\}_{N\times T}$, where $q_{ek}$ denotes the relation between exercise $e$ and concept $k$. If exercise $e$ is related to concept $k$, $q_{ek} = 1$; otherwise, $q_{ek} = 0$. $\mathbf{Q}_e = [q_{e1}, ..., q_{ek}, ..., q_{eT}]$ denotes the relations between exercise $e$ and all concepts. Usually, $\mathbf{Q}$ is pre-defined by experts. Finally, we formulate the CD task as follows,

*Input:* The response logs $R_{train}$ and exercise-concept relations $\mathbf{Q}$.

*Output:* A CD model to infer student modeling $\mathbf{A}$ through response log prediction.

### 3.2 The Existing NCDM Framework

Neural network-based CD models mostly adopt the NCDM framework [7, 11, 12]. The framework uses $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_s, ..., \mathbf{a}_M]^\top \in \mathbb{R}^{M\times T}$ (students' mastery levels on concepts), and $\mathbf{a}_s = [a_{s1}, ..., a_{sk}, ..., a_{sT}]$. Each dimension of $\mathbf{a}_s$ has independent semantics, e.g., $a_{sk}$ denotes student $s$'s comprehension degree on concept $k$. NCDM framework also introduces exercise difficulties $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_e, ..., \mathbf{d}_N]^\top \in \mathbb{R}^{N\times T}$ and discrimination $\mathbf{h}^{disc} = [h_1^{disc}, ..., h_e^{disc}, ..., h_N^{disc}]^\top \in \mathbb{R}^{N\times 1}$. Then, a diagnostic function is used to obtain the hidden representation between student $s$ and exercise $e$, formulated as:

$$\mathbf{x}_{se} = \mathbf{Q}_e \odot (\sigma(\mathbf{a}_s) - \sigma(\mathbf{d}_e)) \times h_e^{disc}, \tag{1}$$

where $\sigma$ denotes the Sigmoid activation function, and $\mathbf{x}_{se}$ denotes the hidden vector for student $s$ and exercise $e$. $\mathbf{Q}_e$ denotes the relations between concepts and exercise $e$, and $\odot$ denotes the element-wise multiplication.

### 3.3 Effects of Student-concept Sparsity Barrier

First, we present a theoretical analysis of the effects of the student-concept sparsity barrier on NCDM. For a response log $\{(s, e, r_{se})\}$, suppose that exercise $e$ is related to concept $k$, i.e., the $k$-th dimension of $\mathbf{Q}_e$ is 1 and other dimensions are 0. Element-wise multiplication $\odot$ operation will lead to the sit-

uation that only the $k$-th dimension value of $(\sigma(\mathbf{a}_s) - \sigma(\mathbf{d}_e)) \times h_e^{disc}$ is not masked. Hence, the given log $\{(s, e, r_{se})\}$ will only update $c_{sk}$. Consequently, the NCDM framework would not update corresponding comprehension degrees if students do not or rarely interact with some concepts.

Second, we conduct a laboratory experiment on real-world datasets in Fig. 2 to further investigate relationships between NCDM framework and the sparsity barrier. In Fig. 2 (a), we intentionally discard some data to simulate different sparsity ratios ($\eta\%$) between students and randomly selected concepts. According to the results, the sparsity barrier has an obvious negative impact on the performance of two representative CD models (NCDM [7] and KaNCD [13]), e.g., KaNCD has an obvious AUC decrease of over 10% when transitioning from discarding 0% to discarding 80% interactions.

Third, the wide presence of sparsity significantly limits CD developments. As visualized in Fig. 2 (b), both two real-world datasets exhibit obvious sparsity. Moreover, in computerized adaptive testing, researchers aim to minimize the exercise number given to students but keep accurate diagnosis, leading to student-concept sparsity [40].

## 4 The Proposed Model

In this section, we draw inspiration from a classic educational work [14]. This work argues that student modeling should at least consider students' learned knowledge and their foundational abilities. In other words, to correctly answer an exercise, a student should not only understand the concept, but also do well in applying her understanding to the exercise. Therefore, we propose that a student's mastery levels on all concepts should consider two parts, formulated as:

$$\forall s \in S, \forall k \in K, a_{sk} = c_{sk} + g_s. \tag{2}$$

$g_s$ denotes student $s$'s application ability. $a_{sk}$ denotes student $a$'s mastery level on concept $k$. $c_{sk}$ denotes student $a$'s comprehension degree on concept $k$. Fig.3 illustrates the overall structure of our proposed ESR-CD. In the remaining parts of this section, we will introduce ESR-CD in detail.

### 4.1 NN based Matrix Factorization Layer

Although student $s$ does not interact with concept $k$, her similar students may provide sufficient interactions. Inspired by this, we propose to utilize similar students to enhance the quality of comprehension degrees. Specifically, we transfer matrix factorization techniques into CD. We first embed entities into latent embeddings of student, exercise, and concept as: $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_s, ..., \mathbf{u}_M]^\top \in \mathbb{R}^{M \times Z}$, $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_e, ..., \mathbf{v}_N]^\top \in \mathbb{R}^{N \times Z}$ and $\mathbf{O} = [\mathbf{o}_1, ..., \mathbf{o}_k, ..., \mathbf{o}_T]^\top \in \mathbb{R}^{T \times Z}$. $Z$ denotes the dimension of free embeddings.

Note that, we cannot adopt any model relying on explicit discrete adjacency matrices, e.g., WRMF [41], or graph-based models [39]. The outputs of matrix factorization, i.e., comprehension degrees on concepts and exercise difficulties, are both continuous latent values without discrete ground truth labels. For these graph-based models, we cannot obtain a discrete adjacency matrix [39]. For WRMF, we cannot assign different weights to each entry without ground truth labels [41]. To this end, we adopt GMF, a neural network based method [42]:

$$c_{sk} = \mathbf{W}(\mathbf{u}_s \odot \mathbf{o}_k), d_{ek} = \mathbf{W}(\mathbf{v}_e \odot \mathbf{o}_k), \tag{3}$$

where student $s$'s comprehension degrees on all concepts $\mathbf{c}_s = [c_{s1}, ..., c_{sk}, ..., c_{sT}]$, where $c_{sk}$ denotes the comprehension degree of student $s$ on concept $k$. Students' comprehension degrees can be represented as $\mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_s, ..., \mathbf{c}_M]^\top \in \mathbb{R}^{M \times T}$. Similarly, $\mathbf{d}_e = [d_{e1}, ..., d_{ek}, ..., d_{eT}]$, $\mathbf{D} = [\mathbf{d}_1, ..., \mathbf{d}_e, ..., \mathbf{d}_N]^\top \in \mathbb{R}^{N \times T}$. $\odot$ denotes the element-wise product, and $\mathbf{W}$ is a linear layer. Even if each row or column of $\mathbf{C}$

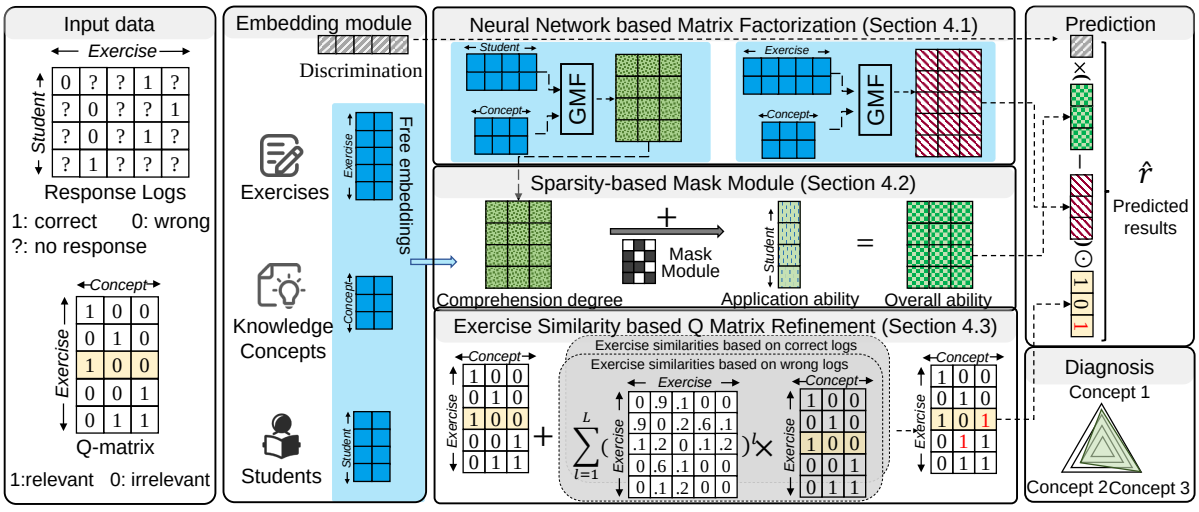



**Fig. 3** Overall structure of our proposed ESR-CD.

only involves a limited number of interactions, we can still optimize corresponding low-dimensional representations in $\mathbf{U}, \mathbf{V}$. Based on optimized low-dimensional representations, GMF enables better predicting $\mathbf{C}$ on non-interacted knowledge concepts.

## 4.2 Sparsity-based Mask Module

To model student $s$'s application ability $g_s$, previous works suggest estimating score deviation, i.e., average student scores [43], or a single learnable parameter [37]. In this paper, we propose that $g_s$ should be reflected in deviations of student performance on different concepts. Note that, $\mathbf{C}$ can also be seen as obtaining student performance on different concepts. An intuitive idea is to obtain student $s$'s application ability $g_s$ from $\mathbf{c}_s$. Furthermore, we find that dimensions in $\mathbf{C}$ corresponding to sparse student-concept interactions are less inaccurate than those corresponding to dense interactions. Therefore, we avoid effects from sparse student-concept interactions when modeling $g_s$. Following [44], we propose to estimate $g_s$ as:

$$g_s = \alpha_s\left(\frac{\sum \mathbf{b}_s \odot \mathbf{c}_s}{\sum \mathbf{b}_s}\right) + \mu_s, \tag{4}$$

where $\alpha_s$ and $\mu_s$ are both student-specific learnable parameters, and they form a student-specific mapping function. $\mathbf{b}_s = [b_{s1}, ..., b_{sk}, ..., b_{sT}]$ represents the mask module for student $s$. $b_{sk} = 1$ denotes that student $s$ has sufficient interactions with concept $k$, otherwise, $b_{sk} = 0$. $\mathbf{b}_s$ is designed for avoiding effects of sparse interactions on practical application ability $g_s$, formulated as $\frac{\sum \mathbf{b}_s \odot \mathbf{c}_s}{\sum \mathbf{b}_s}$. We provide an example to emphasize $\mathbf{b}_s$. Suppose there is a student John and five knowledge concepts. John has provided interactions with the first three concepts, and NCDM judges his mastery levels on these concepts to be 0.2, 0.8, 0.8, 0.5, and 0.5, respectively. Without $\mathbf{b}_s$, John's application ability will be estimated as (0.2 + 0.8 + 0.8 + 0.5 + 0.5) / 5 = 0.56. With $\mathbf{b}_s$, the estimated application ability should be (0.2 + 0.8 + 0.8) / 3 = 0.6. As 0.5 does not mean that John's mastery levels on the latter two knowledge concepts are exactly in an intermediate state, we choose to avoid their effects by adopting $\mathbf{b}_s$.

We argue that the small adjustment of modeling $g_s$ has important effects. To clearly show the effectiveness of $g_s$, we consider an example: student $s$ has not answered exercises related to concept $k$. We

focus on her mastery level on concept $k$ ($a_{sk}$). The estimation of $c_{sk}$ may be close to the initial value. $g_s$ can still be assessed as long as student $s$ provides sufficient interactions with any concepts as shown in Eq.(4). Therefore, we can make a preliminary estimation of $a_{sk}$ based on $g_s$.

### 4.3 Exercise Similarity based **Q** Matrix Refinement Layer

Previous studies have already identified the issue of incomplete annotations in **Q** [8, 13]. A classic work suggests that exercises with similar texts should be related to similar concepts [13]. However, it can not be adapted to datasets without exercise texts. In this paper, we borrow the idea that similar exercises should have a high probability of being assigned to the same concepts, and obtain exercise similarities based on historical response logs as follows:

$$
\begin{aligned}
I_{e_1e_2,0} &= \frac{\sum_{s\in S} \mathbb{1}_{r_{se_1}=r_{se_2}=0}}{\sqrt{|\sum_{s\in S} \mathbb{1}_{r_{se_1}=0}||\sum_{s\in S} \mathbb{1}_{r_{se_2}=0}|}}, \\
I_{e_1e_2,1} &= \frac{\sum_{s\in S} \mathbb{1}_{r_{se_1}=r_{se_2}=1}}{\sqrt{|\sum_{s\in S} \mathbb{1}_{r_{se_1}=1}||\sum_{s\in S} \mathbb{1}_{r_{se_2}=1}|}},
\end{aligned}
\tag{5}
$$

where $\mathbb{1}$ denotes the indicator function ($\mathbb{1}_{true} = 1$, $\mathbb{1}_{false} = 0$). $\sum_{s\in S} \mathbb{1}_{r_{se_1}=r_{se_2}=0}$ denotes the number of the same logs between exercise $e_1$ and $e_2$. $\sqrt{|\sum_{s\in S} \mathbb{1}_{r_{se_1}=0}||\sum_{s\in S} \mathbb{1}_{r_{se_2}=0}|}$ denotes the normalization. Quantities in $I_{e_1e_2,1}$ can be understood analogously. $I_{e_1e_2,0}$,$I_{e_1e_2,1}$ respectively denote similarities between exercise $e_1$ and $e_2$ according to incorrect/correct answers, thus forming $\mathbf{I}_0, \mathbf{I}_1 \in \mathbb{R}^{N\times N}$. Based on $\mathbf{I}_0, \mathbf{I}_1$, we propose to refine **Q** as follows:

$$
\hat{\mathbf{Q}} = (1 - \mathbf{Q}) \odot \Delta(\mathcal{D}_0^{-1}\mathbf{I}_0\mathbf{Q} + \mathcal{D}_1^{-1}\mathbf{I}_1\mathbf{Q}) + \mathbf{Q}, \tag{6}
$$

where $\mathcal{D}_0$ and $\mathcal{D}_1$ denote the degree matrices for $\mathbf{I}_0$ and $\mathbf{I}_1$, respectively. $\mathcal{D}_0^{-1}\mathbf{I}_0\mathbf{Q} + \mathcal{D}_1^{-1}\mathbf{I}_1\mathbf{Q}$ denotes the process of aggregation based on $\mathbf{I}_0, \mathbf{I}_1$. $\odot$ denotes the element-wise, and $(1 - \mathbf{Q}) \odot \Delta(\mathcal{D}_0^{-1}\mathbf{I}_0\mathbf{Q} + \mathcal{D}_1^{-1}\mathbf{I}_1\mathbf{Q})$ denotes that only zero-value elements in original **Q** will be considered. Then, we adopt a threshold operation $\Delta$. $\Delta$ sets entries in $(\mathcal{D}_0^{-1}\mathbf{I}_0\mathbf{Q} + \mathcal{D}_1^{-1}\mathbf{I}_1\mathbf{Q})$ larger than a certain value to 1, while leaving other entries as 0. $\hat{\mathbf{Q}}$ denotes updated exercise-concept relations. Note that, we also consider all one-value elements in original exercise-concept relations, and this refinement layer would ensure that more dimensions of students' comprehension degrees are adequately trained.

### 4.4 Prediction Layer

Following previous work [7, 11], we directly adopt a 1-dimensional embeddings for discrimination $\mathbf{h}^{disc} = [h_1^{disc}, ..., h_e^{disc}, ..., h_T^{disc}]$. Then, we obtain $\mathbf{x}_{se}$ as:

$$
\mathbf{x}_{se} = \hat{\mathbf{Q}}_e \odot (\sigma(\mathbf{a}_s) - \sigma(\mathbf{d}_e)) \times h_e^{disc}, \tag{7}
$$

where $\hat{\mathbf{Q}}_e$ denotes updated relations between exercise $e$ and all concepts. Then, $\hat{r}_{se} = \sigma(MLPs(\mathbf{x}_{se}))$. $\hat{r}_{se}$ is predicted response log. Similar to previous studies [7, 12, 13], we adopt a monotonicity assumption: the probability of correct response to the exercise is monotonically increasing at any dimension of the student's overall all abilities **A**. We make weight parameters in $MLPs$ be nonnegative to fit the assumption. Finally, we adopt the BCE loss:

$$
\mathcal{L} = -\sum_{s,e,r_{se}\in R_{train}} (r_{se}log(\hat{r}_{se}) + (1 - r_{se})log(1 - \hat{r}_{se})). \tag{8}
$$

To enhance clarity and facilitate understanding, we present the overall training procedures of our proposed ESR-CD in detail in Algorithm 1. We first initialize all trainable parameters, calculate exercise similarities to obtain the mask module **B** and exercise-concept relations $\hat{\mathbf{Q}}$. During training, we process batches of data by combining different embeddings based on the diagnosis function in Eq.(7, and mapping $\mathbf{x}_{se}$ to predicted response logs. After that, we minimize the loss to optimize trainable parameters. This process will be conducted repeatedly until the model converges.

**Algorithm 1** Training procedures of ESR-CD.

**Require:** Training response logs $R_{train}$, exercise-concept relations $\mathbf{Q}$.

**Ensure:**

Initialize all trainable parameters,
Calculate exercise similarities $\mathbf{I}_0$ and $\mathbf{I}_1$ (Eq.(5)), and obtain the mask module $\mathbf{B}$ (Eq.(4)),
Update $\hat{\mathbf{Q}}$ based on $\mathbf{I}_0$ and $\mathbf{I}_1$,
**repeat**
  Get a batch of training data $((s, e, r_{se}))$,
  **for** each $(s, e, r_{se})$ in the batch **do**
    Obtain $\mathbf{U}_s$, $\mathbf{V}_e$, $h_e^{disc}$,
    Obtain comprehension degrees $\mathbf{c}_s$ and difficulty $\mathbf{d}_e$ (Eq.(3)),
    Obtain application ability $g_s$ (Eq.(4)),
    Obtain mastery levels $\mathbf{a}_s$ (Eq.(2)),
    Incorporate $\hat{\mathbf{Q}}$, $\mathbf{a}_s$, $\mathbf{d}_e$, and $h_e^{disc}$ (Eq.(7)),
    Obtain prediction $\hat{r}_{se} = \sigma(MLPs(\mathbf{x}_{se}))$, and calculate loss (Eq.(8)),
  **end for**
  Minimize loss to optimize parameters.
**until** Model convergence.

## 5 Experiments

In this section, we will answer the following Research Questions (RQ):

RQ1: Does ESR-CD have consistently superior performance? (Section 5.2)

RQ2: Are all components essential? (Section 5.3)

RQ3: How do choices of hyper-parameters or model structures affect ESR-CD? (Section 5.4)

RQ4: Are mastery levels from ESR-CD consistent with all response logs? (Section 5.5)

RQ5: What are the detailed diagnosis results of ESR-CD? (Section 5.6)

**Table 1** The detailed statistics of two datasets.

| Dataset | ASSIST | MOOC-Radar |
|---|---|---|
| #Students | 2,493 | 14,224 |
| #Exercises | 17,746 | 2,513 |
| #Knowledge concepts | 123 | 580 |
| #Response logs | 267,415 | 898,933 |
| #Response logs per student | 107.266 | 63.198 |
| #Knowledge concepts per exercise | 1.192 | 1 |
| #Sparsity in student-concept interactions | 92.212% | 98.985% |

### 5.1 Experimental Settings

*Datasets.* We select two publicly available datasets, i.e., ASSIST and MOOC-Radar. ASSIST is an open dataset containing abundant student-exercise logs and expert-labeled exercise-concept relationships $\mathbf{Q}$. MOOC-Radar is a recently collected dataset from students' learning records in MOOCs [45], which contains abundant response logs. In addition to response logs and exercise-concept relationships, MOOC-Radar also provides much rich side information, e.g., exercise-related cognitive level labels, exercise texts, interactions between students and videos, and so on. In this paper, we focus on how to mine rich information only from response logs to alleviate data sparsity. The statistics of these two datasets are in Table 1.

For comprehensive validation, we employ two data split settings. One is *Random Split*. We randomly divide the student-exercise response logs into training, validation, and testing sets with a ratio of 7:1:2. The other is *Concept Weak-coverage Split* [13]. For each student, we first identify all the concepts they have encountered. Then, we randomly select 20% of the concepts for each student and assign the student-exercise pairs related to these selected concepts to the testing set. After that, we divide the remaining response logs into the training and validation sets with a ratio of 7:1. In this way, 180,463/25,781/61,171 logs are assigned to the training/validation/testing set on the ASSIST

dataset, respectively. 596,269/85,182/217,482 logs are assigned to the training/validation/testing set on the MOOC-Radar dataset.

*Metrics.* Following previous works [7, 11, 12], we employ Accuracy (ACC) and Area Under the Curve (AUC) to evaluate response log prediction performance. Additionally, to assess the consistency between predicted logs in the testing set and students' mastery levels on concepts, we adopt a common Degree Of Agreement (DOA) metric [1, 7, 13]. We calculate the DOA values over the whole dataset, denoted as DOA@w.

*Baselines.* We choose two kinds of models. One kind is classic models without providing detailed students' comprehension degrees on concepts, therefore, we cannot measure DOA for them.

- IRT [15]. It uses one dimensional variables to model student abilities and exercise difficulties.
- MIRT [46]. It allows for high dimensions to enhance prediction performance.

The other kind is representative neural network based CD models. Apart from prediction, these models can provide students' mastery levels each knowledge concept.

- DINA [21]. It utilizes binary variables to represent student and exercise entities and adopts guess and slip parameters.
- NCDM [7]. It adopts high dimensional representations for comprehension degrees and difficulties with a classic diagnostic function.
- CDMFCK [47]. Compared to NCDM, it considers different effects of knowledge concepts.
- RCD [11]. It further utilizes concept relationships and student-exercise-concept relationships.
- KSCD [48]. It uses probabilistic matrix factorization and devises a novel interaction function.
- KaNCD [13]. Compared to NCDM, it utilizes collaborative information between students and concepts via matrix factorization.

*Hyper-parameters.* For fair comparisons, we set the dimension of latent representations in KSCD, KaNCD, and ESR-CD to 64 on the ASSIST dataset, and 256 on the MOOC-Radar dataset. For the MLP mapping $\mathbf{x}_{se}$ to prediction, we adopt the same structure (hidden dimensions are 256, 128, respectively) for all neural network based CD models. We search the learning rate in the range of {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.02}, and set the batch size to 8,192 for all models. If student $s$ does not interact with concept $k$, we set $b_{sk} = 0$ on both two datasets; otherwise, $b_{sk} = 1$. As for the threshold $\Delta$ in the refinement layer, we set the threshold to 0.5.

## 5.2 Overall Performance (RQ 1)

We report overall performance in Table 2-3. We have several observations from these tables.

- First, ESR-CD has the best performance on two datasets and two experimental settings. E.g., ESR-CD has nearly 0.5% AUC improvements under two settings on the ASSIST dataset. Compared to the representative neural network based CD model KaNCD, ESR-CD has an AUC improvement of over 6% under the concept weak-coverage split on MOOC-Radar dataset.
- Second, ESR-CD shows stronger robustness than other neural network-based CD models. E.g., compared to KaNCD, ESR-CD has a slight AUC improvement of less than 0.5% under the random splits and an AUC improvement of over 7% under the challenging concept weak-coverage split on MOOC-Radar.
- Last but not least, NCDM retains some predictive performance under the concept weak-coverage setting. Most comprehension degrees in NCDM would be close to initial values under this setting, however, NCDM achieves AUC scores of

**Table 2** Overall performance on the ASSIST dataset. We use bold font to emphasize the best results.

| Model | Random Split | | | Concept Weak-coverage Split | | |
|---|---|---|---|---|---|---|
| | AUC ↑ | ACC ↑ | DOA ↑ | AUC ↑ | ACC ↑ | DOA ↑ |
| **IRT** | 0.7298 ± 0.0010 | 0.7096 ± 0.0004 | - | 0.6976 ± 0.0016 | 0.7137 ± 0.0002 | - |
| **MIRT** | 0.7454 ± 0.0003 | 0.7132 ± 0.0006 | - | 0.6936 ± 0.0014 | 0.7108 ± 0.0012 | - |
| **DINA** | 0.7178 ± 0.0017 | 0.6703 ± 0.0109 | 0.5998 ± 0.0034 | 0.6169 ± 0.0031 | 0.4774 ± 0.0211 | 0.5305 ± 0.0173 |
| **NCDM** | 0.7209 ± 0.0002 | 0.6745 ± 0.0021 | 0.5985 ± 0.0124 | 0.6499 ± 0.0007 | 0.6709 ± 0.0006 | 0.5109 ± 0.0041 |
| **CDMFKC** | 0.7431 ± 0.0011 | 0.7038 ± 0.0016 | 0.6040 ± 0.0051 | 0.6742 ± 0.0005 | 0.6798 ± 0.0090 | 0.4995 ± 0.0079 |
| **RCD** | 0.7472 ± 0.0010 | 0.7111 ± 0.0003 | 0.5809 ± 0.0110 | 0.6924 ± 0.0034 | 0.7026 ± 0.0019 | 0.5623 ± 0.0049 |
| **KSCD** | 0.7586 ± 0.0002 | 0.7207 ± 0.0005 | 0.5092 ± 0.0031 | 0.7009 ± 0.0009 | 0.7114 ± 0.0039 | 0.4987 ± 0.0150 |
| **KaNCD** | 0.7599 ± 0.0003 | 0.7271 ± 0.0011 | 0.6486 ± 0.0110 | 0.6952 ± 0.0045 | 0.7098 ± 0.0034 | 0.5808 ± 0.0089 |
| **ESR-CD** | **0.7640 ± 0.0004** | **0.7297 ± 0.0003** | **0.6550± 0.0095** | **0.7051 ± 0.0006** | **0.7153 ± 0.0024** | **0.6104 ± 0.0145** |

**Table 3** Overall performance on the MOOC-Radar dataset. We use bold font to emphasize the best results.

| Model | Random Split | | | Concept Weak-coverage Split | | |
|---|---|---|---|---|---|---|
| | AUC ↑ | ACC ↑ | DOA ↑ | AUC ↑ | ACC ↑ | DOA ↑ |
| **IRT** | 0.8534 ± 0.0010 | 0.8522 ± 0.0008 | - | 0.7625 ± 0.0019 | 0.8328 ± 0.0013 | - |
| **MIRT** | 0.8727 ± 0.0022 | 0.8629 ± 0.0016 | - | 0.7671 ± 0.0036 | 0.8371 ± 0.0017 | - |
| **DINA** | 0.8070 ± 0.0009 | 0.7729 ± 0.0005 | 0.6000 ± 0.0079 | 0.5745 ± 0.0022 | 0.4978 ± 0.0014 | 0.5106 ± 0.0114 |
| **NCDM** | 0.8656 ± 0.0022 | 0.8526 ± 0.0010 | 0.6024 ± 0.0089 | 0.6701 ± 0.0004 | 0.7876 ± 0.0003 | 0.5027 ± 0.0062 |
| **CDMFKC** | 0.8678 ± 0.0028 | 0.8559 ± 0.0015 | 0.6050 ± 0.0080 | 0.6831 ± 0.0037 | 0.7756 ± 0.0014 | 0.4962 ± 0.0136 |
| **RCD** | 0.8699 ± 0.0045 | 0.8634 ± 0.0031 | 0.6344 ± 0.0101 | 0.6903 ± 0.0064 | 0.7814 ± 0.0082 | 0.5332 ± 0.0230 |
| **KSCD** | 0.8754 ± 0.0048 | 0.8598 ± 0.0020 | 0.5283 ± 0.0094 | 0.6952 ± 0.0023 | 0.7935 ± 0.0029 | 0.4949 ± 0.0125 |
| **KaNCD** | 0.8809 ± 0.0008 | 0.8646 ± 0.0005 | 0.6918 ± 0.0115 | 0.7182 ± 0.0022 | 0.8264 ± 0.0013 | 0.6051 ± 0.0032 |
| **ESR-CD** | **0.8836 ± 0.0006** | **0.8671 ± 0.0002** | **0.7249 ± 0.0047** | **0.7718 ± 0.0004** | **0.8430 ± 0.0011** | **0.6601 ± 0.0066** |

0.6499 and 0.6701 on two datasets. The reason is that NCDM still assesses high-quality exercise parameters in Eq.(7) for prediction.

### 5.3 Ablation Study (RQ 2)

We compare these settings: base models (NCDM), NCDM with incorporation (Eq.(2)), NCDM with incorporation and Q refinements (Eq.(6)), and NCDM with incorporation and matrix factorization (Eq.(3)).

There are some observations from Table 4 and Table 5. First, incorporation achieves AUC scores of 0.7004 and 0.7670 on two datasets. The improvements are over 7% compared to the basic model NCDM, demonstrating the effectiveness of $g_s$. Second, incorporation brings a much smaller improvement on the random split setting (nearly 1% improvement on the MOOC-Radar dataset). Third, the refinement leads to obvious improvements. E.g., incorporation with refinement (AUC 0.7619) performs much better than only incorporation (AUC 0.7541) under the random split setting on the ASSIST dataset. Fourth, factorization also has obvious positive effects. E.g., there is a significant performance improvement in the random split setting when comparing $g_s$ with/without MF (As shown in the second and fourth rows in Table 4, AUC 0.7541 → 0.7619). Finally, by combining all modules, ESR-CD achieves the best performance.

### 5.4 Model Analyses (RQ 3)

**Comparisons among Different Matrix Factorization Techniques.** In this part, we compare some feasible methods, i.e., PMF [43], GMF and NCF [42]. As shown in Fig.4, we first find that GMF and NCF demonstrate stable comparable performance on both two datasets. For example, on the MOOC-Radar dataset, GMF achieves AUC scores of 0.8836, and NCF achieves AUC scores of 0.8847 on the ran-

**Table 4** Ablation study on the ASSIST dataset. The last row represents our proposed ESR-CD.

| Incorporation | Refinement | Factorization | Random Split | | | Concept Weak-coverage Split | | |
|---|---|---|---|---|---|---|---|---|
| Eq.(2) | Eq.(6) | Eq.(3) | AUC ↑ | ACC ↑ | DOA ↑ | AUC ↑ | ACC ↑ | DOA ↑ |
| ✗ | ✗ | ✗ | 0.7209 | 0.6745 | 0.5985 | 0.6499 | 0.6709 | 0.5109 |
| ✓ | ✗ | ✗ | 0.7541 | 0.7209 | 0.6462 | 0.7004 | 0.6814 | 0.6014 |
| ✓ | ✓ | ✗ | 0.7598 | 0.7271 | 0.6403 | 0.7014 | 0.7032 | 0.5715 |
| ✓ | ✗ | ✓ | 0.7619 | 0.7283 | 0.6456 | 0.7030 | 0.7120 | 0.5868 |
| ✓ | ✓ | ✓ | **0.7640** | **0.7297** | **0.6550** | **0.7051** | **0.7153** | **0.6104** |

**Table 5** Ablation study on the MOOC-Radar dataset. The last row represents our proposed ESR-CD.

| Incorporation | Refinement | Factorization | Random Split | | | Concept Weak-coverage Split | | |
|---|---|---|---|---|---|---|---|---|
| Eq.(2) | Eq.(6) | Eq.(3) | AUC ↑ | ACC ↑ | DOA ↑ | AUC ↑ | ACC ↑ | DOA ↑ |
| ✗ | ✗ | ✗ | 0.8656 | 0.8526 | 0.6024 | 0.6701 | 0.7676 | 0.5027 |
| ✓ | ✗ | ✗ | 0.8781 | 0.8590 | 0.7131 | 0.7670 | 0.8334 | 0.6510 |
| ✓ | ✓ | ✗ | 0.8827 | 0.8646 | 0.7230 | 0.7710 | 0.8414 | 0.6524 |
| ✓ | ✗ | ✓ | 0.8824 | 0.8653 | 0.7201 | 0.7703 | 0.8364 | 0.6509 |
| ✓ | ✓ | ✓ | **0.8836** | **0.8671** | **0.7249** | **0.7718** | **0.8430** | **0.6601** |

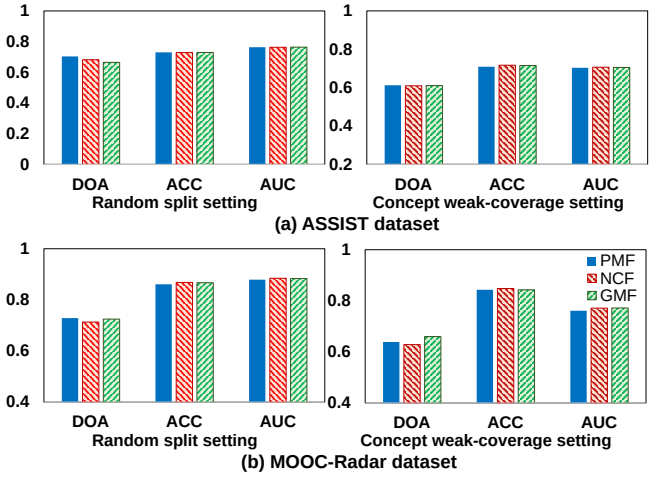



**Fig. 4** Comparisons about different matrix factorization techniques on ASSIST and MOOC-Radar datasets.

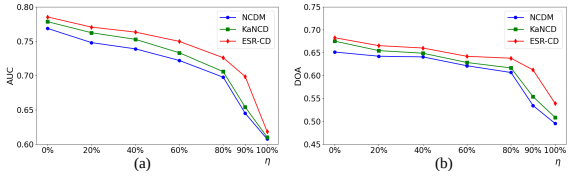



**Fig. 5** (a) AUC and (b) DOA performance with varying sparsity ratios $\eta$ (%) on the **ASSIST** dataset.

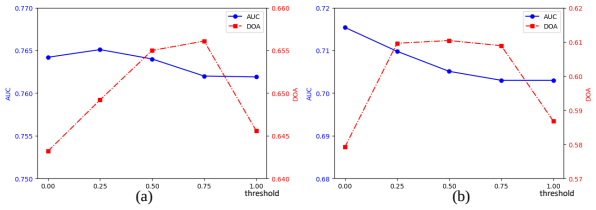



**Fig. 6** Varying threshold $\Delta$ on the ASSIST dataset under (a) random split and (b) concept weak coverage split settings.

dom split setting. Second, we find that these two models slightly outperform PMF on both two datasets especially for accuracy performance (e.g., AUC) under the concept weak-coverage split. E.g., on MOOC-Radar, PMF achieves AUC scores of 0.7610, while GMF and NCF achieve AUC scores of 0.7718 and 0.7711, respectively. In conclusion, GMF and NCF have similar performance and both outperform PMF. Considering the relatively simpler structure of GMF, we choose GMF.

## 5.5 Consistency on the Whole Dataset (RQ 4)

**Varying Sparsity Levels.** To evaluate performance under different data sparsity levels, we conduct a comparative analysis between ESR-CD, NCDM [7] and KaNCD [13]). To simulate sparsity levels $\eta$, we first select 20 concepts for each student. For each student and selected concepts: 1. The interactions between each student and selected concepts are split into a candidate set and a test set at a ra-

**Table 6** Comparisons among $\sigma(c_{sk} \cdot g_s)$, $\sigma(\sigma(c_{sk}) \cdot g_s)$ and $\sigma(c_{sk} + g_s)$ on the ASSIST dataset.

| Formulation | Random Split | | | Concept Weak-coverage Split | | |
|---|---|---|---|---|---|---|
| | AUC ↑ | ACC ↑ | DOA ↑ | AUC ↑ | ACC ↑ | DOA ↑ |
| $\sigma(c_{sk} \cdot g_s)$ | 0.7610 ± 0.0006 | 0.7283 ± 0.0004 | 0.6537 ± 0.0092 | 0.7025 ± 0.0008 | 0.7122 ± 0.0011 | 0.6004 ± 0.0120 |
| $\sigma(\sigma(c_{sk}) \cdot g_s)$ | 0.7556 ± 0.0008 | 0.7239 ± 0.0005 | 0.6506 ± 0.0101 | 0.7045 ± 0.0011 | 0.7147 ± 0.0031 | 0.5959 ± 0.0166 |
| $\sigma(c_{sk} + g_s)$ | **0.7640 ± 0.0004** | **0.7297 ± 0.0003** | **0.6550± 0.0095** | **0.7051 ± 0.0006** | **0.7153 ± 0.0024** | **0.6104 ± 0.0145** |

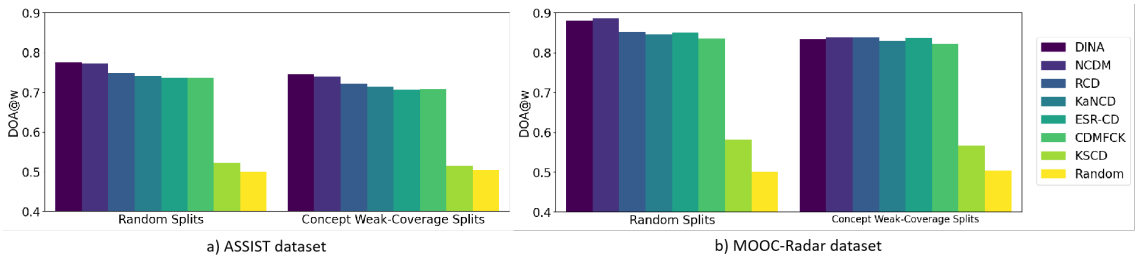



**Fig. 7** DOA@w on both ASSIST and MOOC-Radar datasets.

tio of 8:2. The candidate set is used to augment the training set; 2. To control interaction data sparsity, we randomly remove a ratio ($\eta$) of interactions from the candidate set. The process is consistent with Fig.2 (a). The larger value of $\eta$ denotes the higher sparsity levels.

We have several observations from Fig.5. First, ESR-CD outperforms KaNCD and NCDM under different sparsity levels, demonstrating the effectiveness and robustness of ESR-CD. Second, as the data becomes abundant ($\eta$=0%), KaNCD approaches ESR-CD. When $\eta$=0%, KaNCD achieves an AUC score of 0.7785, and ESR-CD achieves 0.7854. Third, as the data becomes very sparse ($\eta$=100%), all three models have an obvious performance decrease. However, ESR-CD performs much better than both NCDM and KaNCD. These results prove the stable effectiveness of ESR-CD.

**Varying Threshold in the Refinement Layer.** To further explore the effects of the **Q** matrix refinement, we conduct additional experiments by adjusting parameters related to the refinement layer to verify its impacts. Specifically, we adjust the threshold operation $\Delta$ on the ASSIST dataset, setting it to 0, 0.1, 0.25, 0.5, 0.75, and 1. The results are recorded in Figure 6. As the threshold decreases, ESR-CD allows more entries to be marked as 1, which improves student performance prediction (AUC) to some extent. However, this severely impacts the consistency metric (DOA), indicating that the diagnosis performance is compromised. To strike a balance between AUC and DOA, we set the threshold to 0.5 in our experiments.

**Comparing Different Combination Ways.** In this part, we discuss how to combine $g_s$ and $c_{sk}$. Apart from $\sigma(c_{sk} + g_s)$ in ESR-CD, we also consider $\sigma(c_{sk} \cdot g_s)$ and $\sigma(\sigma(c_{sk}) \cdot g_s)$, where $\sigma$ denotes the Sigmoid activation function. If we adopt $a_{sk} = \sigma(c_{sk} \cdot g_s)$, the derivative with respect to $g_s$ becomes: $\frac{\partial a_{sk}}{\partial g_s} = c_{sk} \cdot \sigma(c_{sk} \cdot g_s) \cdot (1 - \sigma(c_{sk} \cdot g_s))$. Obviously, $\sigma(c_{sk} \cdot g_s)$ and $(1 - \sigma(c_{sk} \cdot g_s))$ are both positive. The maintenance of positive correlation depends on whether $c_{sk}$ is positive. One option is to apply another Sigmoid activation to $c_{sk}$. If sigmoid activation is applied to both $c_{sk}$ and $a_{sk}$, the derivative becomes: $\frac{\partial a_{sk}}{\partial g_s} = \sigma(c_{sk}) \cdot \sigma(\sigma(c_{sk}) \cdot g_s) \cdot (1 - \sigma(\sigma(c_{sk}) \cdot g_s))$. The use of consecutive sigmoid layers can result in gradient vanishing.

We also conduct additional experiments to compare $\sigma(c_{sk} \cdot g_s)$, $\sigma(\sigma(c_{sk}) \cdot g_s)$ and $\sigma(c_{sk} + g_s)$. The

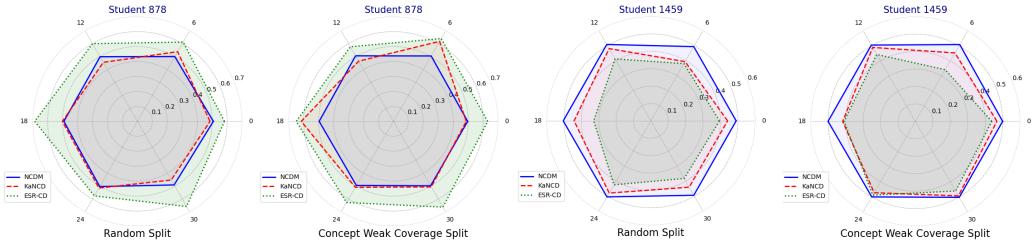



**Fig. 8** Examples of cognitive diagnosis results on the ASSIST dataset.

relevant experimental results are recorded in Table 6. First, we find that $\sigma(c_{sk} + g_s)$ outperforms $\sigma(c_{sk} \cdot g_s)$ and $\sigma(\sigma(c_{sk}) \cdot g_s)$ under both experimental settings. Second, when comparing $\sigma(c_{sk} \cdot g_s)$ and $\sigma(\sigma(c_{sk}) \cdot g_s)$, we find that $\sigma(c_{sk} \cdot g_s)$ achieves better performance under the classic random split setting, indicating that the conventional diagnosis task setup heavily relies on optimizing $c_{sk}$. This is also an answer to why we do not completely remove $c_{sk}$ and only use $g_s$. Third, under the concept weak-coverage setting, $\sigma(\sigma(c_{sk}) \cdot g_s)$ yields better results than $\sigma(c_{sk} \cdot g_s)$, which proves that the positive correlation between $g_s$ and $a_{sk}$ is more crucial if we need to predict a student's mastery levels on non-interacted concepts.

In Table 2-3, we calculate DOA on the testing set. Some works suggest that calculating *DOA* on the *w*hole dataset (DOA@w) [1, 7, 13]. To better display the performance of our proposed ESR-CD, we also report DOA@w. As shown in Figure 7, DINA and NCDM achieve the highest DOA@w on both two datasets. Our proposed ESR-CD maintains similar performance to other neural network-based CD models (KaNCD, RCD, and CDMFCK) (slightly worse than DINA and NCDM), which proves the effectiveness of our proposed ESR-CD. Note that, both KSCD and KaNCD argue that the outputs of matrix factorization represent students' mastery levels on concepts. However, KaNCD directly uses the mastery levels in the diagnostic function, whereas KSCD introduces an additional feature fusion layer, which leads to a decrease in consistency.

### 5.6 Case Study (RQ 5)

To display detailed cognitive diagnosis results, we randomly select two students (878 and 1459) from the ASSIST dataset and record their mastery levels on concepts 0, 6, 12, 18, 24, and 30. As a comparison, we include two classic neural network-based cognitive diagnosis models (NCDM and KaNCD). The results are plotted in radar charts in Figure 8.

As illustrated in Figure 8, our proposed ESR-CD tends to push the mastery levels away from 0.5. It is important to emphasize that the range for mastery levels on concepts is 0-1. From the perspective of the information theoretic, 0.5 represents maximum uncertainty, as it indicates the equal likelihood of mastery or non-mastery. For student 1459, it pushes the values furthest towards 1. For student 878, it pushes furthest towards 0. All these results can prove that our proposed ESR-CD can push mastery levels $a_{sk}$ more away from 0.5 than NCDM and KaNCD, which conveys more useful information for diagnosis.

## 6 Conclusion and Future Work

In this paper, we pointed out that the student-concept sparsity barrier would limit CD development. To

this end, we proposed ESR-CD, which simultaneously considers application abilities and comprehension degrees for modeling students' mastery levels on concepts. One one hand, for application abilities, we proposed a sparsity-based mask module that solely to avoid unwanted effects from sparse entries between students and concepts. On the other hand, we also proposed a neural network based matrix factorization layer and a exercise-concept relation refinement layer to enhance the optimization of comprehension degrees. Extensive experiments on two real-world datasets demonstrated the stable effectiveness of our proposed ESR-CD. In the future, we aim to explore other directions in CD, e.g., exploring concept dependencies by prompting large language models or preventing the amplification of unwanted biases in CD.

**Acknowledgements** This work has been supported in part by grants from the National Science and Technology Major Project (2021ZD0111802), the New Cornerstone Science Foundation through the XPLORER PRIZE, the National Natural Science Foundation of China (72188101, 62376086), and Joint Funds of the National Natural Science Foundation of China (U22A2094).

**Competing interests** The authors declare that they have no competing interests or financial conflicts to disclose.

## References

1. Chen X, Wu L, Liu F, Chen L, Zhang K, Hong R, Wang M. Disentangling cognitive diagnosis with limited exercise labels. 2023, 36: 18028–18045
2. Liu Y J, Zhang T C, Wang X C, Yu G, Li T. New development of cognitive diagnosis models. Frontiers of Computer Science, 2023, 17(1): 171604
3. Shen J, Qian H, Zhang W, Zhou A. Symbolic cognitive diagnosis via hybrid optimization for intelligent education systems. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 14928–14936
4. Gao W, Wang H, Liu Q, Wang F, Lin X, Yue L, Zhang Z, Lv R, Wang S. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval. 2023, 983–992
5. Liu H Y, Zhang T C, Li F, Yu M H, Yu G. A probabilistic generative model for tracking multi-knowledge concept mastery probability. Frontiers of Computer Science, 2024, 18(3): 183602
6. Dai H, Zhang Y, Yun Y, An R, Zhang W, Shang X. Adaptive meta-knowledge dictionary learning for incremental knowledge tracing. Engineering Applications of Artificial Intelligence, 2024, 132: 107969
7. Wang F, Liu Q, Chen E, Huang Z, Chen Y Y, Yin Y, Huang Z, Wang S. Neural cognitive diagnosis for intelligent education systems. In: Proceedings of the AAAI conference on artificial intelligence. 2020, 6153–6161
8. Liu S, Qian H, Li M, Zhou A. Qccdm: A q-augmented causal cognitive diagnosis model for student learning. In: Proceedings of the 26th European Conference on Artificial Intelligence. 2023, 1536–1543
9. Wang F, Gao W, Liu Q, Li J, Zhao G, Zhang Z, Huang Z, Zhu M, Wang S, Tong W, Chen E. A survey of models for cognitive diagnosis: New developments and future directions. arXiv preprint arXiv:2407.05458, 2024
10. Liu Q. Towards a new generation of cognitive diagnosis. In: IJCAI. 2021, 4961–4964
11. Gao W, Liu Q, Huang Z, Yin Y, Bi H, Wang M C, Ma J, Wang S, Su Y. Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. 2021, 501–510
12. Li J, Wang F, Liu Q, Zhu M, Huang W, Huang Z, Chen E, Su Y, Wang S. Hiercdf: A bayesian network-based hierarchical cognitive diagnosis framework. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022, 904–913
13. Wang F, Liu Q, Chen E, Huang Z, Yin Y, Wang S, Su Y. Neuralcd: a general framework for cognitive diagnosis. IEEE Transactions on Knowledge and Data Engineering, 2022
14. Burke K. The Mindful School: How To Assess Thoughtful Outcomes. K-College. ERIC, 1993
15. Embretson S E, Reise S P. Item response theory. Psychology Press, 2013
16. Ma H, Wang C Q, Zhu H S, Yang S S, Zhang X M, Zhang X Y. Enhancing cognitive diagnosis using uninteracted exercises: A collaboration-aware mixed sampling approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 8877–8885
17. Gao W, Liu Q, Wang H, Yue L, Bi H, Gu Y, Yao F, Zhang Z, Li X, He Y. Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives. In: Proceed-

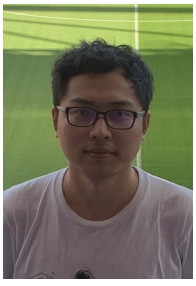
ings of the AAAI Conference on Artificial Intelligence. 2024, 8417–8426

18. Li M, Qian H, Lv J L, He M, Zhang W, Zhou A. Foundation model enhanced derivative-free cognitive diagnosis. Frontiers of Computer Science, 2025, 19(1): 191318
19. Zhang Z, Liu Q, Jiang H, Wang F, Zhuang Y, Wu L, Gao W, Chen E. Fairlisa: Fair user modeling with limited sensitive attributes information. In: Thirty-seventh Conference on Neural Information Processing Systems. 2023
20. Zhang Y, Dai H, Yun Y, Liu S, Lan A, Shang X. Meta-knowledge dictionary learning on 1-bit response data for student knowledge diagnosis. Knowledge-Based Systems, 2020, 205: 106290
21. De La Torre J. Dina model and parameter estimation: A didactic. Journal of educational and behavioral statistics, 2009, 34(1): 115–130
22. Wang S, Zeng Z, Yang X, Zhang X Y. Self-supervised graph learning for long-tailed cognitive diagnosis. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 110–118
23. Qian H, Liu S, Li M, Li B, Liu Z, Zhou A. Orcdf: An oversmoothing-resistant cognitive diagnosis framework for student learning in online education systems. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024, 2455–2466
24. Wu L, Chen X, Liu F, Xie J, Xia C A, Tan Z T, Tian M, Li J L, Zhang K, Lian D F, Hong R, Wang M. Edustudio: towards a unified library for student cognitive modeling. Frontiers of Computer Science, 2025, 19(8): 198342
25. Liu S, Shen J, Qian H, Zhou A. Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In: Proceedings of the ACM on Web Conference 2024. 2024, 4260–4271
26. Zhang Y, Qin C, Shen D, Ma H, Zhang L, Zhang X, Zhu H. Relicd: a reliable cognitive diagnosis framework with confidence awareness. In: 2023 IEEE International Conference on Data Mining (ICDM). 2023, 858–867
27. Li J, Liu Q, Wang F, Liu J Y, Huang Z, Yao F Z, Zhu L B, Su Y. Towards the identifiability and explainability for personalized learner modeling: An inductive paradigm. In: Proceedings of the ACM on Web Conference 2024. 2024, 3420–3431
28. Zhang Z, Wu L, Liu Q, Liu J, Huang Z, Yin Y, Zhuang Y, Gao W, Chen E. Understanding and improving fairness in cognitive diagnosis. Science China Information Sciences, 2024, 67(5): 152106
29. Zhang D, Zhang K, Wu L, Tian M, Hong R, Wang M. Path-specific causal reasoning for fairness-aware cognitive diagnosis. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024, 4143–4154
30. Zhang Z, Liu Q, Hu Z, Zhan Y, Huang Z, Gao W, Mao Q. Enhancing fairness in meta-learned user modeling via adaptive sampling. In: Proceedings of the ACM on Web Conference 2024. 2024, 3241–3252
31. Chen L, Wu L, Zhang K, Hong R, Lian D, Zhang Z, Zhou J, Wang M. Improving recommendation fairness via data augmentation. In: Proceedings of the ACM Web Conference 2023. 2023, 1012–1020
32. Jiang Y, Ma H, Zhang X, Li Z, Chang L. Incorporating metapath interaction on heterogeneous information network for social recommendation. Frontiers of Computer Science, 2024, 18(1): 181302
33. Gao C, Wang S, Li S, Chen J, He X, Lei W, Li B, Zhang Y, Jiang P. Cirs: Bursting filter bubbles by counterfactual interactive recommender system. ACM Transactions on Information Systems, 2023, 42(1): 1–27
34. Cai M, Hou M, Chen L, Wu L, Bai H, Li Y, Wang M. Mitigating recommendation biases via group-alignment and global-uniformity in representation learning. ACM Transactions on Intelligent Systems and Technology, 2024
35. Cai M, Chen L, Wang Y, Bai H, Sun P, Wu L, Zhang M, Wang M. Popularity-aware alignment and contrast for mitigating popularity bias. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024, 187–198
36. Huang L L, Ma H F, He X C, Chang L. Multi-affect (ed): improving recommendation with similarity-enhanced user reliability and influence propagation. Frontiers of Computer Science, 2021, 15: 1–11
37. Shuai J, Zhang K, Wu L, Sun P, Hong R, Wang M, Li Y. A review-aware graph contrastive learning framework for recommendation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022, 1283–1293
38. Zhang J, Zhu Y, Liu Q, Wu S, Wang S H, Wang L. Mining latent structures for multimedia recommendation. In: Proceedings of the 29th ACM international conference on multimedia. 2021, 3872–3880
39. He X, Deng K, Wang X, Li Y, Zhang Y, Wang M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval. 2020, 639–648
40. Zhuang Y, Liu Q, Zhao G, Huang Z, Huang W, Pardos Z, Chen E, Wu J, Li X. A bounded ability estimation

for computerized adaptive testing. In: Thirty-seventh Conference on Neural Information Processing Systems. 2023
41. Hu Y, Koren Y, Volinsky C. Collaborative filtering for implicit feedback datasets. In: 2008 Eighth IEEE International Conference on Data Mining. 2008, 263–272
42. He X, Liao L, Zhang H, Nie L, Hu X, Chua T S. Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. 2017, 173–182
43. Mnih A, Salakhutdinov R R. Probabilistic matrix factorization. Advances in neural information processing systems, 2007, 20
44. Wu C, Wang X, Lian D, Xie X, Chen E. A causality inspired framework for model interpretation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023, 2731–2741
45. Yu J F, Lu M Y, Zhong Q Y, Yao Z J, Tu S Q, Liao Z S, Li X Y, Li M L, Hou L, Zheng H T, Li J Z, Tang J. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. SIGIR '23. 2023, 2924–2934
46. Reckase M D. The past and future of multidimensional item response theory. Applied Psychological Measurement, 1997, 21(1): 25–36
47. Li S, Guan Q, Fang L, Xiao F, He Z, He Y, Luo W. Cognitive diagnosis focusing on knowledge concepts. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022, 3272–3281
48. Ma H, Li M, Wu L, Zhang H, Cao Y, Zhang X Y, Zhao X. Knowledge-sensed cognitive diagnosis for intelligent education platforms. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022, 1451–1460

Pengyang Shao is currently pursuing a PhD degree at Hefei University of Technology (HFUT), China. He received his Bachelor's degree in 2019 from the same university. His research interest lies on data mining, and large language models. He has published several papers in leading conferences and journals, including KDD, WWW, ACM TOIS and SCIS.

Kun Zhang received a PhD degree in computer science and technology from the University of Science and Technology of China in 2019. He is currently a faculty member at the Hefei University of Technology (HFUT), China. His research interests include Natural Language Understanding and Recommender Systems. He has published several papers in referred journals and conferences, such as IEEE TSMC:S, IEEE TKDE, ACM TKDD, AAAI, KDD, ACL, and ICDM. He received the KDD 2018 Best Student Paper Award.



Chen Gao is now a Faculty Member (Research-track AP) of BNRist, Tsinghua University. He obtained his Ph.D. Degree (advised by Prof. Yong Li and Prof. Depeng Jin) and Bachelor's Degree from the Department of Electronic Engineering, Tsinghua University in 2021 and 2016, respectively. His research primarily focuses on data mining (recommender system and spatio-temporal data mining), large language model, embodied agent, etc., with over 60 papers in top-tier venues (50+ CCF-A), attracting over 4,000 citations.

Lei Chen is currently a postdoctoral researcher at Tsinghua University, China. He received his PhD from Hefei University of Technology, China, in 2022. His research primarily focuses on fairness-aware recommender systems and large language model applications. He has published several papers in leading conferences and journals, including WWW, SIGIR, IEEE TKDE and ACM TOIS.

Miaomiao Cai is a Ph.D. candidate at Hefei University of Technology, China, where she also earned her Bachelor's degree in Engineering in 2020. Her research primarily focuses on debiasing techniques for recommender systems. She has published several papers in leading conferences and journals, including ACM KDD, ACM MM, and ACM TIST.

Le Wu is currently a professor at the Hefei University of Technology (HFUT), China. She received her PhD degree from the University of Science and Technology of China (USTC), China. Her general area of research interests are data mining, recommender systems, and responsible user modeling. She has published more than 60 papers in referred journals and conferences, such as IEEE TKDE, NIPS, SIGIR, WWW, and AAAI. Dr. Le Wu is the recipient of the Best of SDM 2015 Award, and the Distinguished Dissertation Award from the China Association for Artificial Intelligence (CAAI) 2017.

Yong Li received the B.S. degree from Huazhong University of Science and Technology in 2007, and the M. S. and the Ph. D. degrees in Electrical Engineering from Tsinghua University, in 2009 and 2012, respectively. Currently, he is a Faculty Member of the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of wireless networking, mobile computing and urban computing. Dr. Li has served as General Chair, TPC Chair, TPC Member for several international workshops and conferences, and he is on the editorial board of four international journals.

Meng Wang is a professor at the Hefei University of Technology, China. He received his BE degree and PhD degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), China in 2003 and 2008, respectively. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He is an associate editor of IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT), IEEE Transactions on Multimedia (IEEE TMM), and IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS).