ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems

Hong Qian
hqian@cs.ecnu.edu.cn
School of Computer Science and Technology, and Shanghai Institute of AI Education
East China Normal University
Shanghai, China

Shuo Liu
shuoliu@stu.ecnu.edu.cn
School of Computer Science and Technology
East China Normal University
Shanghai, China

Mingjia Li
52275901007@stu.ecnu.edu.cn
School of Computer Science and Technology, and Shanghai Institute of AI Education
East China Normal University
Shanghai, China

Bingdong Li
bdli@cs.ecnu.edu.cn
School of Computer Science and Technology, and Shanghai Institute of AI Education
East China Normal University
Shanghai, China

Zhi Liu
zhliu@cs.ecnu.edu.cn
School of Computer Science and Technology, and Shanghai Institute of AI Education
East China Normal University
Shanghai, China

Aimin Zhou*
amzhou@cs.ecnu.edu.cn
School of Computer Science and Technology, and Shanghai Institute of AI Education
East China Normal University
Shanghai, China

ABSTRACT

Cognitive diagnosis models (CDMs) are designed to learn students' mastery levels using their response logs. CDMs play a fundamental role in online education systems since they significantly influence downstream applications such as teachers' guidance and computerized adaptive testing. Despite the success achieved by existing CDMs, we find that they suffer from a thorny issue that the learned students' mastery levels are too similar. This issue, which we refer to as oversmoothing, could diminish the CDMs' effectiveness in downstream tasks. CDMs comprise two core parts: learning students' mastery levels and assessing mastery levels by fitting the response logs. This paper contends that the oversmoothing issue arises from that existing CDMs seldom utilize response signals on exercises in the learning part but only use them as labels in the assessing part. To this end, this paper proposes an oversmoothing-resistant cognitive diagnosis framework (ORCDF) to enhance existing CDMs by utilizing response signals in the learning part. Specifically, ORCDF introduces a novel response graph to inherently incorporate response signals as types of edges. Then, ORCDF designs a tailored response-aware graph convolution network (RGC) that effectively captures the crucial response signals within the response graph. Via ORCDF, existing CDMs are enhanced by replacing the input embeddings with the outcome of RGC, allowing for the consideration of response signals on exercises in the learning part. Extensive experiments on real-world datasets show that ORCDF not only helps existing CDMs alleviate the oversmoothing issue but also significantly enhances the models' prediction and interpretability performance. Moreover, the effectiveness of ORCDF is validated in the downstream task of computerized adaptive testing.

CCS CONCEPTS

• **Applied computing** → Education; • **Computing methodologies** → Machine learning.

KEYWORDS

Cognitive diagnosis, Oversmoothing, Representation, Student performance prediction, Online education systems

ACM Reference Format:
Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. 2024. 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 (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671988

1 INTRODUCTION

Cognitive diagnosis (CD) [19] serves as the foundational element in online intelligent education systems. It exerts an upstream and fundamental influence on subsequent modules such as computer adaptive testing [42], course recommendation [10, 35] and learning path suggestions [25, 26], among others. Specifically, as illustrated in Figure 1, CD aims to learn students' underlying mastery levels (Mas) by analyzing their historical response logs, thereby providing insights into various attributes of exercises, such as difficulty level (Diff) and discrimination (Disc). In recent years, an array of cognitive diagnosis models (CDMs) have emerged, prominently featuring frameworks such as item response theory (IRT) [8] and the neural cognitive diagnosis model (NCDM) [30]. The two core parts of CDM include learning students' Mas and assessing the learned Mas by fitting the response logs. The function used in the latter part is often

*Aimin Zhou is the corresponding author.

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08…$15.00
https://doi.org/10.1145/3637528.3671988

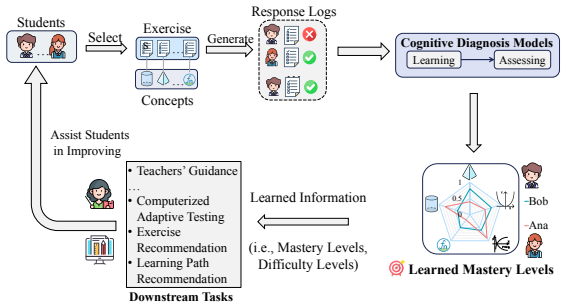


Figure 1: An example of CD, as well as relationships between CD and downstream tasks.

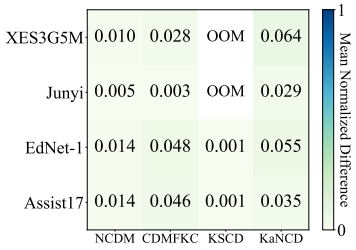


Figure 2: Results of motivation and pilot experiments: the oversmoothing issue in most existing CDMs is highlighted. The degree of oversmoothing is measured by the mean normalized difference (the lower the worse). OOM means the out-of-memory on an NVIDIA 3090 GPU. The vertical axis represents the names of four real-world datasets, and the horizontal axis lists the representative existing CDMs.

referred to as the interaction function (IF). IRT utilizes a latent factor to represent Mas and adopts the logistic function as IF. In contrast, NCDM replaces the traditional IFs with multi-layer perceptrons (MLP) and uses concept-specific vectors (i.e., set the embedding dimension being equal to the number of concepts) to characterize Mas. As embedding-based methods rapidly evolve and gain prominence, there is an increasing trend of representing both students and exercises in a vectorized form, and they are gradually refined by using a variety of advanced techniques [6, 12, 15, 21, 23, 31].

Despite the success, ***this paper, for the first time, identifies that existing CDMs share a potential and thorny issue that the learned Mas of students are too similar***. We refer to this issue as oversmoothing. Oversmoothing could diminish the CDMs' effectiveness in down-stream tasks. To support the motivation of this paper and reveal the oversmoothing issue, we conduct a pilot study on four real-world datasets collected from the online education systems, ensuring a diverse range of circumstances in the students' response logs. To characterize the degree of oversmoothing, inspired by [14], the mean normalized difference (MND) is proposed to measure the Mas learned by CDMs. Intuitively, the larger the MND value, the bigger the difference among students' Mas that learned by CDMs. Details of MND are elaborated in Section 5.1. As shown in Figure 2, although CDMs such as NCDM [30], CDMFKC [15], KSCD [21] and KaNCD [31] achieve commendable prediction performance, the MND values of Mas they have learned are quite small and hard to distinguish. Since CD is an upstream task, addressing this issue is urgent. For instance, if teachers rely on the outcomes of CD to assist student development, exceedingly subtle distinctions could lead to confusion. Intuitively, if MND is 0.005, it implies that the average difference in Mas for two students in a class on certain concepts is merely 0.005 (e.g., 0.51 and 0.515). Such a small margin could potentially bring difficulty to teachers to accurately assess the cognitive state of entire class. This not only fails to aid students but could also result in misguided instruction. Moreover, for downstream algorithms, a diagnosis result plagued by oversmoothing may lead to erroneous recommendations of learning materials, causing irreversible impacts on students.

One straightforward approach is to design a regularization term aimed at amplifying the differences between students. However, achieving a balance between the weight of this regularization term and the binary cross-entropy (BCE) loss during training is challenging. Besides, although this direct approach may help in mitigating the oversmoothing issue, it could compromise the model's prediction performance, since it forcefully amplifies the differences among all students and adversely affects the learning of students' Mas who should, in principle, be closely aligned. In this paper, we contend that the oversmoothing issue arises because existing CDMs seldom utilize response signals in the learning part but only use them as labels in the assessing part. For instance, students with right response on exercises with high difficulty levels should attain higher Mas on corresponding concepts in the learning part. Cooperating response signals in both learning and assessing parts of CDMs can widen the gap among students' Mas as they reserve the unique feature in students' response logs.

To this end, this paper proposes an oversmoothing-resistant cognitive diagnosis framework (ORCDF) to enhance existing CDMs by utilizing response signals in the learning part. Specifically, ORCDF introduces a novel response graph, which utilizes response logs and a $\mathbf{Q}$-matrix, inherently incorporating response signals as types of edges. Then, ORCDF designs a tailored response-aware graph convolution network (RGC) that effectively captures the crucial response signals within the response graph. We reveal that by utilizing the multiple layers of RGC, we achieve a multi-perspective analysis of student mastery. This is accomplished by combining the outcomes from multiple layers of RGC, leading to a more comprehensive understanding of student learning. Via ORCDF, existing CDMs are enhanced by replacing the input embeddings with the outcome of RGC through the transformation layer, allowing for the consideration of response signals on exercises in the learning part. Nevertheless, ORCDF encounters a new challenge: overemphasizing the role of response signals can exacerbate the guess and slip problem. This problem occurs when students guess in order to answer correctly or make mistakes on exercises they actually master, and could potentially lead models to make unreasonable inference of students' Mas. Different from previous methods that introduce extra parameters for guess and slip probabilities [16], this paper addresses the guess and slip problem in student-exercise interactions by considering them as noise edges in the response graph. Specifically, we flip the student-exercise edge in the response

graph with a flip ratio p_f (i.e., changing right to wrong, and wrong to right) in each epoch during the learning phase. We then design a loss function that ensures consistency in learning despite the presence of different noises, thereby mitigating the guess and slip problem. Extensive experiments show ORCDF's superiority over state-of-the-art methods in terms of resisting oversmoothing, enhancing prediction performance, and improving interpretability. Finally, we validate the efficacy of ORCDF in downstream tasks.

The subsequent sections respectively recap the related work, present the preliminaries, introduce the proposed ORCDF, show the empirical analysis and finally conclude the paper.

2 RELATED WORK

Cognitive Diagnosis Models. CDMs involve various approaches, such as latent factor models like IRT and MIRT (multidimensional IRT), or concept mastery pattern models like the deterministic input, noisy and gate (DINA) model, to infer students' mastery levels. DINA, a classic CDM, employs binary variables to represent mastery levels where 0 means unmastered and 1 means mastered. However, recent advances in deep learning have led to significant improvements in handling large-scale interactions. Notably, NCDM uses MLP as its IF, treating mastery patterns as continuous variables ranging from 0 to 1. This evolution in approach has been paralleled by diverse methods in analyzing response logs, including MLP based [13, 15, 21], graph attention networks [28] and Bayesian networks [12, 33], each contributing to a more nuanced understanding of student learning patterns. However, as depicted in Figure 2, these advanced CDMs encounter the oversmoothing issue which could potentially hinder the application of CD in downstream tasks of intelligent education, affecting their performance and consequently impacting student learning. To the best of our knowledge, the oversmoothing issue in the field of CD remains unexplored.

Oversmoothing Issue. The oversmoothing issue [14] is a significant problem in graph representation learning (GRL). Many studies have shown that the layers of graph neural network (GNN) deepen, the representations of graph nodes become increasingly smooth, leading to a substantial decrease in accuracy. This has prompted numerous researchers to employ a variety of methods [22] to address this issue, enabling deeper GNN architectures. The same phenomenon is also observed in various fields where graphs are used for data mining. For example, in recommendation systems, graph collaborative filtering (GCF) [34] faces the oversmoothing problem, which arises for the same reasons as in GRL due to the stacking of GNN layers. However, in the context of CD, oversmoothing is not a result of stacking GNN layers, since most CDMs like NCDM, CDMFKC, KSCD and KaNCD do not utilize GNN. Yet, this issue does exist and is urgent, as shown in Figure 2. Thus, existing solutions to addressing oversmoothing in GRL and GCF are not suitable to resolve the oversmoothing issue in CD.

3 PRELIMINARIES

This section first introduces the fundamental elements of CD and then introduces the formal problem definition of CD and oversmoothing issue in CDMs. We also give abbreviations for terms in Table 5 at the beginning of the Appendix.

Cognitive Diagnosis Basis. Consider an education scenario which contains three sets: $S = \{s_1, \ldots, s_N\}$, $E = \{e_1, \ldots, e_M\}$, and $C = \{c_1, \ldots, c_Z\}$. They symbolize students, exercises and knowledge concepts, with respective sizes of N, M and Z. Q represents the relationship between exercises and knowledge concepts, which can be regarded as a binary matrix $\mathbf{Q} = (\mathbf{Q}_{iz})_{M\times Z}$, where $\mathbf{Q}_{iz} \in \{0, 1\}$ means whether e_i relates to c_z or not. Students from set S, driven by unique interests and requirements, select exercises from E. The results are documented as response logs. Specifically, these logs can be illustrated as triplets $T = \{(s, e, r) \mid s \in S, e \in E, r_{se} \in \{0, 1\}\}$. $r_{se} = 1$ represents correct and $r_{se} = 0$ represents wrong. In this paper, we treat response logs as interaction matrix $\mathbf{I} \in \mathbb{R}^{N\times M}$. It contains three categorical values (1 means right, 0 means no interaction and -1 means wrong). Finally, we give the formal definition of the CD task and oversmoothing issue in CDMs.

Definition 3.1 (Problem Definition). Given interaction matrix $\mathbf{I} \in \mathbb{R}^{N\times M}$*, a binary matrix* $\mathbf{Q} \in \mathbb{R}^{M\times Z}$*, the goal of cognitive diagnosis is to infer* $\mathbf{Mas} \in \mathbb{R}^{N\times Z}$*, which denotes the latent mastery level of students on each concept.*

Definition 3.2 (Oversmoothing in CDMs). Given the learned $\mathbf{Mas} \in \mathbb{R}^{N\times Z}$ *by CDMs, if the difference in students'* **Mas** *is sufficiently small, it indicates the presence of oversmoothing issue in CDMs.*

In this paper, we utilize the mean normalized difference proposed in Section 5.1 to quantify the degree of oversmoothing.

4 METHODOLOGY: THE PROPOSED ORCDF

This section introduces the proposed ORCDF. It starts by introducing the proposed novel response graph, then explores the response-aware graph convolution (RGC), a technique designed to capture the rich information embedded in the response graph. Following this, we introduce a consistency regularization loss function. We also discuss the model training and analyze model complexity. An overview of ORCDF is shown in Figure 3.

Response Graph. As illustrated in Figure 4(a), focusing on responses, the response graph (ResG), denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, comprises three types of nodes and edges. $\mathcal{V} = S \cup E \cup C$ involves students, exercises, and concepts, $\mathcal{E}$ involves interactions between S and E (i.e., "Right"), S and E (i.e., "Wrong"), E and C (i.e., "Related"). Notably, we incorporate the response signal on exercises as the edge types between students' nodes and exercises' nodes. Next, we will introduce how to capture the fruitful response signal information.

4.1 Response-aware Graph Convolution

Construct Embeddings. In CD, the primary data elements are response logs and the $\mathbf{Q}$. It is crucial to deconstruct these complex logs into their fundamental components: students, exercises, and concepts. We encode them with trainable embeddings $\mathbf{H}_s \in \mathbb{R}^{N\times d}, \mathbf{H}_e \in \mathbb{R}^{M\times d}, \mathbf{H}_c \in \mathbb{R}^{Z\times d}$. For instance, $\mathbf{h}_{s_i} \in \mathbf{R}^{1\times d}$ denotes the row vector of the i-th student. To facilitate subsequent convolution processes, we stack the aforementioned embeddings to form $\mathbf{H}^{(0)} \in \mathbb{R}^{(N+M+Z)\times d}$.

Right-Wrong Decomposition. In the ResG, there are two types of response signals existing between student nodes and exercise nodes, as shown in Figure 4(a). To better explore the impact of different response signals on learning Mas, we intuitively decompose

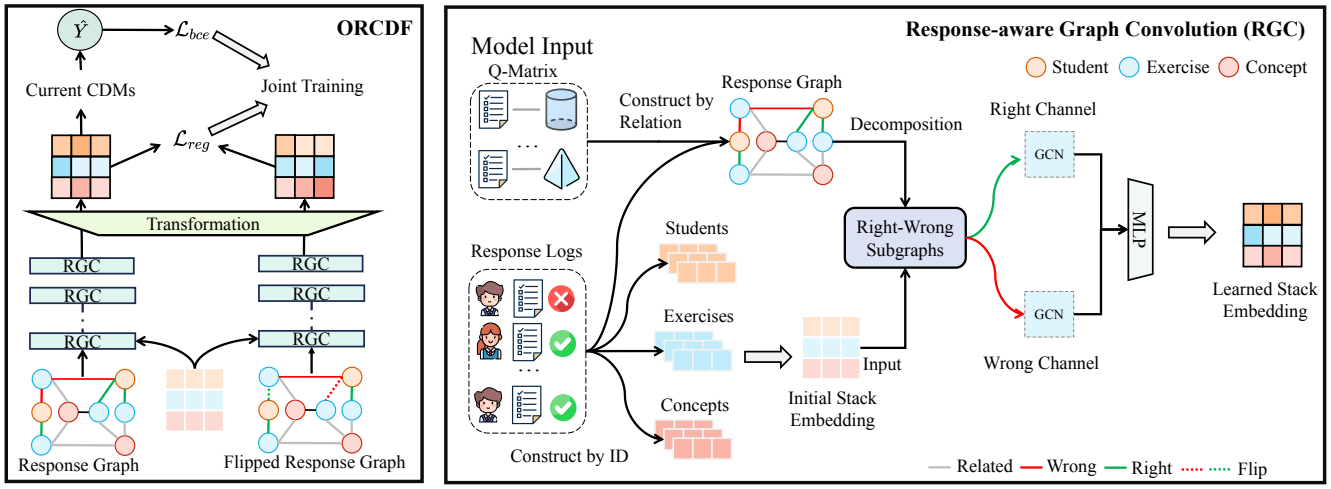


Figure 3: The left side provides an overview of the proposed ORCDF. The right side details the main component of ORCDF.

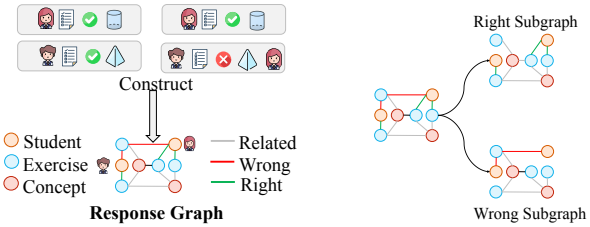


Figure 4: (a) The proposed response graph. (b) Right-wrong decomposition.

the response graph into a right subgraph and a wrong subgraph. From the perspective of adjacency matrix, this involves splitting the interaction matrix $\mathbf{I}$ into $\mathbf{I}_{\text{right}}$ (1 represents right, 0 represents others) and $\mathbf{I}_{\text{wrong}}$ (1 represents wrong, 0 represents others). *For brevity, in the following sections, we will denote "R" for right and "W" for wrong.* Then we construct the right and wrong subgraphs (i.e, $\mathbf{A}_{\text{R}}, \mathbf{A}_{\text{W}}$ as expressed by Eq. (1):

$$\mathbf{A}_{\text{R}} = \begin{pmatrix} \mathbf{O} & \mathbf{I}_{\text{R}} & \mathbf{O} \\ \mathbf{I}_{\text{R}}^{\top} & \mathbf{O} & \mathbf{Q} \\ \mathbf{O} & \mathbf{Q}^{\top} & \mathbf{O} \end{pmatrix}, \quad \mathbf{A}_{\text{W}} = \begin{pmatrix} \mathbf{O} & \mathbf{I}_{\text{W}} & \mathbf{O} \\ \mathbf{I}_{\text{W}}^{\top} & \mathbf{O} & \mathbf{Q} \\ \mathbf{O} & \mathbf{Q}^{\top} & \mathbf{O} \end{pmatrix}. \tag{1}$$

In the ResG, the neighbors of a specific exercise node may include students who either answer the exercise right or wrong. However, after disentangling such response signals in the ResG, in each subgraph, the neighbors of a particular exercise node will only consist of students who displayed the same response signals. For example, if both s_1 and s_2 are connected to e_1, indicating that they both answer e_1 correctly, there may be some shared information explaining why they both got it right. Such crucial response signals will be propagated during the message-passing mechanism by GCN. This process enables a deeper understanding of the nuances in student responses. In the following part, we will introduce a novel graph convolution approach tailored to capture the information from the two disentangled subgraphs in CD.

Embedding Propagation. Considering that in CD, the features of students, exercises, and knowledge concepts are quite simple, consisting only of IDs, we draw inspiration from [7]. As a result, we eliminate linear transformations and nonlinear activation functions, opting to use only the fundamental components of GCN. Hence, the graph embedding propagation layer is designed with the following matrix form

$$\mathbf{H}^{(l)} = \hat{\mathbf{A}}\mathbf{H}^{(l-1)}, \quad \hat{\mathbf{A}} = \left(\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\right), \tag{2}$$

where $\mathbf{A}$ can be $\mathbf{A}_{\text{R}}$ or $\mathbf{A}_{\text{W}}$. The degree matrix $\mathbf{D}$ is a diagonal matrix with size $(N+M+Z)\times(N+M+Z)$, where each entry $\mathbf{D}_{ii}$ representing the number of non-zero entries in the i-th row vector of the matrix $\mathbf{A}$. Using Eq. (2), we can obtain the convolution outcomes from the l-th layer of the disentangled subgraphs, specifically $\mathbf{H}_{\text{R}}^{(l)}$ and $\mathbf{H}_{\text{W}}^{(l)}$. However, right and wrong represent completely opposite response signals, it may be inappropriate to directly plus the results obtained from convolutions performed on the two subgraphs. A sophisticated function capable of aggregating these two types of information is necessary, as the interaction mechanisms between students and exercises are quite intricate. It can be expressed as

$$\mathbf{H}_{F}^{(l)} = \phi(\mathbf{H}_{\text{R}}^{(l)}\mathbf{W}_{\text{rc}} + \mathbf{H}_{\text{W}}^{(l)}\mathbf{W}_{\text{wc}}), \tag{3}$$

where $\mathbf{H}_{F}^{(l)}$ denotes the final representation of the l-th RGC layer and ϕ denotes arbitrary nonlinear activate function. $\mathbf{W}_{\text{rc}}, \mathbf{W}_{\text{wc}} \in \mathbb{R}^{d\times d}$ are trainable parameters. Intuitively, $\mathbf{H}_{\text{R}}^{(l)}\mathbf{W}_{\text{rc}}$ denotes the right channel which obtain the semantic information from right signal. Conversely, $\mathbf{H}_{\text{W}}^{(l)}\mathbf{W}_{\text{wc}}$ represents the opposite. The ultimate embedding $\mathbf{H}_F$ is calculated using a mean pooling operation on the outcomes from each layer of the RGC which can be expressed as

$$\mathbf{H} = \frac{1}{1+L}(\mathbf{H}_{F}^{(0)} + \mathbf{H}_{F}^{(1)} + \ldots + \mathbf{H}_{F}^{(L)}). \tag{4}$$

Discussion. Here, we explain why RGC can alleviate the oversmoothing issue in existing CDMs. Notably, since our goal is to alleviate the oversmoothing issue in CDMs, and given that shallow layers of RGC already achieve satisfactory experimental results, the

oversmoothing issue caused by deep GNN is not addressed in this work and can be considered for future research. First, we analyse the NCDM which directly set Mas as $\mathbf{H}^{(0)} \in \mathbb{R}^{N\times d}$ and $d = Z$. Firstly, the MND of any two students s_1 and s_2 in original NCDM is calculated as

$$\mathrm{MND}_{s_1,s_2} = \|\mathbf{Mas}_{s_1} - \mathbf{Mas}_{s_2}\|_2^2 = \|\mathbf{H}_{s_1}^{(0)} - \mathbf{H}_{s_2}^{(0)}\|_2^2 . \tag{5}$$

Clearly, the difference between the learned Mas of s_1 and s_2 in NCDM reflects the disparity in their individual information. Here, we will use a one-layer RGC incorporated with NCDM as an example. For brevity, we will omit all normalization coefficients, biases and retain only the key components. The MND of students s_1 and s_2 after utilizing RGC is calculated as

$$\mathrm{MND}_{s_1,s_2} = \|\mathbf{Mas}_{s_1} - \mathbf{Mas}_{s_2}\|_2^2 = \|\mathbf{H}_{s_1} - \mathbf{H}_{s_2}\|_2^2 . \tag{6}$$

Via Eq. (2), we can derive that $\mathbf{H}_{s_1}^{(1)}(\mathrm{R}) = \sum_{e_j\in\mathcal{N}^{\mathrm{R}}(s_1)} \mathbf{H}_{e_j}\mathbf{W}_{\mathrm{rc}}$ where $e_j \in \mathcal{N}^{\mathrm{R}}(s_1)$ represents the j-th exercise s_1 practiced correctly and is also the neighbor of s_1 in the right subgraph. Consequently, the term $\mathbf{H}_{s_1}^{(1)}(\mathrm{R})$ represents the normalized summation exercises where s_1 practiced correctly. Similarly, we can derive $\mathbf{H}_{s_2}^{(1)}(\mathrm{R}), \mathbf{H}_{s_1}^{(1)}(\mathrm{W})$ and $\mathbf{H}_{s_2}^{(1)}(\mathrm{W})$ following the same logic. Finally, Via Eq. (2) and Eq. (4), the $\|\mathbf{H}_{s_1} - \mathbf{H}_{s_2}\|_2^2$ can be calculated as $\frac{1}{2}(\|\mathbf{H}_{s_1}^{(0)} - \mathbf{H}_{s_2}^{(0)} + \mathbf{H}_{s_1}^{(1)}(\mathrm{R}) - \mathbf{H}_{s_2}^{(1)}(\mathrm{R}) + \mathbf{H}_{s_1}^{(1)}(\mathrm{W}) - \mathbf{H}_{s_2}^{(1)}(\mathrm{W})\|_2^2)$. Consequently, we can have the following observations:

• The first term $\mathbf{H}_{s_1}^{(0)} - \mathbf{H}_{s_2}^{(0)}$ is the same as Eq. (5) which reflects the disparity of individual information of s_1 and s_2.

• The second term $\mathbf{H}_{s_1}^{(1)}(\mathrm{R}) - \mathbf{H}_{s_2}^{(1)}(\mathrm{R}) + \mathbf{H}_{s_1}^{(1)}(\mathrm{W}) - \mathbf{H}_{s_2}^{(1)}(\mathrm{W})$ captures the difference in the exercises that students s_1 and s_2 practiced correctly and incorrectly.

• The final MND_{s_1,s_2} of RA-NCDM is the mean of the first and second terms which can be interpreted as a comparison of the differences between students from the aforementioned perspectives.

For instance, if both s_1 and s_2 have similar accuracy in their exercises, the first term will be quite small due to the monotonicity assumption in CD. However, if the exercises attempted by s_1 are more challenging compared to those of s_2, the second term will capture this disparity and consequently increase the final difference between s_1 and s_2. This suggests that the RGC is capable of capturing the differences in the exercises practiced by students, resulting in more distinctive Mas for each student.

Notably, as the number of RGC layers increases, the perspectives for considering student differences also multiply. For instance, a two-layer RGC would further compare the differences with other students who have similar exercise performance as the current student. Therefore, by incorporating RGC, CDMs can assess student differences from multiple angles, thereby mitigating the oversmoothing issue. We will validate this conclusion in our ablation study in Section 5.3 and give visualizations of the learned Mas by T-SNE [29] in Section 5.4.

4.2 Consistency Regularization Loss

After the graph convolution by multiple RGC layers, we can get the final representation $\mathbf{H}$ via Eq. (4). However, as we disentangle the response signal and capture student differences from various perspectives, it may exacerbate the notorious impact of the guess and slip problem on CDMs [16, 36]. Previous methods, as referenced in [4], model the guess and slip probabilities for each exercise as fixed parameters. Evidently, this approach is somewhat brute-force and might overlook the individual impact of students. This is because the probability of guessing or slipping is likely to vary for each person across different exercises. Contrary to the aforementioned methods, in this paper, we treat guess and slip as noise edges within the ResG. Specifically, we flip the student-exercise edge type (i.e., from R to W or W to R) with a probability p_f in the ResG. This noised version of the ResG, where some edges are flipped, is referred to as the flipped ResG, as illustrated in the left part of Figure 3. We aim for the representations derived from the original ResG and the flipped ResG to be similar, in order to ensure that the CDMs remain effective even when subject to the disturbances caused by guess and slip problem. It can be formulated as

$$\mathcal{L}_{\mathrm{reg}} = -\sum_{s_a\in S} \log\left(\exp\left(\mathbf{h}'_{s_a}\mathbf{h}_{s_a}^{\top}/\tau\right)\right), \tag{7}$$

where $\mathbf{h}'_{s_a}$ is the representation derived from flipped ResG, and $\mathbf{h}'_{s_a}\mathbf{h}_{s_a}^{T}$ denotes the similarity score the representation derived from the ResG and flipped ResG. τ is the hyperparameter which controls the degree of smoothness utilized in various methods [38–40].

4.3 Model Training

Given input embeddings, existing CDMs predict the performance of students practicing exercises, which can be formulated as

$$\hat{y}_{ij} = \mathcal{M}_{CD}(\mathbf{H}_{s_i}, \mathbf{H}_{e_j}, \mathbf{H}_c), \tag{8}$$

where $\mathcal{M}_{CD}(\cdot)$ denotes the CDMs, and $\mathbf{H}$ represents the input embedding that contains the representation of the student, exercises and concepts.

Transformation Layer. To facilitate the integration of ORCDF with the majority of existing CDMs, we need to transform dimensions to suit the specific type of CDM in use. If the embedding size of CDMs is a latent dimension (e.g., KaNCD), we directly utilize $\mathbf{H}$ as the input embedding for incorporated CDMs. Otherwise (e.g., NCDM), we introduce a transformation layer which can be formulated as

$$\mathbf{H}_{\mathrm{t}} = \mathbf{H}\mathbf{W}_{\mathrm{t}} + \mathbf{b}_{\mathrm{t}}, \tag{9}$$

where $\mathbf{H}_{\mathrm{t}}$ will be employed as input embedding for incorporated CDMs and $\mathbf{W}_{\mathrm{t}} \in \mathbb{R}^{d\times Z}$, $\mathbf{b}_{\mathrm{t}} \in \mathbb{R}^{(N+M+Z)\times 1}$ are trainable parameters. As a result, unlike the previous RCD which sets $d = Z$, we can choose d as a latent dimension (e.g., 64). This significantly reduces the time complexity of graph convolution, a point that will be further analyzed in the subsequent subsection.

Joint Training. The primary loss employed in CD task is to calculate the BCE loss between the model's predictions and the true response scores in a mini-batch. The aforementioned consistency regularization loss is incorporated jointly optimized with the CD task. The overall loss can be expressed as

$$\mathcal{L}_{\mathrm{BCE}} = -\sum_{(s,e,r_{se})\in T} \left[r_{se}\log\hat{y}_{se} + (1-r_{se})\log(1-\hat{y}_{se})\right], \tag{10}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}} . \tag{11}$$

λ_{reg} is a hyperparameter that governs the relative importance of the consistency regularization loss.

4.4 Model Complexity Analysis

Theoretically, we reveal that the graph convolution in ORCDF takes $O(4|\mathcal{E}|Ld)$ time complexity. L denotes the number of RGC layers, and d denotes the dimension of embeddings. By leveraging the lightweight backbone and the transformation, our method is significantly lower in time complexity compared with the recent GNN-based approach RCD [6]. Specifically, RCD has the complexity of $O(2|\mathcal{E}|LZ^2)$, where Z represents the number of concepts ($d \ll Z$). It suggests that ORCDF is more suitable for current online education scenario on ground of the increasing granularity of knowledge concepts. Indeed, ORCDF showcases a notable speed advantage, being up to ***18 times*** faster than RCD on the Assist17 dataset (i.e., $Z = 102$). This dataset is collected from ASSISTment online tutoring systems and extensively utilized CDMs [12]. This improvement comes along with enhanced performance and lower GPU memory usage. For detailed information, please refer to Appendix A.

5 EXPERIMENTS

In this section, we first describe four real-world datasets and evaluation metrics. Then, through extensive experiments, we aim to verify the superiority of ORCDF, which not only assists existing CDMs in mitigating the oversmoothing issue but also enhances the models' prediction performance and interpretability performance. To ensure the reliability and reproducibility of our experiments, they are independently repeated ten times with different seeds and our code is available at https://github.com/lswhim/ORCDF.

5.1 Experimental Settings

Datasets Description. The experiments are conducted using four real-world datasets: Assist17, EdNet-1, Junyi, and XES3G5M. The Assist17 dataset is provided by the ASSISTment web-based online tutoring systems [5] and are widely used for cognitive diagnosis tasks [30]. EdNet-1 [3] is the dataset of all student-system interaction collected over 2 years by Santa, a multi-platform AI web-based tutoring service with more than 780K users in Korea. Junyi [2] is an online math practice log dataset offered by Junyi Academy. XES3G5M [20] is a knowledge tracing benchmark dataset with auxiliary information. For more detailed statistics on these four datasets, please cf. Table 1. Notably, "Sparsity" refers to the sparsity of the dataset, which is calculated as $|T|/(|S||E|)$. "Average Correct Rate" represents the average score of students on exercises, and "Q Density" indicates the average number of concepts per exercise.

Evaluation Metrics. To assess the efficacy of ORCDF, we utilize both score prediction, interpretability and oversmoothing metrics.

Table 1: Statistics of real-world datasets for experiments.

Datasets	Assist17	EdNet-1	Junyi	XES3G5M
#Students	1709	1776	10000	4000
#Exercises	3162	11925	734	7191
#Knowledge Concepts	102	189	734	832
#Response Logs	390,311	616,193	408,057	1,174,514
Sparsity	0.072	0.029	0.055	0.04
Average Correct Rate	0.815	0.662	0.687	0.799
Q Density	1.22	2.25	1.0	1.16

• **Score Prediction Metrics:** Assessing the effectiveness of CDMs poses difficulties owing to the absence of the true Mas. A common approach to address this challenge is to learn the Mas within the train data and then evaluate the models based on their learned Mas to predict students' performance on exercises in the test data. In line with prior CDM studies, we partition the response logs of students into train, valid and test data with 7:1:2 following the previous researches [30] and assess CDMs' performance on the test data using classification metrics such as Area Under the ROC Curve (AUC), Accuracy (ACC). Crucially, we build the ResG solely based on the train data.

• **Interpretability Metric:** Diagnostic results are highly interpretable hold significant importance in CD. In this regard, we employ the degree of agreement (DOA), which is consistent with the approach used in [17, 18, 24]. The underlying intuition here is that, if s_a has a greater accuracy in answering exercises related to c_k than student s_b, then the probability of s_a getting c_k should be greater than that of s_b. Namely, $\mathbf{Mas}_{s_a,c_k} > \mathbf{Mas}_{s_b,c_k}$. Details about DOA can be found in Appendix B. Consistent with [12], we compute the average DOA for the top 10 concepts with the highest number of response logs in Assist17, EdNet-1, Junyi and XES3G5M.

• **Oversmoothing Metric:** We employ the proposed MND to measure the Mas learned by CDMs. In CD, since the Mas of students learned by CDMs with concept mastery pattern lies within the range of 0 to 1, we utilize the l_2 norm of the difference between two students' mastery level vectors to describe the disparity between them. It can be formulated as follows:

$$\text{MND} = \frac{1}{|S|}\frac{1}{|S|-1}\sum_{s_u\in S}\sum_{s_v\in S}\frac{\left\|\mathbf{Mas}_{s_u}-\mathbf{Mas}_{s_v}\right\|_2^2}{|C|}, \tag{12}$$

where S, C represent the set of students and knowledge concepts, respectively, and $\mathbf{Mas}_{s_u}$ stands for the learned Mas of student s_u by CDMs. A larger MND value indicates greater difference in the Mas that learned by CDMs, implying that the oversmoothing issue is more adequately addressed.

Implementation Details. For parameter initialization, we employ the Xavier [7], and for optimization purposes, Adam [11] is adopted. For fair comparison, the embedding size is uniformly set to 32 for MIRT, KaNCD, and KSCD, and to Z for NCDM and CDMFKC. The batch size is set as 4096 for all datasets. To regulate the impact of the regularization term, we adjust the flip ratio p_f within the range $\{0.05, 0.15, 0.1, 0.2\}$, λ_{reg} within the range $\{10^{-4}, 10^{-3}, \ldots, 10^{-1}\}$, τ within the range $\{0.1, 0.5, 1.0, 3.0, 5.0\}$. Analysis regarding the aforementioned hyperparameters can be found in Section 5.6 and Appendix C.

5.2 Student Performance Prediction

To showcase the effectiveness of ORCDF, we integrate it with various CDMs, as described in the subsequent part.

• IRT [8] is a classic model of latent factor CDMs, which uses one dimension θ to model Mas and utilize logistic function as IF to predict the student score performance.

• MIRT [27] is a representative model of latent factor CDM, which uses multidimensional $\boldsymbol{\theta}$ to model Mas.

• NCDM [30] is the first recent deep-learning based CDM which utilizes MLP to replace the traditional manually designed IFs.

• CDMFKC [15] employs a sophisticatedly designed neural network to model the impact of knowledge concepts on students' score performance.

• KSCD [21] also delves into the implicit relationships among knowledge concepts and employs a concept-augmented IF.

• KaNCD [31] is an enhanced version of NCDM, delving into the implicit relationships among concepts to tackle the knowledge coverage issue.

Details. To ensure fairness in comparison, we adhere to the hyperparameter settings and IFs as specified in their original publications. IRT and MIRT are non-interpretable models, namely latent factor CDMs, the Mas they learn cannot be correlated directly with specific knowledge concepts. Therefore, they are not suitable for calculating DOA and MND. In Table 2, we use "-" to indicate this inapplicability. If CDMs signify out-of-memory on an NVIDIA 3090 GPU, we use the term "OOM" to denote this occurrence.

Experimental Results. The main observations are as follows. As shown in Table 2, the proposed ORCDF consistently and substantially improves the MND values of all the base CDMs across all datasets. This confirms the efficacy of ORCDF in alleviating the oversmoothing problem. We will validate the improvement of downstream tasks resulting from mitigating the oversmoothing issue in Section 5.5. Besides, from Table 2, ORCDF significantly benefits the base CDMs. It substantially enhances both the prediction performances of all the base CDMs across various datasets. It is validated that the ORCDF effectively alleviates the oversmoothing issue without compromising the prediction performance of the CDMs. Notably, DOAs of CDMs have improved substantially. This suggests that learning Mas from multiple perspectives aligns more closely with the monotonicity assumption prevalent in educational measurement, indicating enhanced interpretability performance.

Furthermore, we conduct comparisons with other competitive frameworks including other graph-based methods or related approaches to further validate the effectiveness of the ORCDF. The results are in Figure 5.

• RCD [6] is the first method to employ GAT in addressing tasks within the field of CD. It uses standard GAT to delve into the intricate relationships among students, exercises, and concepts.

• LightGCN [9] is a recent classic model that employs GCN in collaborative filtering. Since LightGCN is a lightweight graph neural network suitable for heterogeneous graphs with solely ID as features, we chose it as the representative baseline.

• HierCDF [12] utilizes the Bayesian network to model the mastery pattern with directed acyclic graph of knowledge concepts. For a fair comparison, we integrate the aforementioned methods with the NCDM and conduct the experiments on Assist17 and Junyi under the same settings as previously described. Since RCD has already shown superiority over some heterogeneous graph representation learning methods such as HetG [41], and HAN [32], we do not include these methods in our comparative analysis. Since the RCD experiences OOM issue on Junyi, we do not report the results for this dataset. As shown in Figure 5, ORCDF outperforms other chosen frameworks, whether they are specifically tailored for CD or other fields, in the task of predicting student performance. The superiority of ORCDF in terms of MND further confirms that learning Mas from multiple perspectives is beneficial for alleviating the oversmoothing issue.

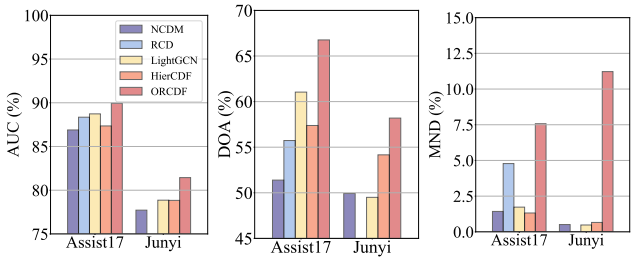


Figure 5: Comparison with other related frameworks incorporating NCDM.

5.3 Ablation Study

In this subsection, we scrutinize and evaluate each key individual component of ORCDF to comprehend their respective impacts and significance on the overall performance of the model. The ablation analysis is conducted using the following three versions.

• OR-w/o-rgc: This ablation of ORCDF does not integrate the response-aware graph convolution. Instead, it directly perform convolution on the entire response graph without decomposition.

• OR-w/o-reg: This ablation of ORCDF does not utilize the proposed consistency regularization loss $\mathcal{L}_{\text{reg}}$.

• OL: It represents the base CDMs, which can be considered as the one without the inclusion of response-aware graph convolution and consistency regularization loss.

Due to space constraints, we only present the ablation study using OR-NCDM as an example. This choice is motivated by the fact that NCDM is often employed as a classic CDM in downstream tasks. It is worth noting that the results from incorporating other CDMs are generally similar.

Experimental Results. As indicated in Table 3, the proposed method outperforms the other two versions, suggesting that each component plays a significant role in enhancing the model's overall effectiveness. OR-w/o-rgc performs significantly worse, further validating the superiority of the proposed RGC in capturing the information within the response graph. We empirically find that although the consistency regularization loss is designed to alleviate the guess and slip problem, it not only improves the prediction and interpretability performance but also achieves a higher MND than the original version. This indicates that the guess and slip problem indeed exists in real-world scenarios, and addressing this problem is crucial for the effectiveness of CDMs.

5.4 In-Depth Analysis of ORCDF's Advantages

In this subsection, we analyze the proposed ORCDF from two perspectives: generalization performance and robustness performance.

Generalization Performance. To assess the efficacy of ORCDF in addressing the generalization issue, we conduct experiments on three datasets with varying test ratios $p_t = \{10\%, 20\%, 30\%, 40\%, 50\%\}$. As p_t increases which is consistent with [6], the generalization ability of CDMs is tested more stringently. As depicted in Figure 6 of Appendix B, with an increasing test ratio p_t, the number of response logs used for training decreases. However, OR-NCDM consistently outperforms NCDM, illustrating that ORCDF can provide more accurate diagnosis results with fewer student response records. This

Table 2: Overall student score prediction performance. "OL" stands for "original", referring to the original method, and "OR" denotes the proposed ORCDF enhancement applied to the original method. Within each method, the entry that exhibits the highest mean value is highlighted in bold. The standard deviation is not shown in the table since it is very small (less than 0.01). If the mean value significantly differs from the original method, passing a t-test with a significance level of 0.01, then we denote it with "*" at the corresponding position. "-" indicates that the model is not suitable of calculating this metric. "OOM" signifies out-of-memory occurring on a single NVIDIA 3090 GPU. All metrics are ideally larger for better results.

Dataset	Metric (%)	IRT		MIRT		NCDM		CDMFKC		KSCD		KANCD	
		OL	OR	OL	OR	OL	OR	OL	OR	OL	OR	OL	OR
Assist17	AUC	88.95	**89.60***	91.42	**91.95***	86.89	**89.94***	87.30	**90.02***	88.56	**89.68***	88.56	**90.33***
	ACC	86.11	**86.75***	88.15	**88.51***	84.56	**87.10***	85.15	**87.2***	86.14	**86.75***	86.06	**87.56***
	DOA	-	-	-	-	51.39	**66.76***	54.69	**66.67***	65.86	**68.05***	62.86	**67.01***
	MND	-	-	-	-	1.43	**7.57***	4.64	**20.7***	0.05	**2.21***	3.51	**14.08***
EdNet-1	AUC	73.18	**74.56***	74.41	**74.68***	72.86	**74.81***	73.05	**74.85***	73.74	**74.66***	74.42	**75.11***
	ACC	70.89	**71.85***	71.70	**71.89***	70.68	**71.98***	70.79	**71.95***	71.42	**71.85***	71.75	**72.07***
	DOA	-	-	-	-	59.31	**64.29***	60.45	**64.01***	64.55	**65.07***	63.02	**65.47***
	MND	-	-	-	-	1.42	**4.29***	0.82	**4.05***	0.05	**2.45***	5.48	**7.12***
Junyi	AUC	80.35	**81.46***	80.87	**81.46***	77.72	**81.44***	78.27	**81.30***	OOM		79.12	**81.72***
	ACC	76.65	**77.52***	77.28	**77.54***	74.49	**77.59***	74.95	**77.28***			75.57	**77.71***
	DOA	-	-	-	-	49.92	**58.19***	49.92	**60.74***			53.59	**60.85***
	MND	-	-	-	-	0.51	**11.22***	0.34	**17.18***			2.86	**12.82***
XES3G5M	AUC	79.18	**80.13***	80.43	**80.66***	75.46	**80.22***	74.15	**79.98***	OOM		79.68	**80.41***
	ACC	81.52	**82.51***	82.31	**82.52***	81.21	**82.49***	80.17	**82.28***			82.23	**82.44***
	DOA	-	-	-	-	68.01	**73.93***	69.03	**73.89***			73.50	**73.62***
	MND	-	-	-	-	1.04	**19.37***	2.83	**35.26***			6.43	**16.67***

Table 3: Ablation study of ORCDF. Details are as same as Table 2.

Dataset	Metric	NCDM			
		OL	OR-w/o-rgc	OR-w/o-reg	OR
Assist17	AUC	86.89	88.73	89.91	**89.94**
	ACC	84.56	86.19	87.07	**87.10**
	DOA	51.39	63.74	65.26	**66.76**
	MND	1.43	2.53	6.90	**7.57**
EdNet-1	AUC	72.86	74.77	74.76	**74.81**
	ACC	70.86	71.94	71.86	**71.98**
	DOA	59.31	63.73	64.23	**64.29**
	MND	1.42	2.26	3.35	**4.29**
Junyi	AUC	77.72	80.23	81.14	**81.44**
	ACC	74.49	76.52	77.22	**77.59**
	DOA	49.92	57.96	58.14	**58.19**
	MND	0.51	4.96	7.96	**11.22**
XES3G5M	AUC	75.46	80.24	80.22	**80.32**
	ACC	81.21	82.46	82.46	**82.49**
	DOA	68.01	73.45	73.93	**73.94**
	MND	1.04	5.79	10.71	**19.37**

is particularly suitable for current online learning scenarios, where students often have limited response logs.

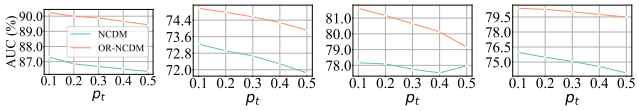


Figure 6: Performance under different p_t on four datasets.

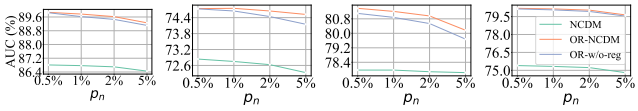
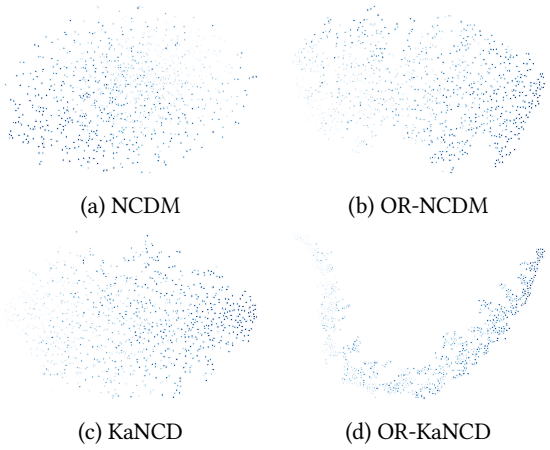


Figure 7: Performance under different p_n on four datasets.

Robustness Performance. Guess and slip problem [16, 37] in CD can significantly affect the accurate determination of students' Mas. These noise interactions typically stem from two main factors: guess and slip. To showcase the capacity of our proposed method, ORCDF, in mitigating the issue of guess and slip problem, we can conceptually introduce noise into the training datasets while keeping the test dataset unchanged. Specifically, to inject noise into the train datasets, we can randomly select student responses and flip them to the opposite. For example, correct responses can be flipped to incorrect ($1 \rightarrow 0$) and vice versa ($0 \rightarrow 1$) at a certain noise ratio, represented as p_n. As illustrated in Figure 7 of Appendix B, as the noise ratio p_n increases, the fact that OR-NCDM outperforms NCDM indicates its effectiveness in giving reasonable diagnosis result, especially when there is noises in the students' response logs. Notably, OR-NCDM shows a lesser performance drop than OR-w/o-rgc as the noise ratio increases, validating the effectiveness of our proposed loss function.

The Distrubution of Students' Mas. Indeed, students can naturally be grouped into categories based on their performance, such as those with low and high correct rates. This classification reflects intrinsic differences in their Mas. We employ t-SNE [29], a renowned dimensionality reduction method, to map the **Mas** onto a two-dimensional plane. By shading the scatter plot according



(a) NCDM (b) OR-NCDM

(c) KaNCD (d) OR-KaNCD

Figure 8: Visualizations of the learned Mas on the EdNet-1 dataset.

Table 4: Performance in computerized adaptive testing.

Dataset		Math2		
		Metric@Step		
Strategy	CDM	AUC/ACC@5	AUC/ACC@10	AUC/ACC@15
Random	IRT	70.80/64.35	74.51/66.76	77.63/69.01
	OR-IRT	**76.58/68.86**	**78.73/70.11**	**81.12/72.47**
	NCDM	70.16/63.92	73.89/66.36	77.39/69.17
	OR-NCDM	**77.10/69.03**	**79.72/71.06**	**81.78/73.25**
MAAT	IRT	70.53/63.67	74.49/66.78	77.80/69.15
	OR-IRT	**77.89/70.26**	**79.74/72.02**	**81.59/71.81**
	NCDM	71.35/64.85	74.50/66.90	77.25/69.26
	OR-NCDM	**78.76/69.38**	**80.53/72.00**	**81.90/73.23**
BECAT	IRT	70.66/64.06	74.36/65.42	77.68/68.72
	OR-IRT	**76.11/68.34**	**78.75/70.22**	**81.71/73.09**
	NCDM	70.23/65.74	74.93/68.93	77.99/69.82
	OR-NCDM	**76.48/68.16**	**79.55/70.81**	**81.62/73.13**

to the corresponding correct rates, with deeper shades of blue indicating higher correct rates, we achieve a visual representation of the students' Mas distribution. From Figure 8, it is clear that OR-NCDM and OR-KaNCD clusters all students S with high accuracy rates more cohesively than NCDM and KaNCD.

5.5 Validation on the Downstream Task

As an upstream task in the field of intelligent education, CD is applied in various downstream tasks. To validate the effectiveness of ORCDF, we chose to test it in the context of computerized adaptive testing (CAT) [42, 43]. Specifically, we integrate the commonly employed IRT and NCDM with our ORCDF, denoting these as OR-IRT and OR-NCDM, respectively. Our experimental settings align with recent research [43], which adopts a 7:2:1 split for students in the response logs of each dataset. Details can be found in Appendix C. As illustrated in Table 4, OR-NCDM performs better than OR-IRT, which validates the superiority of deep learning-based methods in CAT which is consistent with [42, 43]. OR-IRT and OR-NCDM significantly outperform their original versions. This validates the effectiveness of ORCDF in downstream tasks.

5.6 Hyperparameter Analysis

Effect of L. As shown in Figure 10 in Appendix D, a larger L decreases the model's training speed, while a smaller L results in poor performance. The recommended values of L are 3 or 4, which can yield relatively good performance. Notably, as L increases, the MND does not continually decrease, a phenomenon that seems different from what is observed in graph representation learning. We contend this could be related to the heterogeneity of the response graph and the complexity of student interactions, which we leave for future work.

The Effect of p_f. As depicted in Figure 11 in Appendix D, OR-NCDM is influenced by the flip ratio parameter. A too high flip ratio introduces more noise, deteriorating the model's performance. Typically, a p_f =0.15 yields better prediction performance, aligning with the established fact that everyone has a probability of guessing correctly or slipping, neither too high nor too low.

The Effect of $\lambda_{\mathbf{reg}}$. As illustrated in Figure 12 in Appendix D, this parameter controls the impact of guess and slip on model training, which varies across different datasets and requires tuning. It is observable that as the number of response logs in the dataset gradually increases, the optimal parameter value decreases. We recommend setting it to $1e^{-3}$.

The Effect of τ**.** As illustrated in Figure 13 in Appendix D, the temperature parameter τ affects the similarity between representations learned from the response graph and those from the flipped response graph. As the size of the dataset gradually increases, the better temperature value also gradually increases. Here, we recommend choosing 0.5 when the number of students is small and opting for 3.0 when there is a larger student population.

6 CONCLUSION

This paper proposes an oversmoothing-resistant cognitive diagnosis framework (ORCDF), where most existing CDMs can be integrated and thus enhanced. We, for the first time, identify the oversmoothing in CD and then address it by learning students' Mas from multiple perspectives, utilizing the proposed response graph and response-aware graph convolution network. Besides, we reformulate the guess and slip problem as noise edges in the response graph and deign a loss function to alleviate the problem. As long as the oversmoothing is addressed in CD, it greatly helps provide distinctive and personalized diagnostic results for students and teachers. However, ORCDF, while effective, is still not sufficiently interpretable enough in the field of intelligent education. More interpretable methods are expected to be developed to mitigate the oversmoothing issue explicably in cognitive diagnosis.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive comments. We also would like to thank Xinyue Ma for the reliable help. The algorithms and datasets in the paper do not involve any ethical issue. This work is supported by the National Natural Science Foundation of China (No. 62106076), National Social Science Fund of China (No. BEA230071), and Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901).

REFERENCES

[1] Haoyang Bi, Haiping Ma, Zhenya Huang, Yu Yin, Qi Liu, Enhong Chen, Yu Su, and Shijin Wang. 2020. Quality meets Diversity: A Model-Agnostic Framework for Computerized Adaptive Testing. In *Proceedings of the 20th IEEE International Conference on Data Mining*. Sorrento, Italy, 42–51.

[2] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach. In *Proceedings of the 8th Educational Data Mining*. Madrid, Spain, 532–535.

[3] Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *Proceedings of 21st Artificial Intelligence in Education*. Ifrane, Morocco, 69–73.

[4] Jimmy De La Torre. 2009. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics* 34, 1 (2009), 115–130.

[5] Mingyu Feng, Neil T. Heffernan, and Kenneth R. Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-adapted Interaction* 19, 3 (2009), 243–266.

[6] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. 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*. Virtual Event, 501–510.

[7] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. Sardinia, Italy, 249–256.

[8] Shelby J Haberman. 2005. Identifiability of parameters in item response models with unconstrained ability distributions. *ETS Research Report Series* 2005, 2 (2005), i–22.

[9] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. 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*. Virtual Event, 639–648.

[10] Lu Jiang, Kunpeng Liu, Yibin Wang, Dongjie Wang, Pengyang Wang, Yanjie Fu, and Minghao Yin. 2023. Reinforced Explainable Knowledge Concept Recommendation in MOOCs. *ACM Transactions on Intelligent Systems and Technology* 14, 3 (2023), 43:1–43:20.

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California.

[12] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Virtual Event, 904–913.

[13] Mingjia Li, Hong Qian, Jinglan Lv, Mengliang He, Wei Zhang, and Aimin Zhou. 2024. Foundation Model Enhanced Derivative-Free Cognitive Diagnosis. *Frontiers of Computer Science* (2024).

[14] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. In *Proceedings of the 32rd AAAI Conference on Artificial Intelligence*. New Orleans, LA, 3538–3545.

[15] Sheng Li, Quanlong Guan, Liangda Fang, Fang Xiao, Zhenyu He, Yizhou He, and Weiqi Luo. 2022. Cognitive Diagnosis Focusing on Knowledge Concepts. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta, GA, 3272–3281.

[16] Qi Liu, Runze Wu, Enhong Chen, Guandong Xu, Yu Su, Zhigang Chen, and Guoping Hu. 2018. Fuzzy Cognitive Diagnosis for Modelling Examinee Performance. *ACM Transactions on Intelligent Systems and Technology* 9, 4 (2018), 1–26.

[17] Shuo Liu, Hong Qian, Mingjia Li, and Aimin Zhou. 2023. QCCDM: A Q-Augmented Causal Cognitive Diagnosis Model for Student Learning. In *Proceedings of the 26th European Conference on Artificial Intelligence*. Kraków, Poland, 1536–1543.

[18] Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024. Inductive Cognitive Diagnosis for Fast Student Learning in Web-Based Intelligent Education Systems. In *Proceedings of the ACM on Web Conference 2024*. Singapore, 4260–4271.

[19] Yingjie Liu, Tiancheng Zhang, Xuecen Wang, Ge Yu, and Tao Li. 2023. New development of cognitive diagnosis models. *Frontiers of Computer Science* 17, 1 (2023), 171604.

[20] Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2023. XES3G5M: A Knowledge Tracing Benchmark Dataset with Auxiliary Information. In *Advances in Neural Information Processing Systems 37*. New Orleans, LA.

[21] Haiping Ma, Manwei Li, Le Wu, Haifeng Zhang, Yunbo Cao, Xingyi Zhang, and Xuemin Zhao. 2022. Knowledge-Sensed Cognitive Diagnosis for Intelligent Education Platforms. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. Atlanta, GA, 1451–1460.

[22] Yimeng Min, Frederik Wenkel, and Guy Wolf. 2020. Scattering GCN: Overcoming Oversmoothness in Graph Convolutional Networks. In *Advances in Neural Information Processing Systems 33*. Virtual Event.

[23] Pengyang Shao, Chen Gao, Lei Chen, Yonghui Yang, Kun Zhang, and Meng Wang. 2024. Improving Cognitive Diagnosis Models with Adaptive Relational Graph Neural Networks. *arXiv preprint arXiv:2403.05559* (2024).

[24] Junhao Shen, Hong Qian, Wei Zhang, and Aimin Zhou. 2024. Symbolic Cognitive Diagnosis via Hybrid Optimization for Intelligent Education Systems. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 14928–14936.

[25] Shuanghong Shen, Qi Liu, Zhenya Huang, Yonghe Zheng, Minghao Yin, Minjuan Wang, and Enhong Chen. 2024. A Survey of Knowledge Tracing: Models, Variants, and Applications. *IEEE Transactions on Learning Technologies* (2024).

[26] Jianwen Sun, Fenghua Yu, Sannyuya Liu, Yawei Luo, Ruxia Liang, and Xiaoxuan Shen. 2023. Adversarial Bootstrapped Question Representation Learning for Knowledge Tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa, Canada, 8016–8025.

[27] James B Sympson. 1978. A model for testing with multidimensional items. In *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN.

[28] Shiwei Tong, Jiayu Liu, Yuting Hong, Zhenya Huang, Le Wu, Qi Liu, Wei Huang, Enhong Chen, and Dan Zhang. 2022. Incremental Cognitive Diagnosis for Intelligent Education. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Washington, DC, 1760–1770.

[29] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[30] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural Cognitive Diagnosis for Intelligent Education Systems. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY.

[31] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2023. NeuralCD: A General Framework for Cognitive Diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2023).

[32] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *Proceedings of the 28th World Wide Web Conference*. San Francisco, CA, 2022–2032.

[33] Ting Wu, Hong Qian, Ziqi Liu, Jun Zhou, and Aimin Zhou. 2023. Bi-objective evolutionary Bayesian network structure learning via skeleton constraint. *Frontiers of Computer Science* 17, 6 (2023), 176350.

[34] Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy X. Huang. 2022. Hypergraph Contrastive Collaborative Filtering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Madrid, Spain, 70–79.

[35] Wei Xu and Yuhan Zhou. 2020. Course video recommendation with multimodal information in online learning platforms: A deep learning framework. *British Journal of Educational Technology* 51, 5 (2020), 1734–1747.

[36] Shangshang Yang, Haoyu Wei, Haiping Ma, Ye Tian, Xingyi Zhang, Yunbo Cao, and Yaochu Jin. 2023. Cognitive Diagnosis-Based Personalized Exercise Group Assembly via a Multi-Objective Evolutionary Algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 3 (2023), 829–844.

[37] Shangshang Yang, Xiaoshan Yu, Ye Tian, Xueming Yan, Haiping Ma, and Xingyi Zhang. 2023. Evolutionary Neural Architecture Search for Transformer in Knowledge Tracing. In *Advances in Neural Information Processing Systems 36*. Louisiana, NO.

[38] An Zhang, Wenchang Ma, Xiang Wang, and Tat-Seng Chua. 2022. Incorporating Bias-aware Margins into Contrastive Loss for Collaborative Filtering. In *Advances in Neural Information Processing Systems 35*. New Orleans, LA.

[39] An Zhang, Leheng Sheng, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2023. Empowering Collaborative Filtering with Principled Adversarial Contrastive Loss. In *Advances in Neural Information Processing Systems 36*. Louisiana, NO.

[40] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. 2023. Invariant Collaborative Filtering to Popularity Distribution Shift. In *Proceedings of the ACM Web Conference 2023*. ACM, Austin,TX, 1240–1251.

[41] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. 2019. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Anchorage, AK, 793–803.

[42] Yan Zhuang, Qi Liu, Zhenya Huang, Zhi Li, Shuanghong Shen, and Haiping Ma. 2022. Fully Adaptive Framework: Neural Computerized Adaptive Testing for Online Education. In *Proceeddings of the 36th AAAI Conference on Artificial Intelligence*. Virtual Event, 4734–4742.

[43] Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. 2023. A Bounded Ability Estimation for Computerized Adaptive Testing. In *Advances in Neural Information Processing Systems 37*. New Orleans, LA.

APPENDIX

The appendix is organized as follows:

• Appendix A analyzes the ORCDF's time complexity and compares it with other frameworks.

• Appendix B presents the detailed settings of compared baselines and other details about student performance perdition.

• Appendix C presents the detailed settings of the downstream tasks, namely, computerized adaptive testing.

• Appendix D further supplements the analysis with additional details regarding the hyperparameter analysis.

Notably, our code is available at https://github.com/lswhim/ORCDF.

Table 5: Abbreviations for terms.

Term	Abbreviation
Mastery Levels	**Mas**
Difficulty Levels	**Diff**
Cognitive Diagnosis	CD
Cognitive Diagnosis Model	CDM
Degree of Agreement	DOA

A TIME COMPLEXITY ANALYSIS

In this section, we present a detailed time complexity analysis of our proposed model OR-NCDM. We compare our time complexity with that of RCD, as RCD is the only CDM based on GNN.

Time Complexity Analysis of ORCDF. We take OR-NCDM as an exmaple. In OR-NCDM, we construct a response graph (ResG) $\mathcal{G}$ with three node and edge types based on $\mathbf{I}$ and $\mathbf{Q}$. Given that we do not employ the non-linear activation and feature transformation usually found in GNNs, the time complexity can be straightforwardly computed as $O(2|\mathcal{E}|Ld)$ for RGC, where L denotes the number of RGC' layers. d stands for the size of the embeddings. Due to the need for computing representations through the flipped ResG, the total time complexity amounts to $O(4|\mathcal{E}|Ld)$.

Time Complexity Analysis of RCD. In RCD, it construct three relation maps. Namely, an exercise-concept graph is constructed using $\mathbf{Q}$ and a student-exercise graph is formed using $\mathbf{I}$. Given that RCD employs the graph attention network, which necessitates the computation of attention coefficients between every pair of connected nodes, its time complexity belongs to $O(2|\mathcal{E}|LZ^2)$. Herein, Z represents the number of concepts ($d \ll Z$).

OR-NCDM evidently takes less time compared to RCD due to two main reasons. Firstly, due to the transformation layer reduces the embedding dimension to d, where d is much smaller than Z. Secondly, by removing complex operations like linear transformations in GNN, the graph convolution of RGC's computation become much faster than the GAT used in RCD.

In the experiment, we incorporate NCDM into all frameworks and use the speed of NCDM as the baseline, set at 1.0x. As shown in the figure, our proposed ORCDF is **18 times** faster than RCD and offers better prediction performance. When the number of knowledge concepts continuously increases, RCD tends to train too slowly and runs into out-of-memory issues, especially with large sets of knowledge concepts. In contrast, ORCDF maintains good performance, as demonstrated in scenarios like XES3G5M with 832 knowledge concepts on a single NVIDIA 3090 GPU, as detailed in Table 2.

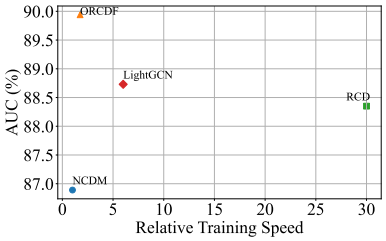


Figure 9: Relative training speed of different related frameworks on the Assist17 dataset.

B EXPERIMENTAL DETAILS

Interpretability Metric. DOA is defined as Eq. (13)

$$\mathrm{DOA}_k = \frac{\sum_{a,b \in S} \delta\left(\mathbf{Mas}_{s_a,c_k}, \mathbf{Mas}_{s_b,c_k}\right) \frac{\sum_{j=1}^{M} \mathbf{Q}_{jk} \wedge \varphi(j,a,b) \wedge \delta(r_{aj}, r_{bj})}{\sum_{j=1}^{M} \mathbf{Q}_{jk} \wedge \varphi(j,a,b) \wedge I(r_{aj} \neq r_{bj})}}{Z}, \tag{13}$$

where $Z = \sum_{a,b \in S} \delta(\mathbf{Mas}_{s_a,c_k}, \mathbf{Mas}_{s_b,c_k})$, $\mathbf{Q}_{j,k}$ indicates exercise e_j's relevance to concept c_k, $\phi(j,a,b)$ checks if both students s_a and s_b answered e_j, $r_{a,j}$ represents the response of s_a to e_j, and $I(r_{a,j} \neq r_{b,j})$ verifies if their responses are different, $\delta(r_{a,j}, r_{b,j})$ is 1 for a right response by s_a and a wrong response by s_b, and 0 otherwise.

Implementation Details. This section delineates the detailed settings when comparing our method with the baselines and state-of-the-art methods in both transductive scenario and inductive scenario. All experiments are run on a Linux server with two 3.00GHz Intel Xeon Gold 6354 CPUs and one RTX3090 GPU. All the models are implemented by PyTorch. For all methods that involve using MLP as the interaction function, we adopt the commonly used two-layer tower structure with hidden dimensions of 512 and 256. Additionally, we employ the approach used in NCDM to ensure that it satisfies the monotonicity assumption.

In the following, we elaborate on some details regarding the utilization of compared methods.

• DINA [4] is a representative CDM which models the mastery pattern with discrete variables (0 or 1).

• MIRT [27] is a representative model of latent factor CDMs, which uses multidimensional $\boldsymbol{\theta}$ to model the latent abilities. We set the latent dimension as 16 which is the same as [30]

• NCDM [30] is a deep learning based CDM which uses MLPs to replace the traditional interaction function (i.e., logistic function). We adopt the default parameters which are reported in that paper.

• RCD [6] leverages GNN to explore the relations among students, exercises and knowledge concepts. Here, to ensure a fair comparison, we solely utilize the student-exercise-concept component of RCD, excluding the dependency on knowledge concepts.

• KANCD [31] improves NCDM by exploring the implicit association among knowledge concepts to address the problem of knowledge coverage. Here, we adopt the default parameters which are reported in that paper.

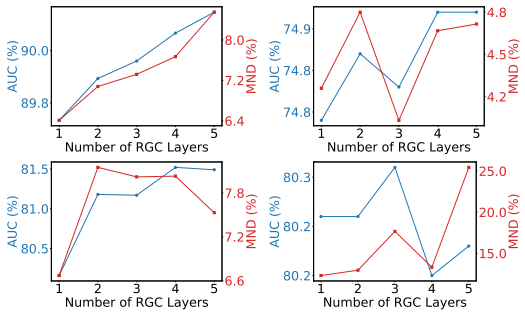


Figure 10: Effect of L on four datasets.

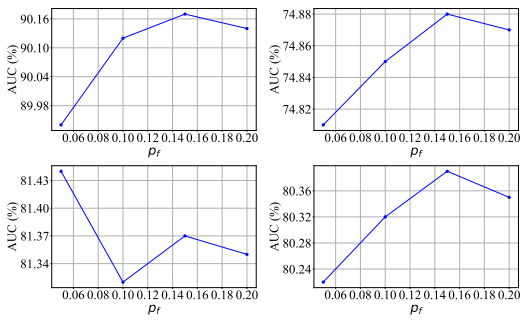


Figure 11: Performance under different p_f on four datasets.

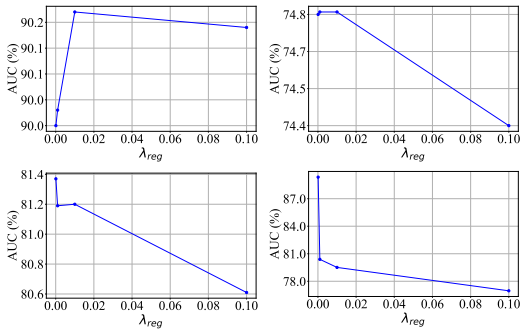


Figure 12: Performance under different $\lambda_{\mathbf{reg}}$ on four datasets.

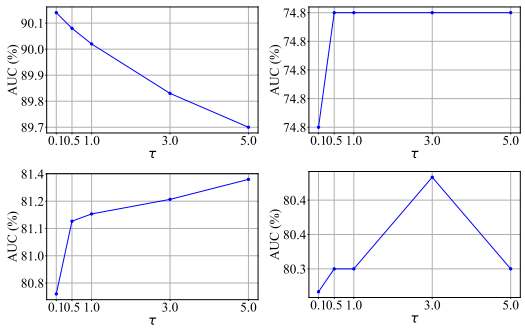


Figure 13: Performance under different τ on four datasets.

• KSCD [21] also explores the implicit association among knowledge concepts and leverages a knowledge-enhanced interaction function. Due to the absence of open-source code online, we have independently replicated KSCD.

• LightGCN [9] is a recent classic model that employs GCN in CF. In our context, we straightforwardly consider users as students and items as exercises. We set dimension as 32, the number of GCN layers as 3 which is the same as OR-NCDM for a fair comparison.

• HierCDF [12] is also a cognitive diagnosis framework that employs a Bayesian network, requiring a directed acyclic graph (DAG) to delineate the dependencies between knowledge concepts. It enables cognitive diagnosis models to learn mastery levels that adhere to the DAG structure, better aligning with the assumption of relationships between knowledge concepts in educational theory. We use the hyperparameters recommended in the original paper.

The implementation of DINA, MIRT, NCDM and KANCD comes from the public repository https://github.com/bigdata-ustc/EduCDM. For RCD, we adopt the implementation from the authors in https://github.com/bigdata-ustc/RCD. For LightGCN, we also use the code from the authors https://github.com/gusye1234/LightGCN-PyTorch. For HierCDF, we also use the code from the authors https://github.com/CSLiJT/HCD-code.

C DETAILS ABOUT COMPUTERIZED ADAPTIVE TESTING.

Computerized adaptive testing (CAT) primarily comprises CDM and item selection strategies. Its aim is to accurately determine students' mastery levels (Mas) with as few exercises as possible. The core of CAT often lies in designing a more effective item selection strategy [1, 43]. They often opt for simple and classic CDMs like IRT or NCDM. However, in reality, these diagnostic models suffer from the oversmoothing issue and tend to underperform.

In this study, we employ a classic dataset, known as Math2, which consists of 3911 students, 16 exercises, and 16 concepts. This dataset has been widely used in various researches as referenced in studies such as [16, 30]. The objective of CAT is to accurately estimate a student's Mas using the fewest possible steps (i.e., the smallest number of exercises). However, as the true Mas cannot be obtained as ground truth, we, like previous methods, use the student performance prediction task to validate the learned Mas. For more detailed information, we recommend the readers refer to [1, 43]. Here, we utilize three commonly selected strategies which can be applied on both IRT and NCDM. These strategies can be formulated as follows.

• Random is a simply strategy which select exercises randomly for each student in CAT.

• MAAT [1] utilizes the proposed expected model change to select exercises that are likely to have a significant impact on the student's Mas.

• BECAT [43] employs the concept of Coreset and utilizes expected gradient difference approximation to select exercises.

D DETAILS OF HYPERPARAMTER ANALYSIS

All figures correspond to datasets are in the order of Assist17, EdNet-1, Junyi and XESG35M. In all analyses regarding hyperparameters, we use OR-NCDM as an example.