ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230434

QCCDM: A Q-Augmented Causal Cognitive Diagnosis Model for Student Learning

Shuo Liu[a], Hong Qian[a;*], Mingjia Li[a] and Aimin Zhou[a]

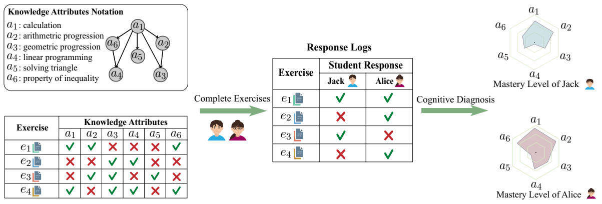
[a]Shanghai Institute of AI for Education and School of Computer Science and Technology, East China Normal University, Shanghai 200062, China

Abstract. Cognitive diagnosis is vital for intelligent education to determine students' knowledge mastery levels from their response logs. The Q-matrix, representing the relationships between exercises and knowledge attributes, improves the interpretability of cognitive diagnosis models. However, completing the Q-matrix poses an expensive and challenging task due to the fine-grained division of knowledge attributes. Moreover, a manually sparse Q-matrix can also compromise the accuracy and interpretability of deducing students' mastery levels, especially for infrequently observed or unseen knowledge attributes. To address this issue, this paper proposes a Q-augmented Causal Cognitive Diagnosis Model (QCCDM) for student learning. Specifically, QCCDM incorporates the structure causal model (SCM) to capture the causality between students' mastery levels on different attributes, which enables to infer their proficiency on rarely observed knowledge attributes with better accuracy and interpretability. Notably, with SCM, one can guide students on how to realize their self-improvement through intervention. Furthermore, we propose to augment the Q-matrix in QCCDM, which uses the manual Q-matrix as a prior to deduce the relationships between exercises and explicit as well as latent knowledge attributes, resulting in a complete and comprehensive assessment of students' abilities. We assess the efficacy of Q-augmentation across the widely-used Q-based cognitive diagnosis models and conduct the ablation study. The extensive experimental results on real-world datasets show that QCCDM outperforms the compared methods in terms of both accuracy and interpretability.

1 Introduction

Cognitive diagnosis (CD) is the cornerstone of many real-world applications such as medical diagnosis [9], course recommendation [34] and face-identification proficiency [13]. Particularly, in intelligent educational systems [19], cognitive diagnosis, as shown in Figure 1, endeavors to unearth the cognitive states of students via the response logs (e.g., scores on exercises), especially their proficiency levels (a.k.a., mastery levels) in specific knowledge attributes/concepts.

Over the course of recent decades, a myriad of cognitive diagnosis models (CDMs) have been proposed. The existing CDMs can be roughly dichotomized into two categories: Q-based and Q-irrelevant, where the Q-matrix (abbr. Q) represents the relationship between exercises and knowledge attributes, indicating which knowledge attributes are tested in a certain exercise. For Q-irrelevant CDMs, representative methods IRT [21] and MIRT [28] are known as latent factor models (LFMs). An example of IRT is the use of a single scalar θ to represent students' latent mastery level, along with the usage of logistic function as the interactive function. As an illustration, the 3PL-IRT model [10] is formulated as $P_i(\theta_j|r=1) = c_i + (1-c_i)\frac{\exp(a_i(\theta_j - b_i))}{1+\exp(a_i(\theta_j - b_i))}$, where $P_i(\theta_j|r=1)$ is the probability of correctly answering an exercise e_i by a student s_j, θ_j represents the latent trait level of s_j, and a_i, b_i and c_i are the discrimination, difficulty and guessing parameters for e_i, respectively. For Q-based CDMs, representative methods include DINA [5] and NCDM [31]. DINA uses the expectation-maximization algorithm to estimate a binary matrix that identifies students' mastery of specific knowledge attributes based on their responses (response logs) to assessment exercises with Q-matrix. To overcome the limitations of manually-designed interactive functions (e.g., logistic function), NCDM utilizes neural networks as interactive functions in conjunction with Q-matrix, achieving remarkable performance.



Exercise	Knowledge Attributes					
	a_1	a_2	a_3	a_4	a_5	a_6
e_1	✓	✓	✗	✗	✗	✓
e_2	✗	✗	✓	✓	✗	✗
e_3	✗	✓	✓	✗	✓	✗
e_4	✓	✗	✓	✓	✗	✓

Response Logs

Exercise	Student Response	
	Jack	Alice
e_1	✓	✓
e_2	✗	✓
e_3	✓	✗
e_4	✗	✓

Figure 1. An example of cognitive diagnosis.

Since Q-matrix explicitly expresses the relationship between exercises and knowledge attributes, it makes the inferred mastery levels more accurate and substantially enhances the interpretability of CDMs. However, it could pose a challenge owing to the sparsity and potentially incomplete annotation of Q-matrix. For example, in the task of finding the value of $\sin(x)$ given that $\cos(x) = (\sqrt{5})^{-1}$ and $0 \le x \le 0.5\pi$, this problem is closely related to the knowledge attribute of trigonometric functions, and human experts classify it as such in Q-matrix. However, some students may encounter difficulties in computation instead of trigonometric functions, such as calculating square roots or applying the Pythagorean theorem. As Q-matrix is a binary matrix, CDMs may overlook the impact of such computational skills, resulting in inaccurate estimations of students' mastery levels. Therefore, in this situation, existing Q-based CDMs may not fully capture the intrinsic relationship between latent knowledge attributes and exercises. As knowledge attributes become increasingly fine-grained [33], Q-matrix becomes more sparse and infrequently observed or unseen knowledge attributes occur. The small amount

[*] Corresponding Author. E-mail: hqian@cs.ecnu.edu.cn.

of exercises related to certain knowledge attributes may impede the accuracy of assessing students' mastery levels in corresponding fields, particularly when students choose not to try the specific types of exercises. This could significantly compromise the accuracy and interpretability of CDMs.

With this issue in mind, this paper proposes a Q-augmented Causal Cognitive Diagnosis Model (QCCDM) to address it. Specifically, a structure causal model (SCM) layer is incorporated as the component of QCCDM. SCM models the causality between knowledge attributes, i.e., attribute hierarchy in cognitive diagnosis. SCM enables more powerful and accurate inference of the mastery levels of the infrequent or unseen knowledge attributes in response logs. Notably, with SCM, one can guide students on how to realize their self-improvement through intervention. Furthermore, in QCCDM a data-driven continuous Q-augmentation component is proposed to uncover the relationship between exercises and explicit as well as latent knowledge attributes. Q-augmentation enhances the traditional manual Q-matrix (e.g., annotated by human experts) and retains the inherent relationship between exercises and explicit knowledge attributes through the usage of mask technique. With QCCDM, the students' abilities could be assessed completely and comprehensively. The extensive experimental results on real-world datasets show that Q-augmentation can significantly improve the performance of Q-based CDMs (e.g., DINA and NCDM), which implies that the proposed Q-augmentation is a plug-in component and possesses the merit of versatility. Besides, QCCDM outperforms the compared methods in terms of both accuracy and interpretability.

The subsequent sections respectively present the preliminaries, introduce the proposed QCCDM, show the experiment results and analysis, and finally conclude the paper.

2 Related Work

Cognitive Diagnosis Models. CDMs are used to evaluate student profiles by employing either latent factor models, such as item response theory (IRT) [21] and MIRT [28], or models based on patterns of attribute mastery (in this scenario Q-matrix can be applied), such as deterministic input, noisy and gate model (DINA) [5], neural cognitive diagnosis model (NCDM) [31]. For instance, DINA, a prominent example of CDMs, utilizes binary independent variables to represent mastery states, where 0 indicates an unmastered state and 1 represents a mastered state. NCDM is another well-known CDM that employs neural interactive functions and represents mastery patterns as continuous variables within the range of $[0, 1]$.

Knowledge Coverage Problem. The knowledge coverage problem is a crucial and challenging in cognitive diagnosis, as highlighted in [32]. With the increasing granularity of knowledge attributes, the Q-matrix becomes more sparse, resulting in a shortage of corresponding exercises for certain knowledge attributes. This inadequacy could lead to unreliable diagnostic results when knowledge coverage is incomplete for students. Recently, KANCD [32] has contributed to addressing this issue by exploring implicit knowledge association. This approach can deduce the mastery level of knowledge attributes, even when certain students have not completed sufficient exercises on them. However, this approach may encounter the issue of interpretability as it assumes that all knowledge attributes are related to one another, which may be too strong and may defy common sense in certain cases (e.g., trigonometric functions and three-variable linear equations). Instead, this work leverages the powerful reasoning and interpretability tool causal models to realize the informed inference of the mastery levels of different knowledge attributes with good confidence and reliability.

Attribute Hierarchy Methods. Attribute hierarchy (AH) describes the dependency among knowledge attributes. Attribute hierarchy model [16] is a class of CDMs that utilizes a rule-based approach to describing students' cognitive states under AH which characterizes it through the hierarchical cognitive assumption (HCA). It posits that mastery of parent attributes is a prerequisite for mastery of child attributes. The hierarchical diagnostic classification model (HCDM) [29] is an exemplary AHM that treats mastery patterns as a discrete space (i.e., 1 represents certain knowledge attribute is mastered, while 0 represents unmastered). However, HCDM is time consuming. HIERCDF [17] uses a discrete Bayesian network to model the dependency relationship between knowledge attributes with directed acyclic graph (DAG). RCD [8] utilizes graph neural network to model both directed and undirected dependency of knowledge attributes. Both of them are interpretable. However, HIERCDF and RCD may not fulfill effective self-improvement for students in certain educational cases, whereas the proposed QCCDM be equipped with SCM which enables intervention could help.

3 Preliminaries

3.1 Task Overview

Cognitive Diagnosis. Let $S = \{s_1, \ldots, s_N\}$, $E = \{e_1, \ldots, e_L\}$ and $A = \{a_1, \ldots, a_K\}$ denote the sets of students, exercises, and knowledge attributes, respectively. $|S| = N$, $|E| = L$ and $|A| = K$ are the size of each set. Given their individual interests and demands, students select appropriate exercises from a set of exercises to practice. The results of their practice sessions, which are recorded in response logs, can be represented as a set of triplets $R = \{(s, e, y) | s \in S, e \in E, y \in \{0, 1\}\}$, where y represents the score of a particular log, i.e., 1 means right and 0 means wrong. Q-matrix (manually annotated by human experts) represents the relationship among exercises and knowledge attributes, which can be regarded as a binary matrix $Q = (Q_{ij})_{L\times K}$, where $Q_{ij} \in \{0, 1\}$ means whether e_i relates to a_j or not and Q_{ij} is the element in the i-th row and j-th column of Q.

Attribute Hierarchy. The attribute hierarchy (AH) refers to the structure of cognitive dependencies among attributes in the cognitive states. Specifically, an attribute a_0 is an ancestor of an attribute a_1 in AH only if the acquisition of knowledge related to a_0 is a prerequisite for the acquisition of knowledge related to a_1. Formally, the attribute hierarchy considered is a directed acyclic graph (DAG), which is a causal graph $\mathcal{G}$ in this paper. $\mathcal{G} = (\mathcal{A}, \mathcal{E})$ denotes the causality among attributes, where $\mathcal{A}$ stands for nodes of attributes and $\mathcal{E}$ stands for the prerequisite relationships among attributes nodes in $\mathcal{G}$.

Problem Definition. Given the students' triplet logs R, a binary matrix $Q \in \mathbb{R}^{L\times K}$, and an attribute hierarchy represented by a DAG $\mathcal{G} \in \mathbb{R}^{K\times K}$, the goal is to infer Mas $\in \mathbb{R}^{N\times K}$, which denotes the latent mastery level of students on each attribute.

3.2 Structure Causal Model

Structural causal model (SCM) [24] is a powerful class of probabilistic graphical models used to represent and reason about causality between variables. SCM consists of a set of structural equations that captures the functional relationships between variables, allowing for the analysis of how changing of one variable affects the other variables. Recently, SCM has been applied extensively in disentangled representation learning and generative models, particularly in the context of causal models in the latent space [35, 25]. It has shown the impressive success in controllable image generation [22, 15], where the ability of manipulating specific attributes of an image is crucial.

While SCM has been used for achieving the causation of disentangled semantic factors through the weakly supervised ground-truth labels and randomly sampled noises in causal disentangled representation, this paper proposes an alternative approach. Specifically, the coefficients of the causal graph and latent representations are learned during the end-to-end training process.

4 Causality and Q-Augmentation in QCCDM

This section introduces two key components in QCCDM, which are the causal cognitive diagnosis model and Q-augmentation. Figure 2 sketches the overall framework of QCCDM.

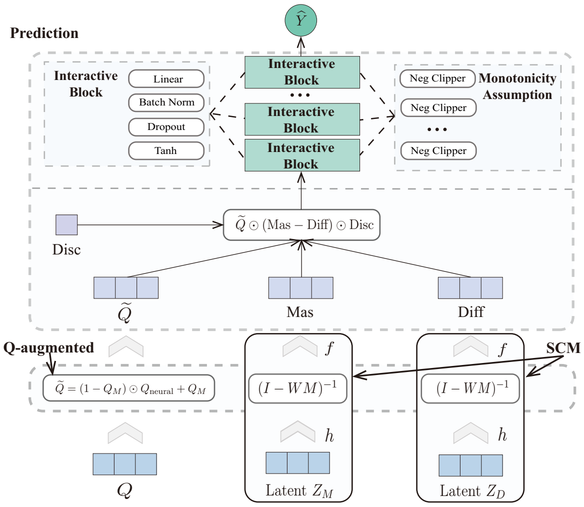


Figure 2. The framework of Q-augmented causal cognitive model.

4.1 Causal Cognitive Diagnosis Model

4.1.1 Causal Attribute Hierarchy

Following the hierarchical cognitive assumption (HCA) [16], we employ SCM to capture the causality between knowledge attributes based on the causal graph given by experts. This paper assumes that the aforementioned causal graph $\mathcal{G}$ is accurate, since the experts always highlight the explicit causality between knowledge attributes. Based on the assumption of accurate causal graph, the definition of causal hierarchical cognitive assumption is shown as Definition 1.

Definition 1 (Causal Hierarchical Cognitive Assumption, CaHCA)**.** *Given an attribute hierarchy which is characterized by an SCM with causal graph $\mathcal{G}$, the values of Mas on attributes are generated by $\mathcal{G}$.*

Definition 1 describes how the relationship between mastery levels on knowledge attributes can be better represented by a causal graph $\mathcal{G}$, which is more appropriate and dependable in the educational contexts. A simple example is univariate calculus and multivariable calculus. If a student has a good understanding of univariate calculus, it is likely that he or she will find it easier to handle multivariable calculus compared with a student with poor calculus skills. Therefore, incorporating the causality between knowledge attributes can assist CDMs in deducing mastery levels from response logs in a reasonable and interpretable manner.

4.1.2 SCM Layer

To model the causation among knowledge attributes, we utilize the causal graph without coefficients as a prior and design an SCM layer to capture the relationship among mastery level on knowledge attributes with a latent factor Z (e.g., diligence, learning strategy and learning environment). Specifically, $\text{Mas}_{i,a_c} = \sum_{a_j \in \text{Pa}(a_c)} \text{Mas}_{i,a_j} \mathcal{E}_{jc} + Z_{i,a_c}$. Herein $\text{Pa}(a_c)$ denotes the parent set of attribute a_c, $\mathcal{E}_{jc}$ denotes the edge weight of a_j to a_c. This means that the mastery level of a student s_i on a_c is constituted by its mastery level on the parent attributes and a latent personalized factor Z_{i,a_c}. However, such linear SCM [4, 14, 26] may be insufficient to capture the intricate causality among attributes and non-linear SCM [2, 37] may be inefficient to solve. Furthermore, non-linear functions commonly used in causal discovery [38] (e.g., MIM, MLP, GP) or Bayesian structure learning [18] may not be suitable for educational contexts (e.g., the sin function may represent when the parent attribute level goes somewhat higher, the child node may descend). Therefore, the generalized linear SCM [36] is utilized as a layer in neural networks in this paper to model the complex relation among attributes. The difficulty level of exercises can be obtained similarly. This results in the following formulas [36]

$$\text{Mas} = f\big((I - WM)^{-1} h(Z_M)\big), \text{Diff} = f\big((I - WM)^{-1} h(Z_D)\big). \tag{1}$$

$\text{Mas} \in \mathbb{R}^{N \times K}$ is the mastery level of students and $\text{Diff} \in \mathbb{R}^{L \times K}$ is the difficulty level of exercises. The functions $f(\cdot)$ and $h(\cdot)$ are applied to perform non-linear transformations on each element individually in the input matrix. $W \in \mathbb{R}^{K \times K}$ denotes the trainable matrix. With the given causal graph (without coefficients) as a prior, $\mathcal{G}$ is used as a graph mask, i.e., M in Figure 2. Z_M and Z_D denote latent factors for Mas and Diff, respectively. We can estimate the value of edge weight among knowledge attributes during the training process.

In the context of education, the non-linear functions f and h should satisfy two fundamental and mild requirements. Firstly, the non-linear functions should be monotonically increasing, as it aligns with the conventional intuition that improvements in a parent attributes should not cause a decrease in its child attributes. Secondly, the non-linear functions should be (highly) interpretable. This paper selects the Sigmoid function as the non-linear functions for f and h, as it meets the aforementioned requirements and can effectively constrain the mastery level within the range of $[0, 1]$, which is consistent with existing CDMs [17, 31]. We acknowledge that more complex and interpretable non-linear functions may be developed in the future work, and we look forward to exploring such possibilities.

4.2 Q-Augmentation

Educational datasets commonly employ the manually annotated Q-matrices that are binary. Q-matrix represents the relationship between knowledge attributes and exercises. As granularity of knowledge attribute categorization increases and big data becomes increasingly prevalent, Q-matrix is becoming more sparse which negatively impacts the interpretability of CDMs. Besides, as some exercises correspond solely to one fine-grained knowledge attribute, the sparsity issue worsens. In real-world educational situations, an exercise can correlate to multiple knowledge attributes while only a few of them may receive expert labeling. This paper assumes that the Q-matrix provided by experts is always correct but may be incomplete.

We call the labeled (e.g., by human experts) knowledge attributes as explicit knowledge attributes and unlabeled ones as latent knowledge attributes. It is reasonable to explore the relationship between exercises and latent attributes. To express the connection between exercises, explicit attributes and latent attributes, we propose to define the Q-augmentation matrix as Definition 2.

Definition 2 (Q-Augmentation)**.** *Q-augmentation represents the relationship between exercises and knowledge attributes, encompassing both latent and explicit knowledge attributes. Denote Q-augmentation as* $\widetilde{Q}$*, where* $\widetilde{Q}_{ij} = 1$ *if* $Q_{ij} = 1$*;* $\widetilde{Q}_{ij} = \epsilon \in [0,1]$ *if* $Q_{ij} = 0$*.*

One straightforward approach is to train the Q-augmentation as trainable parameters Q_{neural} through a data-driven process without constraints. However, there are two limitations. Firstly, the explicit knowledge attributes labeled by experts may be obscured during training process. Secondly, the resulting Q may be too dense, indicating that each exercise is associated with almost all attributes.

To incorporate expert knowledge, we use the prior Q (manually annotated by experts) as a mask Q_M and utilize $(1-Q_M)$ to represent the latent relationships between exercises and attributes which should be updated during the training process. Let $\odot$ represent the element-wise multiplication. With $(1 - Q_M) \odot Q_{\text{neural}}$, we can solely derive relationships between latent attributes and exercises. To sum up, this paper proposes to augment the Q-matrix by the following formula

$$\widetilde{Q} = (1 - Q_M) \odot Q_{\text{neural}} + Q_M \,. \tag{2}$$

To prevent the Q-augmentation $\widetilde{Q}$ learned from data-driven methods from being too dense, we propose the regularization term $\Omega(\widetilde{Q})$ as Eq. (3) to ensure its sparsity with entry-wise matrix norm $\|C\|_{1,1} = \sum_i \sum_j |C_{ij}|$. $\Omega(\widetilde{Q})$ helps reveal the relationship between exercises and truly relevant latent attributes, while avoiding being misled by spurious attributes. Q_M is not incorporated so as to keep the explicit attributes invariant. $\Omega(\widetilde{Q})$ is scaled to an appropriate magnitude.

$$\Omega(\widetilde{Q}) = \frac{\|(1 - Q_M) \odot Q_{\text{neural}})\|_{1,1}}{L \times K} \,. \tag{3}$$

5 Other Details in QCCDM

As mentioned in [6], the student module, exercise module, and their interactive function are crucial components of CDMs. In this paper, we adopt embedding and neural network to model each of these components, which is a prevalent technique in current CDMs. The whole algorithm is shown in Algorithm 1.

Embedding Module. Given a dataset with N students $S = \{s_1, s_2, \ldots, s_N\}$, we initialize $s_i \in \mathbb{R}^K$ as the factor of i-th student, which will be learned automatically during the training process. The exercises' difficulty and discrimination are modeled in the same way. In line with previous work [17, 31], we also model the discrimination of exercises, denoted as Disc $\in \mathbb{R}^{L \times 1}$. Each vector in the student embedding can be considered a latent factor Z_M of knowledge attributes, as mentioned in Section 4.1. Notably, this approach differs from previous CDMs [8, 17, 31] which directly model the mastery level of students. With the aid of the SCM layer, we can obtain the inferred mastery level by performing forward propagation on the causal graph, as presented in Eq. (1).

Monotonicity Assumption. The monotonicity assumption suggests that as a student's mastery level of a certain attribute increases, the probability of answering a related question correctly also increases. For instance, if two students attempt a calculus question but one has a higher level of mastery than the other, the former is more likely to answer the question correctly. To enforce this assumption during implementation, we use ReLU [1] or softplus [7] to ensure non-negative weights of linear layer, which is referred to as Neg Clipper in Figure 2.

Algorithm 1: Q-augmented Causal Cognitive Model

```
Input: Response logs: $R = \{(s, e, y)\}$, $M$: Graph mask, $Q_M$:
       Q-matrix, $T$: Maximum number of epochs, IB:
       Interactive blocks.
Output: Mas: Students' mastery level, $\widehat{Y}$: Predicted response
        set.
Initialize $Z_M$, $Z_D$, Disc, IB, $W$, $Q_{\text{neural}}$;
for $t = 1, 2, \ldots, T$ do
    Sample a subset $B \subset R$;
    Mas ← $\text{Sigmoid}((I - WM)^{-1}\text{Sigmoid}(Z_M))$;
    Diff ← $\text{Sigmoid}((I - WM)^{-1}\text{Sigmoid}(Z_D))$;
    Disc ← $\text{Sigmoid}(\text{Disc})$;
    $\widetilde{Q} \leftarrow (1 - Q_M) \odot Q_{\text{neural}} + Q_M$;
    state ← $\widetilde{Q}_B \odot (\text{Mas}_B - \text{Diff}_B) \odot \text{Disc}_B$;
    $\hat{y} \leftarrow \text{Sigmoid}\big(\text{IB}(\text{state})\big)$;
    $\mathcal{L} \leftarrow$
        $-\frac{1}{|B|}\sum_{i=1}^{|B|} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) + \lambda \cdot \Omega(\widetilde{Q})$;
    Update parameters Z, Diff, Disc, IB, $W$, $Q_{\text{neural}}$;
end
Mas ← $\text{Sigmoid}((I - WM)^{-1}\text{Sigmoid}(Z_M))$;
Add $\hat{y}$ to $\widehat{Y}$;
Return Mas and $\widehat{Y}$.
```

Interactive Block. Existing approaches before NCDM usually mine linear interactions of student exercising process by manually-designed function (e.g., logistic function), which is not sufficient for capturing complex relations in reality. Therefore, NCDM leverages MLP with monotonicity assumption as interactive function and achieves tremendous performance on large-scale datasets with high interpretability. However, the exclusive use of Sigmoid in NCDM may result in vanishing gradients, which can lead to poor performance in certain cases. We replace Sigmoid with tanh, as it has demonstrated superior performance in our experiments than other activation function. Furthermore, we incorporate batch normalization [12] and dropout [27] to mitigate vanishing gradients issue. The structure of interactive block is detailedly shown in Figure 2. These modifications improve the model's performance and enhance its robustness.

Loss Function. The primary loss function employed in QCCDM involves calculating the binary cross-entropy loss between the model's predictions and the true response scores in a mini-batch. In addition, to promote the sparsity and interpretability of Q-augmentation, we utilize regularization, as discussed in Subsection 4.2. Here, λ serves as a hyperparameter for regulating the density of $\widetilde{Q}$. In summary, the model's loss function can be expressed as $\mathcal{L} = -\sum_i y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) + \lambda \cdot \Omega(\widetilde{Q})$.

6 Experiments

This section first introduces the real-world datasets and experimental setup. Then, extensive experiments are conducted on these datasets in order to answer the following questions. The code and appendix of this paper is available at https://github.com/lswhim/CDM_ILOG.

Q1: How does QCCDM perform when compared with existing methods in predicting students' scores?
Q2: How does QCCDM perform when compared with existing approaches in terms of interpretability?
Q3: How do SCM and Q-augmentation contribute to the performance of QCCDM respectively?
Q4: Can QCCDM mitigate the knowledge coverage problem?
Q5: Is Q-augmentation genuinely appropriate for Q-based CDMs?
Q6: How does SCM guide students' self-improvement by intervention?
Q7: How does the hyperparameter λ affect QCCDM?

6.1 Dataset Description

The experiments are conducted on three real-world datasets, i.e., Junyi [3], Math1 and Math2 [20]. Junyi is an online math practicing log dataset provided by the Junyi Academy. Math1 and Math2 are collected from a high school math final exam which respectively contain senior one and senior two students' response logs in their exams. For Junyi, just as HIERCDF [17], to guarantee an adequate number of response logs for each student and reduce the dataset size, students with less than 15 response logs are filtered out, and 10000 students are selected randomly. Details of these datasets are shown in Table 1.

Table 1. Statistics of real-world datasets for experiments. Q-sparsity means the average number of attributes per exercise.

Datasets	Junyi	Math1	Math2
#Students	10000	4209	3911
#Exercises	734	15	16
#Attributes	734	11	16
#Response Logs	408,057	63135	58665
#Edges	927	21	36
Q-sparsity	1	3.2	3.25

6.2 Experimental Setup

This subsection introduces the experimental setup including compared methods, performance metrics, interpretability metric and AH metric. The details of experiment settings are provided in Appendix 1.

Baselines and State-of-the-Art Methods. CDMs can be classified into Q-based CDMs and Q-irrelevant CDMs according to whether they rely on Q or not. All the experiments are repeated with 10 seeds.

- MIRT [28] is selected as the representative model of Q-irrelevant CDMs, which uses multidimensional θ to model the latent abilities.
- DINA [5] is a typical CDM who models the mastery pattern with discrete variables (0 or 1).
- NCDM [31] is a cognitive diagnosis model that uses a neural network to replace the traditional interactive function (i.e., logistic function) with a monotonicity assumption, which is well-suited for large-scale datasets.
- HIERCDF [17] utilizes the Bayesian network to model the mastery pattern with DAG.
- KANCD [32] investigates the implicit association between knowledge attributes to mitigate the issue of knowledge coverage, which is a challenge that arises from the sparsity of Q.
- RCD [8] considers the complex relationships among students, exercises and attributes, and models them with graph neural networks.

Score Prediction Metrics. Evaluating the performance of CDMs can be a challenging task as it is often infeasible to obtain the true mastery levels of students. A commonly adopted practice is to assess the diagnostic models by predicting students' test scores. Therefore, akin to previous CDMs [20, 31, 17], we evaluate the performance of our model on a test set of students' correctness by utilizing classification and regression metrics such as AUC, accuracy (ACC), RMSE, and F1 score after partitioning the data into train and test sets. We adopt the same test size that is 0.2 as in previous work [17].

Interpretability Metric. While evaluation metrics can assess the accuracy of CDMs, obtaining interpretable diagnostic results is also a crucial aspect of CD. To this end, we utilized the degree of agreement (DOA), which is the same as [17, 31]. Intuitively, if a student s_u has a higher accuracy in answering questions related to an attribute a_k than a student s_v, then the probability of s_u mastering a_k should be higher than that of s_v, i.e., $\text{Mas}_{uk} > \text{Mas}_{vk}$. DOA is defined as Eq. (4)

$$\text{DOA}_k = \frac{\sum_{u,v \in S} \delta\left(\text{Mas}_{uk}, \text{Mas}_{vk}\right) \frac{\sum_{j=1}^{L} q_{jk} \wedge J(j,u,v) \wedge \delta\left(y_{uj}, y_{vj}\right)}{\sum_{j=1}^{L} q_{jk} \wedge J(j,u,v) \wedge I\left(y_{uj} \neq y_{vj}\right)}}{\Phi}, \tag{4}$$

where Φ is a normalization factor calculated as the sum of delta function $\delta(\text{Mas}_{uk}, \text{Mas}_{vk})$ over all student pairs, q_{jk} represents whether an exercise e_j is related to an attribute a_k, $J(j,u,v)$ indicates whether both s_u and s_v answered e_j, y_{uj} denotes the response of a student s_u to e_j, and $I\left(y_{uj} \neq y_{vj}\right)$ indicates whether their responses are different. $\delta\left(y_{uj}, y_{vj}\right)$ is 1 if the response of student s_u is correct and that of s_v is incorrect, and 0 otherwise. On Math1 or Math2, DOA is computed as the mean of all attributes, whereas on Junyi, we use the average of the top ten attributes with the highest number of logs [17] due to its large number of attributes.

AH Metric. To assess whether the end-to-end training process effectively models the causality between attributes, we use the Spearman rank coefficient [11] which is formulated as Eq. (5).

$$r_s(\text{Mas}_{\bullet,a_p}, \text{Mas}_{\bullet,a_c}) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2-1\right)}, \tag{5}$$

where $\text{Mas}_{\bullet,k}$ stands for the k-th column of matrix Mas. Herein a_p and a_c must be a parent attribute and a child attribute respectively, and $d_i = \text{rank}(\text{Mas}_{\bullet,a_p}) - \text{rank}(\text{Mas}_{\bullet,a_c})$. We calculate the average coefficient of all the given causality.

6.3 Experimental Results

Predict Performance (Q1). Table 2 shows that QCCDM almost outperforms both compared Q-irrelevant and Q-based CDMs. Due to the typically low standard deviation (lower than 0.001) of CDMs, we do not report them in the table. As RCD requires both directed and undirected graphs, we only utilize the directed part for a fair comparison. While RCD has a higher F1 score on Math1, KANCD obtains a lower RMSE on Math2, the other metrics are significantly lower compared to QCCDM. This indicates that QCCDM outperforms existing CDMs in predicting student scores.

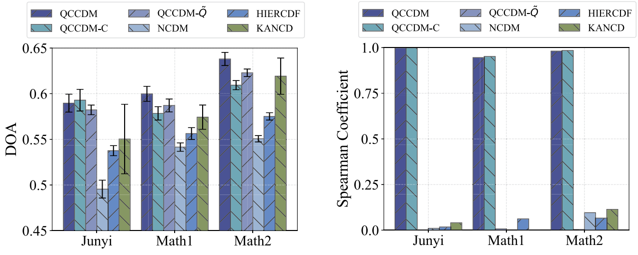
Interpretability Performance (Q2). We assess the interpretability of the inferred mastery levels of different methods by DOA. As shown in Figure 3(a), QCCDM outperforms NCDM, KANCD and HIERCDF in all three datasets in terms of DOA, indicating that QCCDM can accurately infer the mastery level while maintaining high interpretability. We also verify the QCCDM's capability of clustering students with different levels using t-SNE [30], a visualization technique employed to map students according to their Math2 test scores (with 16 being the highest score attainable). Specifically, we assign each student a label based on their test score, and apply t-SNE to

Table 2. Overall predict performance. In each column, an entry is bold if its mean value is the best. An entry is marked with solid circle if it is significantly worse than QCCDM and marked with hollow circle if it is significantly better than QCCDM by t-test with significance level 5%.

	Junyi				Math1				Math2			
Algo.	AUC	ACC	RMSE	F1	AUC	ACC	RMSE	F1	AUC	ACC	RMSE	F1
DINA	0.6353●	0.5002●	0.5263●	0.5786●	0.6795●	0.4744●	0.5456●	0.2501●	0.6831●	0.5076●	0.5390●	0.1934●
MIRT	0.7949●	0.7626●	0.4057●	0.8354●	0.7438●	0.6784●	0.4565●	0.7179	0.7683●	0.6962●	0.4495●	0.6916
NCDM	0.7839●	0.7485●	0.4106●	0.8254●	0.7450●	0.6774	0.4536●	0.7164●	0.7708●	0.6949●	0.4443●	0.6874 ●
HIERCDF	0.7848●	0.7516●	0.4098●	0.8292●	0.7426●	0.6727●	0.4528	0.7373○	0.7689●	0.6897●	0.4522●	0.6941
KANCD	0.7992●	0.7610●	0.4034●	0.8346●	0.7514	0.6835	0.4456	0.7330	0.7798	0.7005	**0.4380**	0.7001
RCD	0.8145●	0.7716●	0.3963	0.8348●	0.7410●	0.6779●	0.4497	**0.7412**○	0.7740●	0.6967●	0.4484	0.7011
QCCDM	**0.8171**	**0.7762**	**0.3928**	**0.8445**	**0.7553**	**0.6856**	**0.4446**	0.7232	**0.7815**	**0.7013**	0.4401	**0.7020**

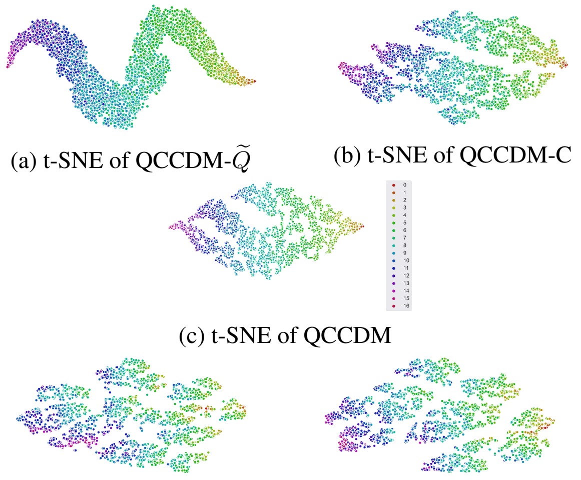
Table 3. Abaltion Study. QCCDM-C refers to the variant of QCCDM that solely utilizes the SCM layer, while QCCDM-$\widetilde{Q}$ corresponds to the variant that exclusively employs Q-augmentation. The meanings of bold, solid circle and hollow circle are the same of Table 2.

	Junyi				Math1				Math2			
Algo.	AUC	ACC	RMSE	F1	AUC	ACC	RMSE	F1	AUC	ACC	RMSE	F1
QCCDM-C	0.8060●	0.7673●	0.4013●	0.8406●	0.7462●	0.6821	0.4512	**0.7425**○	0.7702●	0.6912●	0.4554●	**0.7140**
QCCDM-$\widetilde{Q}$	0.8165●	0.7736●	0.3949	0.8439	0.7478●	0.6831	0.4504●	0.7304●	0.7762	0.6987●	**0.4440**	0.7042
QCCDM	**0.8171**	**0.7762**	**0.3928**	**0.8445**	**0.7553**	**0.6856**	**0.4446**	0.7232	**0.7815**	**0.7013**	0.4401	0.7020



(a) DOA (b) Spearman rank coefficient

Figure 3. DOA and Spearman rank coefficient of each method.



(a) t-SNE of QCCDM-$\widetilde{Q}$ (b) t-SNE of QCCDM-C

(c) t-SNE of QCCDM

(d) t-SNE of NCDM (e) t-SNE of HIERCDF

Figure 4. t-SNE visualizations of inferred Mas by different methods on Math2 dataset.

reduce the dimensionality of the inferred mastery levels obtained by NCDM, HIERCDF, and QCCDM. As shown in Figure 4, we have the following three key observations. 1) We can discern that in Figure 4(c), students who have answered the same number of questions correctly exhibit a discernible pattern from left to right. Conversely, this is not the case in NCDM from Figure 4(d) or HIERCDF from Figure 4(e). 2) Notably, the average-level students in QCCDM exhibit relatively higher dispersion. This finding is rational in real-world scenarios, since students with similar average score of the entire class may differ in their specific fields of strengths and weaknesses. 3) QCCDM is capable of accurately capturing the polarization phenomenon that may exist within a class, where students with exceptionally high or low scores are more tightly clustered than those with average scores.

Ablation Study (Q3). To understand how the SCM and Q-augmentation components affect the performance of the QCCDM, we conducted an ablation study where different versions of QCCDM are tested. QCCDM-C refers to the variant of QCCDM that solely utilizes the SCM layer, while QCCDM-$\widetilde{Q}$ corresponds to the variant that exclusively employs Q-augmentation. The results, which are shown in Table 3 and Figure 3(a), reveal that both components have a substantial positive impact on the performance and interpretability of the model. Although QCCDM-C outperforms QCCDM in terms of DOA on the Junyi, this can be largely attributed to the significantly larger number of knowledge attributes that pose a greater challenge for precisely identifying the latent knowledge attributes and exercises, thereby affecting the performance of QCCDM. Besides, the high Spearman rank coefficients observed across all three datasets suggest that incorporating the causality between knowledge attributes during the training process is an effective strategy, as depicted in Figure 3(b). In contrast, NCDM, KANCD and HIERCDF, exhibit poor performance, suggesting that the inferred mastery levels of individual attributes in these methods may not be strongly correlated with one another. We employ t-SNE to visually represent the inferred mastery levels of various versions of QCCDM applied to the Math2, which are shown in Figure 4(a)(b)(c). The findings suggest that the Q-augmentation can achieve more compact clustering of students with comparable test scores, while the SCM layer provides a more profound understanding of the distinctions between individual students. Figure 4(c) demonstrates that the integration of SCM and Q-augmentation components in QCCDM yields more comprehensive clustering results, suggesting a complementary relationship between the two components.

Knowledge Coverage Problem (Q4). Like previous work [32], in contrast to randomly splitting the logs, we choose to designate e_2, e_3, e_{16} in Math2 dataset [20] as the test set, whereas considering all other exercises as training set. This design ensures that the CDMs will not encounter certain attributes (unseen attributes a_1, a_2, infre-

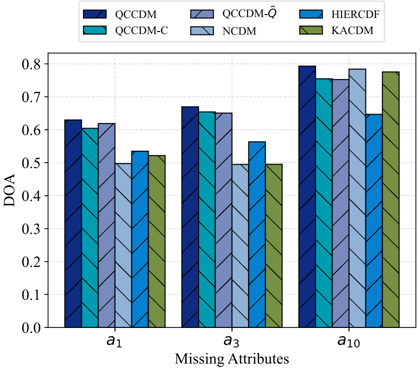


Figure 5. DOA of missing attributes. QCCDM-$\widetilde{Q}$ and QCCDM-C are the ablation of QCCDM.

quent attributes a_{10}) associated with e_2, e_3, e_{16} during the training phase. As shown in Figure 5, existing CDMs may not get reasonable diagnosis result, but all versions of QCCDM achieves better interpretability with high DOA. KANCD leverages knowledge association to mitigate the knowledge coverage problem. However, its interpretability may be limited by the assumption of implicit relationships between knowledge attributes. Causal model can provide more reasonable diagnostic results. For instance, a student may have a strong grasp of calculus, but the knowledge of multivariable calculus may be limited or absent from a standard exam. Due to insufficient updates during training, current CDMs may generate inexplicable diagnostic results for multivariable calculus.

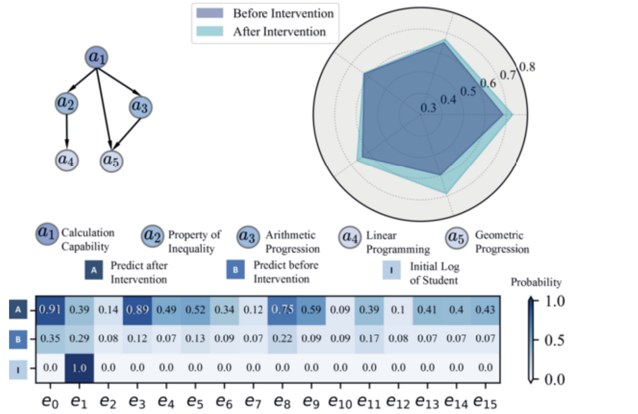
However, given the causality between calculus and multivariable calculus, students who exhibit strong aptitude in calculus are likely to excel in multivariable calculus as well. The incorporation of Q-augmented allows us to capture the hidden associations between exercises and multivariable calculus, effectively reducing the impact of selection bias. In summary, both SCM and Q-augmentation approaches can effectively mitigate the knowledge coverage problem and the combination of these methods yields highly efficient results.

Versatility of Q-augmentation (Q5). Building upon the fact that the inclusion of Q significantly improves the interpretability of CDMs, we substitute Q with Q-augmentation in Q-based CDMs and carry out comparative experiments, as presented in Table 4. For instance, $\widetilde{Q}$-DINA denotes DINA with Q-augmentation. Q-augmentation clearly outperforms manually designed Q in CDMs, as evidenced by its higher accuracy and DOA. Notably, NCDM, HIERCDF, and KANCD show significant improvements with Q-augmentation, while DINA only has slight enhancements due to its discrete mastery pattern. We also analyze that Q-augmentation can better describe the relationship of exercise's difficulty and knowledge attributes in Appendix 2.

Self-Improvement (Q6). Intuitively, causality can play a critical role in guiding students' self-improvement through intervention [23]. Figure 6 demonstrates the intervention undertaken on Student 3436 from Math2, who answered only one question correctly (bottom row in Figure 6) due to low calculation capability (CC). Consequently, we intervene on Student 3436's CC (i.e., assign more related exercises, change learning strategies or environment) and calculate the probability of accurate responses to each question in the future (shown in the third row). These probabilities are contrasted against the probabilities prior to the intervention, as depicted in the second row. For instance, e_0 is associated with the properties of inequality and CC. This student shows significant improvement when the CC increase, indicating that he or she may have failed to answer the question due to the low CC.

Table 4. Comparing the mean performance of Q-based CDMs with replaced Q-augmentation (i.e., $\widetilde{Q}$-DINA). The meanings of bold, solid circle and hollow circle are the same of Table 2.

Datasets	Metrics	DINA	$\widetilde{Q}$-DINA	NCDM	$\widetilde{Q}$-NCDM	HIERCDF	$\widetilde{Q}$-HIERCDF	KANCD	$\widetilde{Q}$-KANCD
Junyi	AUC	0.6350●	**0.6438**	0.7779●	**0.8119**	0.7824●	**0.7953**	0.7845●	**0.8142**
	ACC	0.5001●	**0.5186**	0.7432●	**0.7713**	0.7506●	**0.7581**	0.7500●	**0.7721**
	RMSE	0.5261	**0.5232**	0.4145●	**0.3997**	0.4122●	**0.4076**	0.4111●	**0.3954**
	F1	0.5783●	**0.6043**	0.8219●	**0.8377**	0.8261●	**0.8317**	0.8316●	**0.8405**
	DOA	0.5063	**0.5065**	0.5022●	**0.5696**	0.4796●	**0.5033**	0.5612●	**0.6114**
Math1	AUC	0.6795	**0.6839**	0.7457●	**0.7519**	0.7443	**0.7446**	0.7514●	**0.7564**
	ACC	0.4744	**0.4757**	0.6780	**0.6806**	0.6739	**0.6769**	0.6835●	**0.6892**
	RMSE	**0.5456**	0.5462	0.4523	**0.4506**	**0.4524**	0.4567	0.4456	**0.4417**
	F1	**0.2501**	0.2489	0.7320●	**0.7441**	0.7272●	**0.7503**	0.7330●	**0.7378**
	DOA	0.5056	**0.5067**	0.5334●	**0.5594**	0.5593●	**0.5761**	0.5711●	**0.6005**
Math2	AUC	0.6817●	**0.7024**	0.7671●	**0.7759**	**0.7686**	0.7642	0.7798●	**0.7811**
	ACC	0.5065●	**0.5271**	0.6911	**0.6950**	**0.6728**	0.6705	0.7005	**0.7022**
	RMSE	0.5411●	**0.5331**	0.4467	**0.4429**	0.4500	**0.4513**	0.4380	**0.4374**
	F1	**0.1917**○	0.1598	0.6757	**0.6760**	0.6326●	**0.6859**	0.7001	**0.7015**
	DOA	0.5042●	**0.5115**	0.5440●	**0.5738**	0.5762●	**0.5827**	0.6161●	**0.6372**



	e_0	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}
A	0.91	0.39	0.14	0.89	0.49	0.52	0.34	0.12	0.75	0.59	0.09	0.39	0.1	0.41	0.4	0.43
B	0.35	0.29	0.08	0.12	0.07	0.13	0.09	0.07	0.22	0.09	0.09	0.17	0.08	0.07	0.07	0.07
I	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 6. Intervention result of student 3436 in Math2 dataset.

Hyperparameter Analysis (Q7). We conduct a hyperparameter experiment to study the effect of λ on the density of $\widetilde{Q}$ across all datasets. The results, presented in Appendix 3, show that increasing the value of λ leads to more sparse $\widetilde{Q}$, which can negatively impact the model's performance, as indicated by the lower F1 score on Math2. Besides, a higher value of λ may result in each exercise being related to all knowledge attributes, which is not reasonable. Thus, it is important to maintain a reasonable range for λ to ensure that the relation between exercises and latent knowledge attributes is effectively captured while maintaining good prediction performance for students.

7 Conclusion

This paper presents a novel framework QCCDM, which leverages a SCM layer to provide more accurate diagnostic results in the context of sparse Q. We also propose continuous Q-augmentation as an enhancement for the manually labelled Q provided by the dataset, which allows for exploiting the relationship between latent knowledge attributes and exercises. Compared with existing CDMs, QCCDM provides superior performance and interpretability. Notably, the causality between knowledge attributes can aid students in personalizing their self-improvement strategies. Q-augmentation is versatile and can be applied to Q-based CDMs. QCCDM is an attempt to take a solid step forward in developing both interpretable and accurate CDMs that can aid in students' personalized learning, hoping to take the best of both worlds. In the future, developing strategies for utilizing SCM without knowing the causal graph as a prior is expected.

Acknowledgements

We would like to thank the anonymous reviewers for constructive comments. The algorithms and datasets in the paper do not involve any ethical issue. This work is supported by “Chenguang Program” sponsored by Shanghai Education Development Foundation and Shanghai Municipal Education Commission (No. 21CGA32), Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901), and Natural Science Foundation of Shanghai (No. 21ZR1420300).

References

[1] Abien Fred Agarap, ‘Deep learning using rectified linear units (ReLU)’, *arXiv preprint arXiv:1803.08375*, (2018).

[2] Ruichu Cai, Jie Qiao, Zhenjie Zhang, and Zhifeng Hao, ‘SELF: structural equational likelihood framework for causal discovery’, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 1787–1794, New Orleans, LA, (2018).

[3] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen, ‘Modeling exercise relationships in e-learning: A unified approach’, in *Proceedings of the 8th Educational Data Mining*, pp. 532–535, Madrid, Spain, (2015).

[4] David Maxwell Chickering, ‘Optimal structure identification with greedy search’, *Journal of Machine Learning Research*, **3**, 507–554, (2002).

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

[6] Louis V DiBello, Louis A Roussos, and William Stout, ‘A review of cognitively diagnostic assessment and a summary of psychometric models’, *Handbook of Statistics*, **26**, 979–1030, (2006).

[7] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia, ‘Incorporating second-order functional knowledge for better option pricing’, pp. 472–478, Denver, CO, (2000).

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

[9] Xuan Guo, Rui Li, Qi Yu, and Anne R Haake, ‘Modeling physicians’ utterances to explore diagnostic decision-making’, in *Proceedings of 26th International Joint Conference on Artificial Intelligence*, pp. 3700–3706, Melbourne, Australia, (2017).

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

[11] Andréas Heinen and Alfonso Valdesogo, ‘Spearman rank correlation of the bivariate student t and scale mixtures of normal distributions’, *Journal of Multivariate Analysis*, **179**, 104650, (2020).

[12] Sergey Ioffe and Christian Szegedy, ‘Batch normalization: Accelerating deep network training by reducing internal covariate shift’, in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, Lille, France, (2015).

[13] Géraldine Jeckeln, Ying Hu, Jacqueline G Cavazos, Amy N Yates, Carina A Hahn, Larry Tang, P Jonathon Phillips, and Alice J O’Toole, ‘Face identification proficiency test designed using item response theory’, *arXiv preprint arXiv:2106.15323*, (2021).

[14] Markus Kalisch and Peter Bühlmann, ‘Estimating high-dimensional directed acyclic graphs with the PC-algorithm’, *Journal of Machine Learning Research*, **8**, 613–636, (2007).

[15] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, and Sriram Vishwanath, ‘CausalGAN: Learning causal implicit generative models with adversarial training’, in *Proceedings of the 6th International Conference on Learning Representations*, British Columbia, Canada, (2018).

[16] Jacqueline P Leighton, Mark J Gierl, and Stephen M Hunka, ‘The attribute hierarchy method for cognitive assessment: A variation on tatsuoka’s rule-space approach’, *Journal of Educational Measurement*, **41**(3), 205–237, (2004).

[17] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang, ‘HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework’, in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 904–913, Virtual, (2022).

[18] Mingjia Li, Hong Qian, and Aimin Zhou, ‘Hybrid Bayesian network structure learning via evolutionary order search’, *Computer Science*, (2023).

[19] Qi Liu, ‘Towards a new generation of cognitive diagnosis’, in *Proceedings of 30th International Joint Conference on Artificial Intelligence*, pp. 4961–4964, Montreal, Canada, (2021).

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

[21] Frederic Lord, ‘A theory of test scores’, *Psychometric Monographs*, (1952).

[22] Raha Moraffah, Bahman Moraffah, Mansooreh Karami, Adrienne Raglin, and Huan Liu, ‘CAN: A causal adversarial network for learning observational and interventional distributions’, *arXiv preprint arXiv:2008.11376*, (2020).

[23] Judea Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2000.

[24] Judea Pearl, *Causality*, Cambridge University Press, 2009.

[25] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang, ‘Weakly supervised disentangled generative causal representation learning’, *Journal of Machine Learning Research*, **23**, 1–55, (2022).

[26] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan, ‘A linear non-gaussian acyclic model for causal discovery’, *Journal of Machine Learning Research*, **7**(10), (2006).

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, ‘Dropout: a simple way to prevent neural networks from overfitting’, *Journal of Machine Learning Research*, **15**(1), 1929–1958, (2014).

[28] James B Sympson, ‘A model for testing with multidimensional items’, in *Proceedings of the 1977 Computerized Adaptive Testing Conference*, Minneapolis, MN, (1978).

[29] Jonathan Templin and Laine Bradshaw, ‘Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies’, *Psychometrika*, **79**, 317–339, (2014).

[30] Laurens van der Maaten and Geoffrey Hinton, ‘Visualizing data using t-SNE’, *Journal of Machine Learning Research*, **9**, 2579–2605, (2008).

[31] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang, ‘Neural cognitive diagnosis for intelligent education systems’, in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, (2020).

[32] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su, ‘Neuralcd: A general framework for cognitive diagnosis’, *IEEE Transactions on Knowledge and Data Engineering*, **35**(8), (2023).

[33] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al., ‘Instructions and guide for diagnostic questions: The neurips 2020 education challenge’, *arXiv preprint arXiv:2007.12061*, (2020).

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

[35] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang, ‘CausalVAE: Disentangled representation learning via neural structural causal models’, in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, Virtual, (2021).

[36] Yue Yu, Jie Chen, Tian Gao, and Mo Yu, ‘DAG-GNN: DAG structure learning with graph neural networks’, in *Proceedings of the 36th International Conference on Machine Learning*, pp. 7154–7163, Long Beach, CA, (2019).

[37] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing, ‘DAGs with NO TEARS: continuous optimization for structure learning’, in *Advances in Neural Information Processing Systems 31*, pp. 9492–9503, Montreal, Canada, (2018).

[38] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing, ‘Learning sparse nonparametric dags’, in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425, Sicily, Italy, (2020).