Deep Generative Modeling for Cognitive Diagnosis via Exploratory DeepCDMs

Jia Liu and Yuqi Gu

Department of Statistics, Columbia University

Abstract

Deep generative modeling is a powerful framework in modern machine learning, renowned for its ability to use latent representations to predict and generate complex high-dimensional data. Its advantages have also been recognized in psychometrics. In this paper, we substantially extend the Deep Cognitive Diagnostic Models (DeepCDMs) in Gu (*Psychometrika*, 89:118–150, 2024) to challenging exploratory scenarios with deeper structures and all Q-matrices unknown. The exploratory DeepCDMs can be viewed as an adaptation of deep generative models (DGMs) toward diagnostic purposes. Compared to classic DGMs, exploratory DeepCDMs enjoy critical advantages including identifiability, interpretability, parsimony, and sparsity, which are all necessary for diagnostic modeling. We propose a novel layer-wise expectation-maximization (EM) algorithm for parameter estimation, incorporating layer-wise nonlinear spectral initialization and L_1 penalty terms to promote sparsity. From a parameter estimation standpoint, this algorithm reduces sensitivity to initial values and mitigates estimation bias that commonly affects classical approaches for deep latent variable models. Meanwhile, from an algorithmic perspective, our method presents an original layer-wise EM framework, inspired by modular training in DGMs but uniquely designed for the structural and interpretability demands of diagnostic modeling. Extensive simulation studies and real data applications illustrate the effectiveness and efficiency of the proposed method.

Keywords: Exploratory Cognitive Diagnosis; Deep Generative Modeling; Deep Cognitive Diagnostic Models (DeepCDMs); Identifiability; Layer-wise EM Algorithm.

1 Introduction

Over the past two decades, Cognitive Diagnosis Models (CDMs) have become increasingly prominent in educational and psychological measurement (e.g., Junker and Sijtsma, 2001; von Davier, 2008; Henson et al., 2009; Rupp et al., 2010; de la Torre, 2011; Chen et al., 2015; von Davier and Lee, 2019). CDMs are a class of psychometric models that use item response

data to infer examinees' mastery status on multiple discrete latent *attributes*, such as skills, subskills, or diagnostic criteria. In most applications, each attribute is assumed to be binary, representing the presence or absence of a specific cognitive ability or psychological trait. By estimating an individual's profile across these attributes, CDMs facilitate detailed diagnostic reporting. This information enables practitioners and educators to identify students' strengths and weaknesses at a granular level, supporting the design of targeted interventions and more individualized feedback.

Recently, interest in adopting higher-order structures for CDMs has grown, aiming to capture interdependencies between the latent attributes (de la Torre and Douglas, 2004; Templin et al., 2008; de la Torre and Song, 2009). Most existing models adopt a single layer of higher-order continuous latent traits to explain correlations among lower-level latent attributes (e.g., de la Torre and Douglas 2004; Templin et al. 2008; Bradshaw and Templin 2014; Ma 2022; Liu et al. 2025). Although these single-layer higher-order models offer an interpretable and simplified representation of attribute dependencies, they may be limited in modeling deeper latent hierarchies or providing more granular cognitive diagnoses. To model deeper level cognitive processes, the recent Deep Cognitive Diagnostic Models (DeepCDMs) proposed by Gu (2024) employ a deep architecture to capture probabilistic relationships across *multiple discrete latent* layers. DeepCDMs flexibly let each of these layers deliver diagnostic information at a distinct level of granularity. Despite this added depth, DeepCDMs remain parsimonious through compact parameterization and are mathematically identifiable under intuitive conditions.

In this paper, we show that the advantages of DeepCDMs can be further leveraged by generalizing them to an exploratory setting, where the attribute relationships between adjacent layers (i.e., all the $\mathbf{Q}$-matrices) are unknown. The exploratory DeepCDMs can be viewed as an adaptation of deep generative models (DGMs) for psychometrics and educational measurement, with additional constraints imposed to serve diagnostic purposes. DeepCDMs share structural similarities with several existing DGMs, such as deep belief networks (DBNs; Hinton et al. 2006); see Section 2.3 for further discussion. This connection highlights the expressive power of DeepCDMs from the perspective of DGMs. In particular, DeepCDM's layered architecture defines a hierarchical generative process, suitable for

modeling students' hierarchical and heterogeneous cognitive processes behind data. This structure enables DeepCDMs to approximate highly complex response distributions while maintaining a tractable form for layer-wise learning.

Although usual DGMs (Hinton et al., 2006; Salakhutdinov and Hinton, 2009) excel at predictive and generative performance, their architectures and estimation algorithms are often heuristically designed and lack rigorous statistical foundations. Importantly, whether the parameters underlying the latent representations are uniquely identifiable is largely unknown for DGMs. This gap motivates us to introduce exploratory DeepCDMs, which are built for diagnostic purposes and are *fully identifiable*. Exploratory DeepCDMs are identifiable under transparent conditions on the between-layer $\mathbf{Q}$-matrices (Gu, 2024). Identifiability ensures that no two distinct parameter sets yield the same marginal distribution of the observed responses, thereby guaranteeing consistent parameter estimation. As a consequence, DeepCDMs can provide statistically reliable personalized diagnoses of hierarchical latent abilities. The identifiability conditions naturally imply an interpretable shrinking-ladder-shaped sparse deep architecture, enabling the model to capture the latent skills from fine-grained (shallower and closer to the response data layer) to coarse-grained (deeper and more higher-order). Statistically, such architectures also induce parsimonious parameterizations, crucial for reflecting test design constraints in real-world educational assessments.

Parameter estimation is a challenging issue for exploratory DeepCDMs, as the parameters and $\mathbf{Q}$-matrices across all layers are need to be estimated. The commonly used estimation methods for related hierarchical models are Markov chain Monte Carlo (MCMC; Robert and Casella, 2004) method and EM algorithm (Dempster et al., 1977). Gu (2024) employed MCMC for confirmatory DeepCDMs with known $\mathbf{Q}$-matrices. For exploratory DeepCDMs, MCMC can, in principle, be developed by incorporating additional sampling steps for the $\mathbf{Q}$-matrix entries. However, when the $\mathbf{Q}$-matrices are unknown and the latent structure involves more than two layers—as in the settings considered in this work—significant practical challenges such as initialization sensitivity, slower convergence, MCMC mixing difficulties, and increased computational cost may limit its scalability and efficiency. The classical EM, as explained later in Section 3.4, though faster, suffers from (a) extreme sensitivity to initialization—since all parameters must be initialized simultaneously in a highly nonconvex,

multi-layer parameter space; and (b) cyclic bias accumulation, where errors in one layer's estimation propagate through both the E- and M-steps into other layers over successive iterations. On the other hand, although many algorithms have been proposed for general DGMs in machine learning (e.g., Hinton et al., 2006; Hinton and Salakhutdinov, 2006; Ranganath et al., 2015; Le Roux and Bengio, 2008; Salakhutdinov and Hinton, 2009), they are not directly applicable to DeepCDMs, as their typically overparameterized architectures do not satisfy the parsimony and identifiability requirements of diagnostic modeling and are not designed to promote sparsity or interpretability.

In this work, we propose a novel layer-wise EM algorithm for regularized maximum likelihood estimation with a layerwise L_1 penalty in exploratory DeepCDMs. The algorithm estimates parameters and $\mathbf{Q}$-matrices sequentially, starting from the bottom layer, where a one-layer EM algorithm is used to estimate both the coefficient and proportion parameters. These proportion parameters are then used to generate pseudo-samples of latent attributes, which serve as input for the next layer. We will continue this process one layer after another until all layers are estimated. This strategy is not only intuitive but also grounded in the model's generative structure: marginalizing out deeper layers naturally yields a standard one-layer CDM at the bottom, justifying the use of a one-layer EM for its estimation. In higher layers, each step builds on the most informative signals from the previous one—either as estimated distributions or generated pseudo-observations—thus respecting the model's hierarchical nature. Interestingly, the identifiability proof shows that identifiability can be examined and established in a layer-by-layer manner, thanks to the formulation of the directed graphical model and the discrete nature of the latent attributes. This theoretical insight also supports the design of our proposed algorithm and provides a solid foundation for treating imputed attributes as if observed in each step. Additionally, our layerwise estimation strategy conceptually aligns with the modular training principles widely used in deep generative modeling, where complex models are progressively trained through simpler, localized components. A more detailed discussion on this is in Section 3.5.2.

Initialization plays a crucial role in EM-based estimation, particularly in exploratory settings where the $\mathbf{Q}$-matrices are unknown and must be estimated. In such cases, the parameter space becomes more complex, and a well-informed initialization can greatly enhance

convergence stability and estimation quality. To this end, we adopt a fast, non-iterative procedure based on universal singular value thresholding (USVT), which yields reliable starting values with theoretical guarantees under certain conditions (Chatterjee, 2015; Zhang et al., 2020). The initialization is conducted in a sequential, layer-by-layer manner. For each layer, the input matrix is first denoised via truncated SVD, then linearized by applying the inverse link function, and then a second SVD followed by Varimax rotation is applied to recover a sparse coefficient matrix, promoting sparsity and identifiability. We adopt a penalized estimation framework where all $\mathbf{Q}$-matrices are treated as unknown and estimated from data. At each layer, $\mathbf{Q}$-matrix estimation is framed as a latent variable selection problem, with an L_1 penalty imposed on the coefficient parameters to encourage sparsity. The M-step of each layer's EM update is solved via cyclical coordinate descent (Friedman et al., 2010; Tay et al., 2023), efficiently maximizing the penalized log-likelihood. Additionally, as discussed in Section 3.6, although the algorithm is developed under an exploratory framework, it can be readily adapted for confirmatory applications. Our extensive simulation studies demonstrate the good performance of the proposed layer-wise EM in challenging scenarios involving three latent layers. Finally, we illustrate the practical utility of exploratory DeepCDM using data from the 2019 Trends in International Mathematics and Science Study (TIMSS) assessment.

The remainder of this paper is organized as follows. Section 2 introduces the exploratory DeepCDMs framework, discusses its formulation as a deep generative model, and addresses the identifiability issues. Section 3 presents an efficient layer-wise algorithm for parameter estimation for exploratory DeepCDMs. Section 4 presents simulation studies to evaluate the performance of the proposed layer-wise EM algorithm for exploratory DeepCDMs under various measurement models. Section 5 applies the proposed method to empirical data from the TIMSS 2019 assessment. Finally, Section 6 gives concluding remarks.

2 Exploratory DeepCDMs Framework

In this section, we present the exploratory DeepCDM framework. We will build on the concepts of confirmatory DeepCDMs in Gu (2024) and provide additional details for the exploratory setting. We then discuss how exploratory DeepCDMs adapt DGMs for psycho-

metrics, highlighting their architectural similarities and the additional structural constraints for facilitating diagnostic feedback. Finally, we discuss the theoretical identifiability of the model, with formal results provided in the Supplementary Materials.

2.1 Model Setup

The DeepCDM framework is developed to address the need for diagnostic modeling at multiple granularities. It is formally defined using the terminology of probabilistic graphical models (Wainwright et al., 2008; Koller and Friedman, 2009), particularly directed graphical models. These models employ graphs to compactly represent the joint distribution of high-dimensional random variables, where nodes correspond to variables and edges encode their direct probabilistic relationships.

We first review the definition of a *Directed Acyclic Graph* (DAG), also referred to as a Bayesian network (Pearl, 1988). In a DAG, every edge has a direction, and no directed cycles are allowed. Consider M random variables, $X_1, \ldots, X_M$, which correspond to M nodes in the graph. If a directed edge goes from X_ℓ to X_m, we say that X_ℓ is a *parent* of X_m, and X_m is a *child* of X_ℓ. Let $\mathrm{pa}(m) \subseteq \{1, \ldots, M\}$ denote the index set of all parents of X_m. Define $\mathbb{P}(X_m \mid X_{\mathrm{pa}(m)})$ as the conditional distribution of X_m given its parents $X_{\mathrm{pa}(m)}$. Based on this DAG structure, the joint distribution of $X_1, \ldots, X_M$ factorizes as follows:

$$\mathbb{P}(X_1, \ldots, X_M) = \prod_{m=1}^{M} \mathbb{P}(X_m \mid X_{\mathrm{pa}(m)}). \tag{1}$$

We now present the general DeepCDM formulation. For a DeepCDM with D latent layers, we denote the d-th latent layer as $\mathbf{A}^{(d)} = (A_1^{(d)}, \ldots, A_{K_d}^{(d)})$ for each $d = 1, 2, \ldots, D$, where larger d correspond to deeper layers. In DeepCDMs, all edges are directed top-down and occur only between adjacent layers, defining a generative process from high-level latent variables to observed responses. Specifically, the bottom layer consists of the observed response variables for the J items, denoted as $\mathbf{R} = (R_1, \ldots, R_J)$. The first latent layer, right above the bottom layer, captures the most fine-grained latent attributes, represented as $\mathbf{A}^{(1)}$. These are generated from the second latent layer $\mathbf{A}^{(2)}$, and the process continues recursively up to the deepest layer $\mathbf{A}^{(D)}$. Figure 1 gives an example of a DeepCDM with

three latent layers ($D = 3$). Given the variables in the layer right above, the variables within each layer of the DeepCDM are conditionally independent. This structure intuitively models how more specific latent skills are successively derived from more general, higher-level latent "meta-skills." A natural assumption, supported by the model's identifiability conditions, is that deeper layers should consist of fewer latent variables, i.e., $K_1 > K_2 > \cdots > K_D$ (see Theorems 1, 2, and 3 in the Supplementary Material for detailed identifiability conditions).

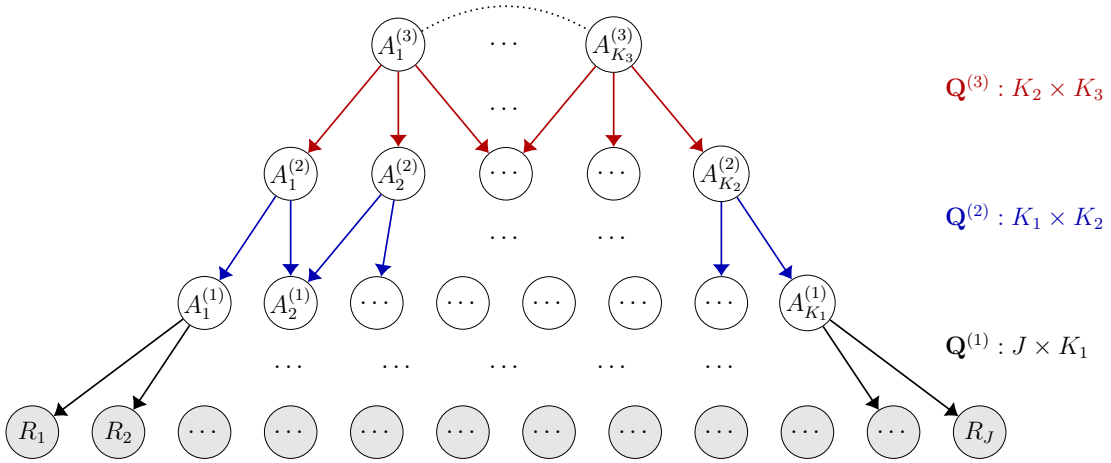

Figure 1: A ladder-shaped three-latent-layer DeepCDM. Gray nodes are observed variables, and white nodes are latent ones. Multiple layers of binary latent variables $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$, and $\mathbf{A}^{(3)}$ successively generate the observed binary responses $\mathbf{R}$. Binary matrices $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, and $\mathbf{Q}^{(3)}$ encode the sparse connection patterns between adjacent layers in the graph.

In traditional CDMs with a single layer of K latent attributes, the $\mathbf{Q}$-matrix (Tatsuoka, 1983) is a fundamental component that specifies the relationship between items and the latent attributes. Specifically, $\mathbf{Q} = (q_{j,k})_{J\times K}$, where $q_{j,k} = 1$ if the item j measures the latent attribute k, and $q_{j,k} = 0$ otherwise. Since the edges in a graphical model reflect direct statistical dependencies, $q_{j,k} = 1$ or 0 also conveys whether the k-th latent node is a parent of the j-th observed node. Consequently, the $\mathbf{Q}$-matrix encodes the bipartite graph structure between the observed and latent layers. Extending this idea to DeepCDMs, with D latent layers, requires D matrices, denoted as $\mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}, \ldots, \mathbf{Q}^{(D)}$, to capture the dependence relationships between any two adjacent layers. Specifically, $\mathbf{Q}^{(1)} = \left(q_{j,k}^{(1)}\right)$ has size $J \times K_1$, similar to the single $\mathbf{Q}$-matrix in traditional CDM, describes the graph between the observed data layer and the shallowest latent layer. While for $d = 2, \ldots, D$, the matrix $\mathbf{Q}^{(d)} = \left(q_{k,\ell}^{(d)}\right)$ has size $K_{d-1} \times K_d$ and represents the dependencies between latent variables

in the $(d-1)$th and dth latent layers. The entry $q_{k,\ell}^{(d)} = 1$ or 0 indicates whether the latent variable $A_\ell^{(d)}$ is a parent of $A_k^{(d-1)}$. In this paper, we consider the challenging setting of *exploratory* DeepCDMs, where all $\mathbf{Q}$-matrices are unknown and need to be estimated.

Based on the general definition of DAGs in (1) and the DeepCDM setup, the *joint distribution* of all variables, including the latent ones, is given by:

$$\mathbb{P}(\mathbf{R}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)}) = \mathbb{P}(\mathbf{R} \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)}) \cdot \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}), \quad (2)$$

$$\text{where} \quad \mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)}) = \prod_{j=1}^{J} \mathbb{P}^{\text{CDM}}(R_j = r_j \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)}), \quad \text{and} \quad (3)$$

$$\mathbb{P}(\mathbf{A}^{(d-1)} = \boldsymbol{\alpha}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)}) = \prod_{k=1}^{K_{d-1}} \mathbb{P}^{\text{CDM}}(A_k^{(d-1)} = \alpha_k^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)}), \quad (4)$$

where $\boldsymbol{r}$ represents an observed response pattern and $\boldsymbol{\alpha}^{(d-1)}$ represents a latent pattern for the $(d-1)$th latent layer. The superscript "CDM" in the conditional distributions of (3) and (4) indicates that the conditional distribution within each layer of the generative process adheres to a CDM. By marginalizing out all latent layers $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)}$ in (2), we obtain the marginal distribution of the observed response vector $\mathbf{R}$:

$$\mathbb{P}(\mathbf{R} = \boldsymbol{r}) = \sum_{\boldsymbol{\alpha}^{(1)}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}} \mathbb{P}(\mathbf{R} = \boldsymbol{r}, \mathbf{A}^{(1)} = \boldsymbol{\alpha}^{(1)}, \ldots, \mathbf{A}^{(D)} = \boldsymbol{\alpha}^{(D)}). \quad (5)$$

This work focuses on binary observed and latent variables, where $\boldsymbol{r} \in \{0,1\}^J$ and $\boldsymbol{\alpha}^{(d)} \in \{0,1\}^{K_d}$. Each observed variable reflects whether a response is correct or incorrect, while each latent variable indicates the presence or absence of a specific skill or higher-level attribute. Similar to traditional CDMs, the latent variables $\mathbf{A}^{(D)}$ in the deepest layer of a DeepCDM are modeled using a categorical distribution:

$$\mathbb{P}(\mathbf{A}^{(D)} = \boldsymbol{\alpha}_\ell) = \pi_{\boldsymbol{\alpha}_\ell}^{(D)}, \quad \forall \boldsymbol{\alpha}_\ell \in \{0,1\}^{K_D}. \quad (6)$$

Here, $K_D > 1$. The proportion parameters $\pi_{\boldsymbol{\alpha}_\ell}^{(D)}$ are subject to the constraint $\sum_{\boldsymbol{\alpha}_\ell \in \{0,1\}^{K_D}} \pi_{\boldsymbol{\alpha}_\ell}^{(D)} = 1$. With this, we complete the specification of a general DeepCDM.

2.2 Specific Examples of DeepCDMs

This subsection presents concrete examples of DeepCDMs that fall under the general framework outlined in Section 2.1. For notational convenience, we also denote the observed response layer $\mathbf{R}$ as $\mathbf{A}^{(0)}$, enabling a unified expression for the layerwise conditional distributions: $\mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)})$ for $d = 1, \ldots, D$. We then define specific DeepCDM variants according to the diagnostic model adopted for each layerwise conditional.

Example 1 (Main-effect DeepCDMs)**.** We use the term "Main-effect DeepCDMs" to refer broadly to DeepCDMs in which each layerwise conditional distribution follows a main-effect diagnostic model. In this setup, the probability that $A_j^{(d-1)} = 1$ is governed by the main effects of its parent attributes, modeled via a link function $f(\cdot)$:

$$\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) = f\Big(\beta_{j,0}^{(d)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(d)} \left\{q_{j,k}^{(d)} \alpha_k\right\}\Big). \tag{7}$$

Here, $\beta_{j,k}^{(d)}$ is nonzero only when $q_{j,k}^{(d)} = 1$. When f is the identity function, Equation (7) reduces to the Additive Cognitive Diagnosis Model (ACDM; de la Torre, 2011). If f is the inverse logit function, Equation (7) gives a Logistic Linear Model (LLM; Maris, 1999).

Example 2 (All-effect DeepCDMs)**.** We refer to DeepCDMs in which the layerwise conditionals follow an all-effect diagnostic model as "All-effect DeepCDMs." In an all-effect diagnostic model, the probability of $A_j^{(d-1)} = 1$ depends on both the main effects and all possible interaction effects of the parent attributes:

$$\begin{aligned}\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) = f\Big(&\beta_{j,0}^{(d)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(d)} \left\{q_{j,k}^{(d)} \alpha_k\right\} \\ &+ \sum\nolimits_{1 \le k_1 < k_2 \le K_d} \beta_{j,k_1 k_2}^{(d)} \left\{q_{j,k_1}^{(d)} \alpha_{k_1}\right\} \left\{q_{j,k_2}^{(d)} \alpha_{k_2}\right\} + \cdots + \beta_{j,12\cdots K_d}^{(d)} \prod\nolimits_{k=1}^{K_d} \left\{q_{j,k}^{(d)} \alpha_k\right\}\Big).\end{aligned} \tag{8}$$

Similar to the main-effect model, not all β-coefficients above are needed. If $\boldsymbol{q}_j^{(d)}$, the j-the row of $\mathbf{Q}^{(d)}$, contains K_j entries of "1", then 2^{K_j} parameters are required in (8). With the identity link function, (8) defines the Generalized DINA model (GDINA; de la Torre, 2011), while the inverse logit function yields the Log-linear CDM (LCDM; Henson et al., 2009).

Example 3 (DeepDINA)**.** The DINA model can be regarded as a special case of the all-effect CDM, where only the highest-order interaction term among the required attributes is retained, and all lower-order effects are constrained to zero:

$$\mathbb{P}^{\text{DINA}}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) = f\Big(\beta_{j,0}^{(d)} + \beta_{j,\mathcal{K}_j^{(d)}} \prod_{k \in \mathcal{K}_j^{(d)}} q_{j,k}^{(d)} \alpha_k\Big), \tag{9}$$

where $\mathcal{K}_j^{(d)} = \Big\{k \in [K];\, q_{jk}^{(d)} = 1\Big\}$ denotes the set of attributes measured by item j. The model assumes that students are capable of an item only if they master all required attributes for that item. So, $\beta_{j,\mathcal{K}_j^{(d)}}$ is the only non-zero coefficient for item j in layer d.

One can also specify a DeepDINO model, a specific type of DeepCDM where the DINO model is used to model each latent layer (Gu, 2024). Due to the duality between DINA and DINO, the identifiability and algorithm applicable to DeepDINA are also applicable to DeepDINO. Therefore, we do not introduce it here and refer readers to Gu (2024) for details.

Example 4 (Hybrid DeepCDMs)**.** A key strength of the DeepCDM framework is its flexibility in allowing different diagnostic models (e.g., DINA, main-effect, all-effect) to be applied across various layers. This is referred as *Hybrid DeepCDMs*, which strike a balance between model complexity and parsimony, offering flexibility in designing diagnostic models based on specific needs. For instance, in practical scenarios, the most general all-effect diagnostic model may be used at the bottom layer to model how fine-grained attributes affect the observed responses, while simpler models like main-effect or DINA could be applied in deeper layers to enhance interpretability and reduce complexity.

As demonstrated earlier, only particular coefficients, determined by the $\mathbf{Q}$-matrices and the specified measurement models, in the generating DeepCDM should be non-zero. However, since all $\mathbf{Q}$-matrices, $\mathbf{Q}^{(d)}, d = 1, \ldots, D$, are unknown, the sparsity pattern of the coefficient vectors is also unknown. Therefore, we assume all coefficients in the model as unknown and estimate them by maximizing a regularized log-likelihood. The $\mathbf{Q}$-matrices can then be inferred by identifying the non-zero coefficients in $\boldsymbol{\beta}^{(d)}$, $d = 1, \ldots, D$. We defer the details of the mechanism for identifying the entries $q_{jk}^{(d)}$ to Section 3.2.

2.3 DeepCDMs as Deep Generative Models (DGMs)

As previously mentioned, the DeepCDM framework can be viewed as an adaptation of DGMs for psychometrics and educational measurement, where additional structural constraints are introduced to enable diagnostic feedback. Exploratory DeepCDMs share architectural similarities with several existing DGMs. For example, when the activation function f is defined as the inverse logit, DeepCDMs resemble DBNs (Hinton et al., 2006) with binary-valued hidden units. However, a key structural difference lies at the top of the network: DBNs assume an *undirected* graph between the top two layers—forming an restricted Boltzmann machine (RBM) —while DeepCDMs adopt a *fully directed*, top-down architecture across all layers. This design enables DBNs to use a heuristic greedy layer-wise pretraining procedure based on contrastive divergence, a technique specific to training undirected models such as RBMs (Hinton et al., 2006; Hinton and Salakhutdinov, 2006). Consequently, such training strategies are not directly applicable to DeepCDMs due to its directed nature, which is more interpretable for modeling hierarchical skill generation.

DeepCDMs also share a top-down generative structure with DEFs (Ranganath et al., 2015), an unsupervised framework using exponential family distributions to model each layer's conditional distribution. DEFs aim to capture compositional semantics through hierarchical latent representations. However, DEFs rely on black-box variational inference methods with neural network–based posterior approximations, which prevent recovery of interpretable parameters—such as $\mathbf{Q}$-matrices—and thus cannot provide individualized diagnostic feedback, a central aspect of cognitive diagnosis.

Another related framework is the deep discrete encoders (DDE; Lee and Gu 2025), a deep generative model designed for rich data types with discrete latent layers. While DDEs and DeepCDMs share architectural and identifiability properties, their goals differ. DDEs aim to address machine learning concerns like overparameterization and lack of interpretability, constructing general-purpose identifiable DGMs. In contrast, DeepCDMs are specifically designed for psychometrics, with each adjacent pair of latent layers constituting a CDM. This structure allows for diagnostic-specific measurement assumptions, addressing varied diagnostic goals and enhancing usability in real-world assessment settings.

Taken together, these distinctions highlight the unique features of exploratory DeepCDMs over conventional DGMs. While most DGMs prioritize data generation or predictive performance and typically lack identifiability and sparsity constraints, DeepCDMs refocus the modeling effort on offering reliable individualized diagnostic feedback and discovering hierarchical latent skill structures. Furthermore, by enforcing sparsity in the coefficient matrices to reflect item–attribute relationships, DeepCDMs provide valuable insights into test design—an essential feature for practical use in educational and psychological assessment, often overlooked in classical DGM frameworks.

2.4 Identifiability

As noted earlier, a key strength of DeepCDMs lies in their formal identifiability guarantees, which apply to both confirmatory and exploratory settings (Gu, 2024). These results are detailed in the Supplementary Materials. In brief, the identification conditions impose explicit structural constraints on the between-layer $\mathbf{Q}$-matrices, offering practical guidance for model design and implementation. Although the specific conditions vary across diagnostic models, they consistently require an increasingly *shrinking latent structure* for deeper layers. That is, the number of latent variables typically decreases with depth, often subject to constraints such as $K_d > c \cdot K_{d+1}$ for some constant $c > 1$ depending on the model. This hierarchy reflects the principle of *statistical parsimony* in DeepCDMs. For instance, in a two-layer DeepDINA model with $K_1 = 7$ and $K_2 = 2$, the number of nonzero parameters is $2K_1 + 2^{K_2} - 1 = 17$, compared to $2^{K_1} - 1 = 127$ in a saturated attribute model without higher-order latent structure. Such substantial reductions in complexity make DeepCDMs especially attractive for applications with fine-grained latent attributes and limited sample sizes. In exploratory settings, while all parameters must be estimated, the identifiability conditions naturally promote sparsity in the true generating model, facilitating parameter recovery and interpretation of the latent attributes.

A central insight underlying these proofs is that the identifiability of DeepCDMs can be established in a layer-by-layer fashion, proceeding from the bottom (shallow) layer to the top (deepest) layer. This approach is justified by the directed nature of the graphical model and the discreteness of latent variables. Two core ideas facilitate this stepwise identifiability.

First, in a multi-layer directed graphical model with only top-down connections between adjacent layers, marginalizing out deeper latent layers yields a Restricted Latent Class Model (RLCM) (Xu, 2017; Gu and Xu, 2020). Once the distribution of the shallowest latent layer is identified through this RLCM, it can be treated *theoretically as if observed* for identifying the next deeper layer. Second, identifiability in RLCMs holds for any marginal distribution of latent attributes, provided the $\mathbf{Q}$-matrix meets specific conditions. This property allows identifiability to propagate upward through layers, even when deeper latent variables introduce complex dependencies.

3 Proposed Estimation Algorithms

In this section, we propose a novel layer-wise EM algorithm for estimating exploratory DeepCDMs. We begin by introducing some notation. Let $\mathbf{R}_{1:N}$ denote the response data matrix of size $N \times J$, representing the observed responses of N students to J items. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(D)})$ and $\mathbf{Q} = (\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)})$ denote the sets of continuous parameters and $\mathbf{Q}$-matrices across all layers, respectively. Let $\boldsymbol{\pi}^{(d)}$ denote a vector composed of the probability mass function $\boldsymbol{\pi}_{d,\boldsymbol{\alpha}^{(d)}} = \mathbb{P}(\mathbf{A}^{(d)} = \boldsymbol{\alpha}^{(d)})$ for all $\boldsymbol{\alpha}^{(d)} \in \{0,1\}^{K_d}$, $d = 1, 2, \ldots, D$. The parameters to be estimated include all continuous parameters in $\boldsymbol{\beta}$, all $\mathbf{Q}$-matrices in $\mathbf{Q}$, and the proportion parameter $\boldsymbol{\pi}^{(D)}$ for the deepest latent layer. Directly maximizing the marginalized log-likelihood to estimate the $\mathbf{Q}$-matrices is computationally prohibitive, even when the number of layers D and the dimensionalities K_d ($d = 0, 1, \ldots, D; K_0 = J$) are of moderate size. This challenge arises from the need to search over an enormous space of possible $\mathbf{Q}$-matrix configurations—specifically, $\prod_{d=1}^{D} 2^{K_{d-1} \cdot K_d}$ combinations—each requiring evaluation of a profile likelihood. To circumvent this combinatorial burden, we instead frame $\mathbf{Q}$-matrix estimation as a structured model selection task, addressed through a regularized likelihood approach that encourages sparsity in the parameter space (Chen et al., 2015).

The regularized marginal log-likelihood is given by:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\pi}^{(D)}, \mathbf{Q} \mid \mathbf{R}_{1:N}) = \sum_{i=1}^{N} \log \left\{ \sum_{\boldsymbol{\alpha}^{(1)}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}} \left[\mathbb{P}(\mathbf{R}_i \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)}, \boldsymbol{\beta}^{(1)}) \right. \right.$$

$$\cdot \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}; \boldsymbol{\pi}^{(D)})\Big]\Bigg\} - N \cdot P_{\boldsymbol{s}}(\boldsymbol{\beta}), \quad (10)$$

where the L_1 penalty function $P_{\boldsymbol{s}}(\boldsymbol{\beta})$ enforces sparsity across layers and is defined as

$$P_{\boldsymbol{s}}(\boldsymbol{\beta}) = \sum_{d=1}^{D} P_s(\boldsymbol{\beta}^{(d)}) = \sum_{d=1}^{D} s_d \sum_{\beta_{j,k}^{(d)} \in \boldsymbol{\beta}^{(d)}} \left|\beta_{j,k}^{(d)}\right|. \quad (11)$$

Here, s_d denotes the regularization parameter for layer d, and each $\boldsymbol{\beta}^{(d)}$ represents the set of model-specific coefficient parameters at that layer, with its structure determined by the chosen measurement model (e.g., main-effect, all-effect, or DINA), as detailed in Section 2.2.

The nested summation over multiple layers of latent attributes in (10) renders direct optimization infeasible. While the classical EM algorithm offers a principled framework for estimating exploratory DeepCDMs, it is not without limitations. In practice, its effectiveness can be hindered by the model's structural complexity and the high dimensionality of the parameter space. These challenges—stemming from the layered latent architecture and the combinatorial nature of $\mathbf{Q}$-matrix estimation—can increase sensitivity to initialization and compromise the scalability and stability of the algorithm. A more detailed discussion of these shortcomings is provided in Section 3.4. These considerations motivate our development of the layer-wise EM algorithm, introduced in the following section.

In the remainder of this section, we first introduce the classical EM algorithm and briefly discuss its limitations in Subsection 3.1. The proposed layer-wise EM algorithm is presented in Subsection 3.2, followed by the initialization strategy in Subsection 3.3. Subsection 3.4 highlights the advantages of the layer-wise EM over the classical EM algorithm. Subsection 3.5 discusses how the layer-wise concept connects to broader principles and algorithms, including identifiability derivation and related methods for DGMs. Finally, Subsection 3.6 discusses the extension of the layer-wise EM to the confirmatory DeepCDM setting.

3.1 The EM Algorithm

Let $\mathbf{A}_{1:N} = (\mathbf{A}_{1:N}^{(1)}, \ldots, \mathbf{A}_{1:N}^{(D)})$ denote the set of latent variables, i.e., the attribute profiles of the N students across D latent layers. The complete data log-likelihood is:

$$\begin{aligned} & l_c^{DeepCDM}(\boldsymbol{\beta}, \boldsymbol{\pi}^{(D)}, \mathbf{Q} | \mathbf{R}_{1:N}, \mathbf{A}_{1:N}) \\ = & \log \left(\mathbb{P}(\mathbf{R}_{1:N} \mid \mathbf{A}_{1:N}^{(1)}, \mathbf{Q}^{(1)}, \boldsymbol{\beta}^{(1)}) \cdot \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}_{1:N}^{(d-1)} \mid \mathbf{A}_{1:N}^{(d)}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) \cdot \mathbb{P}(\mathbf{A}_{1:N}^{(D)}; \boldsymbol{\pi}^{(D)}) \right). \end{aligned} \tag{12}$$

Let $\boldsymbol{\beta}^{(t-1)} = (\boldsymbol{\beta}^{(1,t-1)}, \ldots, \boldsymbol{\beta}^{(D,t-1)})$, $\boldsymbol{\pi}^{(D,t-1)}$, and $\mathbf{Q}^{(t-1)} = (\mathbf{Q}^{(1,t-1)}, \ldots, \mathbf{Q}^{(D,t-1)})$ denote the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\pi}^{(D)}$, and $\mathbf{Q}$ obtained at iteration $t-1$. In each iteration t of the EM algorithm, the following two steps are performed:

E-Step: Compute

$$\widetilde{Q}^{(t)} = \mathbb{E}_{(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)})}[l_c^{DeepCDM}(\boldsymbol{\beta}, \boldsymbol{\pi}^{(D)}, \mathbf{Q} | \mathbf{R}_{1:N}, \mathbf{A}_{1:N}) | \mathbf{R}_{1:N}; \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}^{(D,t-1)}, \mathbf{Q}^{(t-1)}], \tag{13}$$

where the conditional expectation is with respect to $\mathbb{P}(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)} | \mathbf{R}_{1:N}; \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}^{(D,t-1)}, \mathbf{Q}^{(t-1)})$.

M-Step: Update

$$\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(D,t)}, \mathbf{Q}^{(t)}\right) = \arg\max_{\boldsymbol{\beta}, \mathbf{Q}} \widetilde{Q}^{(t)} - N \cdot P_{\boldsymbol{s}}(\boldsymbol{\beta}), \tag{14}$$

where $P_{\boldsymbol{s}}(\boldsymbol{\beta})$ is defined in Equation (11).

For each $d = 1, 2, \ldots, D$, define $\widetilde{\mathbf{A}}^{1:d} = (\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(d)})$, which denote the latent variables shallower than the $(d+1)$th latent layer. Define $\mathbf{A}^{(0)} = \mathbf{R}_{1:N}$, and $\mathbf{A}^{(D+1)} = \emptyset$. According to the conditional independence, the expectation computation in the E-step can be re-expressed as $\widetilde{Q}^{(t)} = \sum_{d=1}^{D+1} \widetilde{Q}^{(d,t)}$, with

$$\begin{aligned} \widetilde{Q}^{(d,t)} = & \quad \mathbb{E}_{\widetilde{\mathbf{A}}^{1:d}}[\log P(\mathbf{A}_{1:N}^{(d-1)} | \mathbf{A}_{1:N}^{(d)}) | \mathbf{R}_{1:N}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}_D^{(t-1)}, \mathbf{Q}^{(t-1)}] \\ = & \sum_{i=1}^{N} \mathbb{E}_{\widetilde{\mathbf{A}}^{1:d}}[\log P(\mathbf{A}^{(d-1)} | \mathbf{A}^{(d)}) | \mathbf{R}_i, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}_D^{(t-1)}, \mathbf{Q}^{(t-1)}] \end{aligned} \tag{15}$$

That is, $\widetilde{Q}^{(t)}$ is decomposed as a summation over layers d and individuals i, where each term is the conditional expectation of $\log P(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)})$ with respect to the partial posterior

distribution $P(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(d)} \mid \mathbf{R}_i, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}_D^{(t-1)}, \mathbf{Q}^{(t-1)})$, which we denote by $\widetilde{P}_i^{1:d}$ for brevity.

Accordingly, the optimization in M-step can be broken into the following parts,

$$(\boldsymbol{\beta}^{(d,t)}, \mathbf{Q}^{(d,t)}) = \arg\max_{\boldsymbol{\beta}^{(d)}, \mathbf{Q}^{(d)}} \widetilde{Q}^{(d,t)} - N \cdot P_{\boldsymbol{s}}(\boldsymbol{\beta}^{(d)}), \quad d = 1, \ldots, D. \tag{16}$$

$$\begin{aligned} \boldsymbol{\pi}^{(D,t)} &= \arg\max \sum_{\boldsymbol{\alpha}^{(D)}} \sum_{i=1}^{N} \log(\mathbb{P}(\mathbf{A}^{(D)} = \boldsymbol{\alpha}^{(D)})) \times \\ &\sum_{(\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(D-1)})} P\left(\mathbf{A}_i^{(1)} = \boldsymbol{\alpha}^{(1)}, \mathbf{A}_i^{(2)} = \boldsymbol{\alpha}^{(2)}, \ldots, \mathbf{A}_i^{(D)} = \boldsymbol{\alpha}^{(D)} | \mathbf{R}_i; \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}_D^{(t-1)}, \mathbf{Q}^{(t-1)}\right) \end{aligned} \tag{17}$$

This decomposition enables the parameters at each layer to be updated via their corresponding optimization problems in (16) and (17), thereby improving the tractability of the M-step. Next, we further look into the $\widetilde{Q}^{(d,t)}$ functions. Recall that $\widetilde{P}_i^{1:d}$ is defined as:

$$\widetilde{P}_i^{1:d} := P(\widetilde{\mathbf{A}}^{1:d} \mid \mathbf{R}_i, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\pi}_D^{(t-1)}, \mathbf{Q}^{(t-1)}) = \frac{P(\mathbf{R}_i \mid \mathbf{A}^{(1)}) P(\widetilde{\mathbf{A}}^{1:d})}{\sum_{\widetilde{\mathbf{A}}^{1:d}} P(\mathbf{R}_i \mid \mathbf{A}^{(1)}) P(\widetilde{\mathbf{A}}^{1:d})}, \tag{18}$$

with

$$\begin{aligned} P(\widetilde{\mathbf{A}}^{1:d}) = \sum_{\boldsymbol{\alpha}^{(d+1)}, \ldots, \boldsymbol{\alpha}^{(D)}} & P\left(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}, \boldsymbol{\beta}^{(t-1)}\right) \times \\ & P\left(\mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}, \ldots, \mathbf{A}^{(D)} = \boldsymbol{\alpha}^{(D)}; \boldsymbol{\beta}^{(t-1)}\right). \end{aligned} \tag{19}$$

As shown, given the parameters from the previous iteration $(t-1)$, the distribution $P(\widetilde{\mathbf{A}}^{1:d})$ is computed by marginalizing out the deeper latent variables $\mathbf{A}^{(d+1)}, \ldots, \mathbf{A}^{(D)}$ from the joint distribution over all latent variables. This marginal, together with the observed response $\mathbf{R}_i$, allows us to evaluate the partial posterior distribution $\widetilde{P}_i^{1:d}$ via Equation (18). With $\widetilde{P}_i^{1:d}$ providing the weights for each possible configuration of $\widetilde{\mathbf{A}}^{1:d}$, the conditional expectation in $\widetilde{Q}^{(d,t)}$ can be computed as a weighted sum over $\log P(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)})$. Specifically, Equation (15) can be written out as:

$$\widetilde{Q}^{(1,t)} = \sum_{i=1}^{N} \sum_{\mathbf{A}^{(1)}} \log P(\mathbf{R}_i | \mathbf{A}^{(1)}) \widetilde{P}_i^{1:1}; \tag{20}$$

and

$$\widetilde{Q}^{(d,t)} = \sum_{\widetilde{\mathbf{A}}^{1:d}} \log P(\mathbf{A}^{(d-1)}|\mathbf{A}^{(d)}) \sum_{i=1}^{N} \widetilde{P}_i^{1:d} = \sum_{(\mathbf{A}^{(d-1)},\mathbf{A}^{(d)})} \log P(\mathbf{A}^{(d-1)}|\mathbf{A}^{(d)}) \sum_{i=1}^{N} \sum_{\widetilde{\mathbf{A}}^{1:d}\setminus\{\mathbf{A}^{(d-1)},\mathbf{A}^{(d)}\}} \widetilde{P}_i^{1:d}, \quad (21)$$

for $d = 2, \ldots, D$. It turns out that, for each d, Equation (16) defines a regularized optimization problem whose objective includes a weighted log-likelihood component over $(\mathbf{A}^{(d-1)}, \mathbf{A}^{(d)})$ pairs. In this formulation, $\mathbf{A}^{(d-1)}$ serves as the outcome, $\mathbf{A}^{(d)}$ as the predictor, and the data point weights are given by $\sum_{i=1}^{N} \sum_{\widetilde{\mathbf{A}}^{1:d}\setminus\left\{\mathbf{A}^{(d-1)},\mathbf{A}^{(d)}\right\}} \widetilde{P}_i^{1:d}$.

We focus on the case where the link function $f(\cdot)$ is the inverse logit, as it is the most commonly used choice for CDMs with binary responses. In this setting, the estimation problem corresponds to a generalized linear optimization problem with a logit link. For other choices of $f(\cdot)$, the problem may fall into the broader categories of linear or generalized linear optimization, depending on the specific functional form. The solution of Equation (17) is that for $\forall \boldsymbol{\alpha}^{(D)} \in \{0,1\}^{K_D}$:

$$\boldsymbol{\pi}_{\boldsymbol{\alpha}^{(D)}}^{(D,t)} = \sum_{i=1}^{N} \sum_{\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(D-1)}} \frac{P(\mathbf{R}_i|\mathbf{A}^{(1)})P(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(D-1)},\mathbf{A}^{(D)}=\boldsymbol{\alpha}^{(D)})}{N \cdot \sum_{\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(D)}} P(\mathbf{R}_i|\mathbf{A}^{(1)})P(\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(D-1)},\mathbf{A}^{(D)})}. \quad (22)$$

These derivations demonstrate that, due to conditional independence, the EM algorithm for exploratory DeepCDMs is both succinct and transparent. However, its practical feasibility is challenged by several issues—particularly sensitivity to initialization and the accumulation of estimation bias across layers and iterations. To address these challenges, we next propose a layer-wise EM algorithm below.

3.2 A Novel Layer-wise EM Algorithm

To elucidate the underlying rationale of the proposed algorithm, we first provide a detailed mathematical derivation of the layer-wise EM procedure. This step-by-step formulation highlights how the algorithm naturally arises from the hierarchical structure of DeepCDMs.

Suppose we have a set of parameters $\boldsymbol{\beta}, \boldsymbol{\pi}^{(D)}, \mathbf{Q}$ that maximize the regularized marginal log-likelihood in Equation (10). Based on the generative formulation of DeepCDMs, we can

marginalize out the deeper latent variables $\mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(D)}$ to derive the implied distribution of the bottom-layer attributes:

$$\boldsymbol{\pi}^{(1)} = \mathbb{P}(\mathbf{A}^{(1)}; \boldsymbol{\pi}^{(1)}) = \sum_{\boldsymbol{\alpha}^{(2)}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}} \Big[\prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)}, \boldsymbol{\beta}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}; \boldsymbol{\pi}^{(D)}) \Big].$$

This recursive marginalization induces a set of shallow-layer parameters $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\pi}^{(1)}, \mathbf{Q}^{(1)})$, which, when substituted into the original model, must also maximize a re-expressed form of the target function:

$$\ell(\boldsymbol{\beta}, \boldsymbol{\pi}^{(1)}, \mathbf{Q} \mid \mathbf{R}_{1:N}) = \sum_{i=1}^{N} \log \left\{ \sum_{\boldsymbol{\alpha}^{(1)}} \Big[\mathbb{P}(\mathbf{R}_i \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)}, \boldsymbol{\beta}^{(1)}) \cdot \mathbb{P}(\mathbf{A}^{(1)}; \boldsymbol{\pi}^{(1)}) \Big] \right\} - N \cdot P_{\boldsymbol{s}}(\boldsymbol{\beta}).$$

This observation forms the conceptual basis of our proposed layer-wise EM algorithm. Rather than estimating all parameters jointly over the entire deep latent architecture, we decompose the problem into a sequence of simpler subproblems, each involving a one-layer structure, and solve them in a bottom-up manner using EM. Although the resulting algorithm appears intuitive, it is grounded in a rigorous use of the model's generative structure. Specifically, each layer-wise step leverages the most reliable information from its immediate lower layer—either in the form of estimated distributions or pseudo-observations—making the estimation process both computationally efficient and statistically reliable.

Focusing on the first layer ($d = 1$), this decomposition implies that the estimates $\boldsymbol{\beta}^{(1)}$ and $\mathbf{Q}^{(1)}$ obtained by maximizing Equation (10) are identical to those obtained under a standard one-layer CDM. Once these first-layer parameters are estimated, the task reduces to estimating the parameters for the remaining $D-1$ latent layers. Unlike the first layer, however, there are no observed realizations of the latent variable $\mathbf{A}^{(1)}$. Let $\mathbf{A}_i^{(1)}$ denote the i-th row of $\mathbf{A}_{1:N}^{(1)}$. A straightforward way to impute $\mathbf{A}_i^{(1)}$ is via the Maximum A Posteriori (MAP) estimate: $\widehat{\mathbf{A}}_i^{(1)} = \arg\max_{\boldsymbol{\alpha}^{(1)} \in \{0,1\}^{K_1}} \mathbb{P}(\mathbf{A}_i^{(1)} = \boldsymbol{\alpha}^{(1)} \mid \mathbf{R}_i, \boldsymbol{\beta}^{(1)}, \mathbf{Q}^{(1)})$, which depends on both the likelihood $\mathbb{P}(\mathbf{R}_i \mid \mathbf{A}_i^{(1)}, \boldsymbol{\beta}^{(1)}, \mathbf{Q}^{(1)})$ and the prior $\mathbb{P}(\mathbf{A}_i^{(1)})$. Alternatively, one can sample from the estimated marginal distribution $\mathbb{P}(\mathbf{A}^{(1)})$ to generate pseudo-observations for the next layer. Compared to MAP, this sampling-based approach introduces less bias and leads to more reliable parameter estimation in deeper layers, particularly during initialization. With

these pseudo-observations in hand, the remaining $D-1$ layers form a structurally similar DeepCDM, and the same one-layer EM algorithm can be recursively applied to estimate $\boldsymbol{\beta}^{(2)}$ and $\mathbf{Q}^{(2)}$, and so on, until the top layer is reached.

The above layer-wise EM algorithm is summarized in Algorithm 1. Proceeding from the bottom up, the algorithm estimates model parameters layer by layer. For each layer $d > 1$, Step 1 imputes the missing data $\mathbf{A}^{(d-1)}$ by drawing samples from the estimated marginal distribution $P(\mathbf{A}^{(d-1)})$, which is computed in Step 3 of the previous layer using Equation (23). In contrast, for the first latent layer ($d = 1$), Step 1 is not required because the observed response data $\mathbf{R}_{1:N}$ (i.e., $\mathbf{A}^{(0)}$) serve directly as the input. Using the initial values obtained in Step 2, Step 3 then applies a one-layer EM algorithm with K_d attributes to estimate the parameters $\boldsymbol{\beta}^{(d)}$, $\mathbf{Q}^{(d)}$, and $\boldsymbol{\pi}^{(d)}$. The initialization procedure is introduced separately in Section 3.3. In Step 3, each M1-step is solved using the coordinate descent algorithm of Friedman et al. (2010), known for its flexibility and effectiveness in handling regularized optimization problems. The algorithm continues this layer-wise procedure until reaching the deepest layer D.

In M2-step, the mechanism for identifying the entries $q_{jk}^{(d)}$ in $\mathbf{Q}^{(d)}$ varies across different measurement models, according to the measurement model utilized in layer d. For *main-effect*-model, $\mathbf{Q}^{(d)}$ can be recovered using the rule $q_{jk}^{(d)} = \mathbf{1}(\beta_{j,k}^{(d)} \neq 0)$, where $\mathbf{1}$ is the indicator function. For *all-effect* model, theoretically, each row of $\mathbf{Q}^{(d)}$ should be identified by the highest-order non-zero coefficient. Specifically, if $\exists S \subseteq [K]$ such that $\beta_{j,S}^{(d)} \neq 0$ and $\beta_{j,S'}^{(d)} = 0$ for all $S' \subseteq [K], S \subset S'$, then $q_{jk}^{(d)} = 1$ for $k \in S$; otherwise, $q_{jk}^{(d)} = 0$. However, this strict rule may not be always applicable because some estimated β-coefficients may be close to zero but not exactly zero. In practice, a more effective approach is either to choose the largest non-zero interaction coefficient or to truncate the coefficients before identifying $\mathbf{Q}^{(d)}$. For the latter approach, we recommend practitioners set the truncation thresholds based on the general magnitude of their estimated coefficients. For the *DINA model*, since there should be only one non-zero coefficient for each item j, the largest non-zero interaction coefficient can be selected, and the corresponding $q_{jk}^{(d)}$ can be identified as equal to one.

Algorithm 1 Layer-wise EM Algorithm for the DeepCDMs

Input: Response matrix $\mathbf{R}_{1:N}$, number of layers D, number of attributes K_d for each layer $d = 1, \ldots, D$. Set $\mathbf{A}^{(0)} = \mathbf{R}_{1:N}$.
For each layer $d = 1, \ldots, D$**, do:**
1. Impute data: If $d > 1$, draw N samples of $\mathbf{A}^{(d-1)}$ from the sample space $\{0,1\}^{2^{K_{d-1}}}$ according to $\boldsymbol{\pi}^{(d-1)}$ obtained in the $d-1$-th step. These samples will be imputed as data for $\mathbf{A}^{(d-1)}$ in the following calculations.
2. Initialization: Using $\mathbf{A}^{(d-1)}$ as input data, apply the USVT-based estimator in Algorithm 2 to obtain the initial values of $\boldsymbol{\beta}^{(d)}$.
3. EM algorithm for the d**-th layer:** Repeat the following steps until convergence (starting from $t_d = 1$):

- **E-Step:** Compute $\widetilde{Q}^{(d,t_d)}$ as defined in Equation (15) using Equations (18)-(21).
- **M-Step:**

  M1. Apply the coordinate descent algorithm to solve the optimization problem defined in Equations (16), yielding updated estimates $\boldsymbol{\beta}^{(d,t_d)}$.

  M2. Estimate $\mathbf{Q}^{(d)}$ as $\mathbf{Q}^{(d,t_d)}$ by identifying the sparse pattern in $\boldsymbol{\beta}^{(d,t_d)}$.

  M3. For each $\boldsymbol{\alpha}^{(d)} \in \{0,1\}^{K_d}$, update entries in $\boldsymbol{\pi}^{(d)}$ as follows:

$$\boldsymbol{\pi}^{(d,t_d)}_{\boldsymbol{\alpha}^{(d)}} = \sum_{i=1}^{N} \frac{P(\mathbf{A}^{(d-1)}|\mathbf{A}^{(d)} = \boldsymbol{\alpha}^{(d)})\boldsymbol{\pi}^{(d,t_{d-1})}_{\boldsymbol{\alpha}^{(d)}}}{N \cdot \sum_{\mathbf{A}^{(d)}} P(\mathbf{A}^{(d-1)}|\mathbf{A}^{(d)})\boldsymbol{\pi}^{(d,t_{d-1})}_{\boldsymbol{\alpha}^{(d)}}}. \tag{23}$$

Output: The updated parameters $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\beta}^{(1,t_1)}, \ldots, \boldsymbol{\beta}^{(D,t_D)})$, $\widehat{\mathbf{Q}} = (\mathbf{Q}^{(1,t_1)}, \ldots, \mathbf{Q}^{(D,t_D)})$, and $\widehat{\boldsymbol{\pi}} = (\widehat{\boldsymbol{\pi}}^{(1,t_1)}, \ldots, \widehat{\boldsymbol{\pi}}^{(D,t_D)})$.

3.3 Initialization Algorithm

As shown in Algorithm 1, the initialization of our DeepCDM method is performed sequentially in D steps, where parameters from the d-th latent layer is initialized after estimating the $(d-1)$-th layer. This approach leverages the information from the estimated distribution $P(\mathbf{A}^{(d-1)})$, obtained during the fitting of the $(d-1)$-th layer, to provide better initial values for the d-th layer. Furthermore, by imputing realizations of $\mathbf{A}^{(d-1)}$ sampled from the estimated $P(\mathbf{A}^{(d-1)})$, the initializations of all laysers reduce to the problem of initializing a one-layer CDM. Specifically, the response data $\mathbf{R}_{1:N}$ is used for the first layer ($d = 1$), while realizations of $\mathbf{A}^{(d)}$ (i.e., the generated pseudo-samples) are used for subsequent layers

($d > 1$). For exploratory DeepCDMs, good initial values should not only be close to the true values but also exhibit a sparse structure similar to the true ones. To achieve this, we apply a method based on USVT (Chatterjee, 2015; Zhang et al., 2020) to estimate the design matrix, followed by a Varimax rotation to promote sparsity. USVT captures the dominant low-rank structure, while Varimax produces a sparse loading pattern that informs the initial $\mathbf{Q}$-matrix. This combination provides informative starting values tailored to the exploratory framework of DeepCDMs. The initialization procedure for each layer d is detailed in Algorithm 2.

Algorithm 2 applies SVD twice. The first application, combined with the inverse transformation (Steps 2-5), is used to denoise and linearize the data. The second application of SVD (Steps 6-7) is performed on the linearized data. Since the loading matrix can only be recovered up to an oblique rotation, analytical methods are needed to resolve the rotational indeterminacy. To address this, we apply Varimax rotation (Kaiser, 1958) in Step 8, a widely used method for obtaining interpretable solutions that has also been justified for enabling statistical inference (Rohe and Zeng, 2023). More importantly, in our context, Varimax is particularly crucial for recovering sparse coefficient structures consistent with the model's identifiability conditions. Note that SVD does not directly recover the binary latent attribute structure. As a result, while it effectively captures the sparse pattern encoded in $\widehat{\mathbf{G}}^1$ (Step 9), the estimated loading matrix $\widetilde{\mathbf{V}}$ may differ from the true one in scale. This issue is addressed by exploiting the discreteness of the latent variables in $\widehat{\mathbf{A}}^{(d)}$ to rescale $\boldsymbol{\beta}^{(d)}$ (Lee and Gu, 2025), as shown in Steps 10-11. This initialization procedure is non-iterative, making it computationally efficient. Moreover, it avoids convergence issues and possesses favorable statistical consistency properties (Zhang et al., 2020).

3.4 Advantages of layer-wise EM compared to EM

As discussed in Hinton et al. (2006), an effective way to learn complex models is by sequentially combining simpler models. The layer-wise EM algorithm applies this principle by breaking down the optimization process into manageable subproblems, each targeting one layer at a time. This strategy helps overcome the key challenges faced by the classical EM algorithm when applied to DeepCDMs.

One major challenge in applying the classical EM algorithm to DeepCDMs is the dif-

Algorithm 2 USVT-Based Initialization Procedure for Layer d

1. Input: The $N \times K_{d-1}$ data matrix $\mathbf{A}^{(d-1)}$, the number of attributes K_d, the function $f(\cdot)$, and a truncation parameter $\epsilon_{N,K_{d-1}} > 0$.

2. Apply SVD to $\mathbf{A}^{(d-1)} = \sum_{j=1}^{K_{d-1}} \tau_j \boldsymbol{u}_j \boldsymbol{v}_j{}^\top$, where $\tau_1 \geq \tau_2 \geq \ldots \tau_{K_{d-1}}$ are the singular values, and $\boldsymbol{u}_j$s and $\boldsymbol{v}_j$s are left and right singular vectors, respectively.

3. Let $\mathbf{X} = (x_{ij})_{N \times K_{d-1}} = \sum_{k=1}^{\widetilde{K}_d} \tau_k \boldsymbol{u}_k \boldsymbol{v}_k{}^\top$, where $\widetilde{K}_d = \max\left\{K_d + 1, \arg\max_{\mathrm{k}}\left\{\tau_{\mathrm{k}} \geq 1.01\sqrt{\mathrm{N}}\right\}\right\}$.

4. Let $\widehat{\mathbf{X}} = (\widehat{x}_{ij})_{N \times K_{d-1}}$ be defined as

$$\widehat{x}_{ij} = \begin{cases} \epsilon_{N,K_{d-1}} & if \quad x_{ij} < \epsilon_{N,K_{d-1}} \\ x_{ij} & if \quad \epsilon_{N,K_{d-1}} \leq x_{ij} \leq 1 - \epsilon_{N,K_{d-1}} \\ 1 - \epsilon_{N,K_{d-1}} & if \quad x_{ij} \geq 1 - \epsilon_{N,K_{d-1}} \end{cases}$$

5. Let $\widetilde{\mathbf{M}} = (\widetilde{m}_{ij})_{N \times K_{d-1}}$, where $\widetilde{m}_{ij} = f(\widehat{x}_{ij})$.

6. Let $\widehat{\boldsymbol{\beta}}_0^{(d)} = (\widehat{\beta}_{1,0}^{(d)}, \ldots, \widehat{\beta}_{j,0}^{(d)}, \ldots, \widehat{\beta}_{K_{d-1},0}^{(d)})$, where $\widehat{\beta}_{j,0}^{(d)} = (\sum_{i=1}^{N} \widetilde{m}_{ij})/N$, $j = 1, \ldots, K_{d-1}$.

7. Apply singular value decomposition to $\widehat{\mathbf{M}} = (\widetilde{m}_{ij} - \widehat{\beta}_{j,0}^{(d)})_{N \times K_{d-1}}$ to have $\widehat{\mathbf{M}} = \sum_{j=1}^{K_{d-1}} \widehat{\tau}_j \widehat{\boldsymbol{u}}_j \widehat{\boldsymbol{v}}_j$, where $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \ldots \widehat{\tau}_{K_{d-1}}$ are the singular values, and $\widehat{\boldsymbol{u}}_j$s and $\widehat{\boldsymbol{v}}_j$s are left and right singular vectors, respectively.

8. Apply varimax to $\widehat{\mathbf{V}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_{K_{d-1}})$, and let $\widetilde{\mathbf{V}}$ be the rotated version of $\widehat{\mathbf{V}}$.

9. Threshold $\widetilde{\mathbf{V}}$ at $\frac{1}{2\sqrt{K_{d-1}}}$ to induce sparsity, and also adjust the column-wise sign flip and rotation. Let $\widehat{G}^1$ be the estimated sparsity pattern.

10. Estimate $\mathbf{A}^0$ by $\widehat{\mathbf{A}}^0 := \widehat{\mathbf{M}}\widetilde{\mathbf{V}}(\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}})^{-1}$. Use this to estimate $\widehat{\mathbf{A}}^{(d)} = (\widehat{A}_{i,k}^{(d)})_{N \times K_d}$: $\widehat{A}_{i,k}^{(d)} = 1(A_{i,k}^0 > 0)$.

11. Estimate $\boldsymbol{\beta}_1{}^{(d)}$ by $\widehat{\boldsymbol{\beta}}_1^{(d)} = \left(\gamma_{\text{scale}} \cdot \left(\widehat{\mathbf{A}}^{(d)\top}\widehat{\mathbf{A}}^{(d)}\right)^{-1} \widehat{\mathbf{A}}^{(d)\top}\widehat{\mathbf{M}}\right) \widehat{\mathbf{G}}^1$. Here, γ_{scale} is an artificial shrinkage factor we introduce to adjust the scale of $\widehat{\boldsymbol{\beta}}_1^{(d)}$, and $\cdot$ is the element-wise product.

12. Output $\widehat{\boldsymbol{\beta}}^{(d)} = (\widehat{\boldsymbol{\beta}}_0^{(d)}, \widehat{\boldsymbol{\beta}}_1^{(d)})$.

ficulty of initialization. The presence of multiple nonlinear latent layers creates a highly nonconvex optimization landscape, often with an exponential number of local optima. As a result, EM is sensitive to starting values—poor initialization can easily lead to convergence at suboptimal local maxima of the penalized log-likelihood function. In the classical EM algorithm, initial values for all parameters across all layers must be specified simultaneously. As the model depth increases, this becomes increasingly challenging, even for moderately deep architectures, due to compounded uncertainty and parameter interactions across layers. In contrast, our layer-wise EM algorithm addresses initialization sequentially by solving one-layer CDMs one at a time. At each stage, it initializes the parameters of the current layer using either observed responses or sampled latent attributes from the estimated marginal distribution informed by the shallower layers. Since these samples are based on estimated proportion parameters which are proved to be identifiable, the initialization for deeper layers is more stable and reliable.

Another major issue of the classical EM algorithm is the accumulation of estimation bias as the algorithm progresses through multiple layers. As the number of latent layers D increases, the computation of quantities like $\sum_{i=1}^{N} \widetilde{P}_i^{1:d}$ for $d = 1, \ldots, D$ becomes more error-prone. These quantities play a crucial role in the M-step, and inaccuracies in their estimation directly affect parameter updates. Furthermore, the iterative nature of the classical EM algorithm introduces a cyclic dependency across all layers, where errors at one layer can propagate forward and backward, reinforcing one another over iterations. This compounding effect often leads the algorithm to converge to suboptimal local maxima, even when reasonably good initial values are provided. In contrast, our proposed layer-wise EM algorithm mitigates this issue by breaking the dependency cycle. Parameters are estimated sequentially from the bottom layer up, so each layer d is only influenced by the estimates from shallower layers $d' < d$. Although some bias may still accumulate, it originates solely from estimation errors in previous layers—not from compounded initialization errors across all layers. This directional, non-cyclic structure significantly reduces the accumulation of error and enhances the robustness of the estimation process.

3.5 Connections to Broader Principles and Algorithms

3.5.1 Connections between the Layer-wise EM and Identifiability

Interestingly, the derivation of the layer-wise EM algorithm aligns with and supports the technical insights in the identifiability proofs of DeepCDMs. The identifiability results in the Supplementary Material and in Gu (2024) demonstrate that identifiability of a DeepCDM can be examined and established in a layer-by-layer manner, proceeding from the bottom up. This follows from the probabilistic formulation of the directed graphical model and the discrete nature of the latent layers. For each layer $d \in \{0, 1, \ldots, D-1\}$, as long as $\mathbf{Q}^{(d+1)}$ satisfies the corresponding identifiability conditions for the one-layer CDM implied by

$$\mathbb{P}(\mathbf{A}^{(d)}) = \sum_{\boldsymbol{\alpha}^{(d+1)} \in \{0,1\}^{K_{d+1}}} \mathbb{P}(\mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}, \mathbf{Q}^{(d+1)}, \boldsymbol{\theta}^{(d+1)}) \cdot \mathbb{P}(\mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}), \quad (24)$$

the parameter set $(\boldsymbol{\beta}^{(d+1)}, \mathbf{Q}^{(d+1)}, \boldsymbol{\pi}^{(d+1)})$ is identifiable. The identifiability of the marginal distribution of $\mathbf{A}^{(d+1)}$ (i.e., $\boldsymbol{\pi}^{(d+1)}$) allows it to be treated theoretically as if it were observed when examining the identifiability for $\mathbf{Q}^{(d+2)}$, $\boldsymbol{\beta}^{(d+2)}$, and the marginal distribution of $\mathbf{A}^{(d+2)}$. Starting from the observed data layer ($d = 0$) and proceeding one layer at a time, the identifiability of DeepCDMs can thus be inductively established.

This layer-wise identifiability rationale supports the procedure of generating samples of $\mathbf{A}^{(d+1)}$ when estimating parameters of the d-th layer, for each d. Specifically, the identifiability of the model at layer d, as shown in Equation (24), implies that the distribution $\mathbb{P}(\mathbf{A}^{(d+1)})$, from which samples of $\mathbf{A}^{(d+1)}$ are drawn, can be uniquely identified. This theoretical guarantee justifies treating the sampled $\mathbf{A}^{(d+1)}$ as observed data in the EM algorithm. The same procedure is applied recursively to the remaining $D-1$ layers.

3.5.2 Connections to Related Algorithms for DGMs

The idea of training deep models in a layer-wise fashion has been widely explored across the deep generative modeling literature, and our layer-wise EM algorithm draws on this foundational principle. For instance, in DBNs, greedy layer-wise pretraining trains each layer locally as a RBM using contrastive divergence (Hinton et al., 2006). This approach improves opti-

mization stability and scalability, particularly in deep models where global joint training is challenging. While DeepCDMs differ from DBNs in having a fully directed architecture and an emphasis on interpretability and identifiability, our method similarly decomposes model training into a sequence of tractable subproblems. DEFs (Ranganath et al., 2015) also adopt a hierarchical top-down structure and rely on recursive variational inference across layers. Although our inference uses EM rather than variational methods, both approaches share the advantage of progressing layer by layer, with each layer conditioned on information derived from the previous one. Our idea also resonates with the framework of modular Bayesian learning (Segal et al., 2005; Joshi et al., 2009), in which large models are decomposed into smaller, interpretable modules that can be estimated sequentially. By aligning with these broader principles, the proposed layer-wise EM adapts a widely used idea—local, modular, progressive learning—for the structured and interpretable setting of cognitive diagnosis, where both identifiability and individualized feedback are essential.

3.6 Extending to Confirmatory DeepCDMs

The layer-wise EM algorithm can be easily modified to fit confirmatory DeepCDMs, which assume all $\mathbf{Q}$-matrices are known. In this case, each M-step solves the following optimization:

$$\boldsymbol{\beta}^{(d,t)} = \arg\max_{\boldsymbol{\beta}^{(d)}} \widetilde{Q}^{(d,t)}, \quad d = 1, \ldots, D \tag{25}$$

Compared to Equation (16), the terms $N \cdot P_{\boldsymbol{s}}(\boldsymbol{\beta}^{(d)})$ are dropped, as all the $\mathbf{Q}$-matrices are known and do not need to be estimated. This makes the confirmatory case simpler than the exploratory case we have focused on. The layer-wise algorithm in Algorithm 1 can be readily used for parameter estimation, incorporating the known $\mathbf{Q}$-matrices during the implementation of coordinate descent in the M-step.

The adaptation of the layer-wise EM algorithm for the confirmatory case is also a novel contribution, representing the first EM-type algorithm for confirmatory DeepCDMs. Unlike the Bayesian approach in Gu (2024), which fits models using an MCMC algorithm that can involve slower convergence and the need for prior specification, the layer-wise EM algorithm offers a computationally more efficient alternative by directly seeking the maximum

likelihood estimator. We point out that the layer-wise EM method could also be extended to incorporate prior distributions for computing maximum posterior estimates, should rich prior information be available, following the approach outlined in Gu (2024).

4 Simulation Studies

In this section, we conduct simulation studies to evaluate the performance of the proposed layer-wise EM algorithm for exploratory DeepCDMs. Specifically, we consider a three-layer DeepCDM ($D = 3$) with the configuration $(J, K_1, K_2, K_3) = (30, 8, 4, 2)$, which represents a challenging scenario due to the large depth with three latent layers and the need to estimate all three unknown $\mathbf{Q}$-matrices across different layers. Three different measurement models are considered for the shallowest layer ($d = 1$): the main-effect model, the all-effect model, and the DINA model. The more parsimonious main-effect model is used to model the two deeper layers ($d = 2, 3$). We denote these three simulation cases as the Main-effect case, All-effect case, and DINA case, respectively. Under each case, three sample sizes—$N = 1000$, 1500, and 2000—are considered. The true $\mathbf{Q}$-matrices are specified in Equation (26) and satisfy the minimal identifiability conditions required for DeepCDMs.

$$\mathbf{Q}^{(1)}_{30\times 8} = \begin{pmatrix} & & & \mathrm{I}_8 & & & & \\ & & & \mathrm{I}_8 & & & & \\ & & & \mathrm{I}_8 & & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{Q}^{(2)}_{8\times 4} = \begin{pmatrix} \mathrm{I}_4 \\ \mathrm{I}_4 \end{pmatrix}, \quad \mathbf{Q}^{(3)}_{4\times 2} = \begin{pmatrix} \mathrm{I}_2 \\ \mathrm{I}_2 \end{pmatrix}. \tag{26}$$

The USVT-based estimator described in Algorithm 2 is used for initialization. The coordinate descent algorithm is implemented using the R package *glmnet* (Hastie et al., 2021) for optimization. The true coefficient values of $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, and $\boldsymbol{\beta}^{(3)}$ are presented in the following subsections for each simulation case, where $\boldsymbol{\beta}^{(1)}$ is set to be larger than those of the two deeper layers, $\boldsymbol{\beta}^{(2)}$ and $\boldsymbol{\beta}^{(3)}$. In CDM, larger coefficient values indicate a stronger relationship

between latent attributes and responses. Therefore, by specifying smaller coefficients for the deeper layers, we introduce more randomness into this relationship. This design reflects the increased uncertainty and complexity involved in mastering higher-level cognitive attributes in psychological or educational assessments. For each d-th layer ($d = 1, 2, 3$), the layer-wise EM algorithm uses the convergence criterion $\max \left|\boldsymbol{\beta}^{(d,t)} - \boldsymbol{\beta}^{(d,t-1)}\right| < \epsilon_d$ over three successive iterations. In this simulation, the thresholds are set as $(\epsilon_1, \epsilon_2, \epsilon_3) = (4^{-2}, 10^{-3}, 10^{-3})$. A more relaxed threshold is used for the first layer to accommodate its relatively larger number of parameters, as it involves both a larger size of $\mathbf{Q}$-matrix ($\mathbf{Q}^{(1)}_{30\times 8}$) and larger coefficient values in $\boldsymbol{\beta}^{(1)}$ compared to the other two deeper layers. This setting ensures stable convergence while maintaining computational efficiency in high-dimensional estimation. For each simulation case, 100 independent replications are conducted. In each replication, the layer-wise EM algorithm is applied to fit the model across a range of regularization parameters, and the one that yields the smallest BIC value is selected for final model fitting. The specific regularization sequences are provided in the Supplementary Material. Root Mean Squared Errors (RMSE) and absolute biases (aBias) are computed to assess estimation accuracy.

Note that the true parameter values differ in magnitude across the three layers, making the RMSE and aBias values not directly comparable. To address this, we report RMSE and aBias for two additional metrics that evaluate the performance of the layer-wise EM algorithm from different perspectives and provide indices that are comparable across layers. The first index is the latent class proportion distribution $\boldsymbol{\pi}^{(d)}$, whose parameter space is defined as $\Delta^{2^{K_d}-1} = \left\{\pi^{(d)}_{\boldsymbol{\alpha}_\ell} : \sum_{\ell=1}^{2^{K_d}} \pi^{(d)}_{\boldsymbol{\alpha}_\ell} = 1,\ \pi^{(d)}_{\boldsymbol{\alpha}_\ell} > 0\right\}$, where $\pi^{(d)}_{\boldsymbol{\alpha}_\ell} = P(\mathbf{A}^{(d)} = \boldsymbol{\alpha}_\ell)$ and $\boldsymbol{\alpha}_\ell \in \{0,1\}^{K_d}$ for each layer $d = 1, \ldots, D$. This metric assesses how accurately the model recovers the true distribution of latent attributes at each layer. The second index is the correct response probability for each layer d, denoted as $P^{(d)}_{CR}$, and defined by Equations (7)–(9) according to the measurement model employed. This metric evaluates how well the model predicts correct responses at the population level for each layer. In the case of a single latent layer, this probability is commonly represented in the literature as $\theta_{j,\boldsymbol{\alpha}} = P(R_j = 1 \mid \mathbf{A} = \boldsymbol{\alpha})$.

4.1 Simulation Studies for Main-Effect DeepCDMs

Throughout this section, let k_{d-1} and k_d be the indices for the $(d-1)$-th and d-th layers, respectively. For the main-effect models, the coefficients $\beta^{(d)}_{k_{d-1},k_d}$ are specified as:

$$\beta^{(d)}_{k_{d-1},0} = c^{(d)}_0, \quad \beta^{(d)}_{k_{d-1},k_d} = \frac{c^{(d)}_1}{\sum_{k_d=1}^{K_d} q^{(d)}_{k_{d-1},k_d}}, \quad \forall\, k_d \in [K_d],\ d = 1,2,3, \tag{27}$$

where $K_0 = J$, and the constants $(c^{(d)}_0, c^{(d)}_1)$ are set to $(6,-3)$, $(3,-1.5)$, and $(3,-1.5)$ for $d = 1,2,3$, respectively. The RMSE and aBias are reported in Table 1. All indices decrease as the sample size increases, indicating improved estimation accuracy. For each sample size, the estimation accuracy is lower in the deeper layers than in the shallower layers, as reflected by the values of $\boldsymbol{\pi}^{(d)}$ and $P^{(d)}_{CR}$. This result is intuitively reasonable, as estimating deeper layers is fundamentally more challenging due to stochastic latent layers. In addition, all RMSE and aBias values remain reasonably small, providing empirical support for the identifiability of the model.

Measurement Model	N	Layer (d)	RMSE			aBias		
			$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P^{(d)}_{CR}$	$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P^{(d)}_{CR}$
Main-effect	1000	1	0.408	0.0190	0.022	0.355	0.0015	0.015
		2	0.189	0.0220	0.052	0.127	0.0170	0.034
		3	0.459	0.0690	0.075	0.319	0.0540	0.056
			Computation time (min): 4.13			Iterations: 78.8		
	1500	1	0.338	0.0016	0.021	0.334	0.0013	0.015
		2	0.186	0.0205	0.047	0.121	0.0150	0.034
		3	0.416	0.0510	0.074	0.306	0.0385	0.052
			Computation time (min): 4.48			Iterations: 75.86		
	2000	1	0.302	0.0014	0.015	0.297	0.0011	0.013
		2	0.179	0.0170	0.044	0.113	0.0138	0.030
		3	0.372	0.0490	0.055	0.243	0.0385	0.042
			Computation time (min): 13.67			Iterations: 71.81		

Table 1: RMSE and aBias for the Main-Effect DeepCDM

To examine the recovery of the $\mathbf{Q}$-matrices, we report the proportion of correctly estimated rows and entries in each $\mathbf{Q}^{(d)}$ for $d = 1,2,3$, as shown in Table 2. Note that these indices are not directly comparable across layers at each sample size, as the proportions depend on both the size of the $\mathbf{Q}$-matrix and the difficulty of parameter estimation. Due to the shrinkage ladder structure, which is supported by the identifiability conditions, the

shallower layers contain larger $\mathbf{Q}$-matrices, making their structures more challenging to recover. However, these layers also benefit from more informative signals, as they are closer to the observed data layer. In contrast, the deeper layers involve smaller $\mathbf{Q}$-matrices that are structurally easier to recover, but they suffer from less informative signals due to the greater amount of uncertainty introduced at deeper levels. Despite these differences, when comparing $\mathbf{Q}$-matrix recovery across sample sizes, it is evident that for each layer d, the estimation accuracy of $\mathbf{Q}^{(d)}$ improves as the sample size increases.

N	Proportion	Layer 1	Layer 2	Layer 3
1000	$P_{\text{Row-wise}}$	0.490	0.697	0.829
	$P_{\text{Entry-wise}}$	0.932	0.956	0.901
1500	$P_{\text{Row-wise}}$	0.548	0.704	0.820
	$P_{\text{Entry-wise}}$	0.945	0.957	0.931
2000	$P_{\text{Row-wise}}$	0.582	0.783	0.845
	$P_{\text{Entry-wise}}$	0.947	0.968	0.935

Table 2: Proportion of Correctly Recovered Rows ($P_{\text{Row-wise}}$) and Entries ($P_{\text{Entry-wise}}$) for the Main-effect DeepCDM

4.2 Simulation Studies for All-Effect DeepCDMs

Denote the true coefficients as $\beta^{(d)}_{k_{d-1},S_d}$, $\forall S_d \subseteq [K_d]\backslash\emptyset$ and $\beta^{(d)}_{k_{d-1},0} = c_0^{(d)}$. In the all-effect DeepCDMs case, denote $\mathcal{K}_{k_{d-1}} = \left\{k_d \in [K_d];\, q^{(d)}_{k_{d-1},k_d} = 1\right\}$, and $K_0 = J$, and the first layer ($d = 1$) is modeled using the all-effect model with the parameters given below:

$$\beta^{(d)}_{k_{d-1},S_d} = \begin{cases} \frac{c_1^{(d)}}{2^{|\mathcal{K}_{k_{d-1}}|}} & \prod_{l\in S_d} q^{(d)}_{k_{d-1},l} = 1 \\ 0 & \prod_{l\in S_d} q^{(d)}_{k_{d-1},l} = 0 \end{cases}$$

The two deeper layers ($d = 2, 3$) are then modeled by the main-effect model, with parameters specified according to Equation (27). The constants $(c_0^{(d)}, c_1^{(d)})$ are specified as $(c_0^{(1)}, c_1^{(1)}) = (6, -3)$, $(c_0^{(2)}, c_1^{(2)}) = (3, -1.5)$, and $(c_0^{(3)}, c_1^{(3)}) = (3, -1.5)$.

The RMSE and aBias values are summarized in Table 3. As expected, all indices decrease with increasing sample sizes, indicating improved estimation accuracy. RMSE and aBias values remain reasonably small across all layers, providing empirical support for the

identifiability of the model. To evaluate the recovery of the $\mathbf{Q}$-matrices, we report the proportions of correctly estimated rows and entries in each $\mathbf{Q}^{(d)}$ for $d = 1, 2, 3$, as shown in Table 4. As in previous cases, these proportions are not directly comparable across layers, as they are influenced by differences in $\mathbf{Q}$-matrix size and estimation difficulty. Nevertheless, for each layer d, the estimation accuracy of $\mathbf{Q}^{(d)}$ improves as the sample size increases.

Measurement Model	N	Layer (d)	RMSE			aBias		
			$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P_{CR}^{(d)}$	$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P_{CR}^{(d)}$
All-effect	1000	1	0.486	0.0018	0.025	0.432	0.0015	0.020
		2	0.197	0.0250	0.047	0.132	0.0198	0.037
		3	0.660	0.0800	0.095	0.427	0.0600	0.070
			Computation time (min): 32.00			Iterations: 75.78		
	1500	1	0.428	0.0015	0.021	0.386	0.0012	0.016
		2	0.191	0.0216	0.044	0.128	0.0175	0.034
		3	0.451	0.0700	0.073	0.326	0.0520	0.058
			Computation time (min): 25.74			Iterations: 80.15		
	2000	1	0.389	0.0013	0.017	0.353	0.0010	0.014
		2	0.144	0.0160	0.033	0.104	0.0130	0.026
		3	0.300	0.0440	0.051	0.205	0.0350	0.039
			Computation time (min): 63.37			Iterations: 68.76		

Table 3: RMSE and aBias for the All-Effect DeepCDM

N	Proportion	Layer 1	Layer 2	Layer 3
1000	$P_{\text{Row-wise}}$	0.808	0.662	0.745
	$P_{\text{Entry-wise}}$	0.986	0.942	0.867
1500	$P_{\text{Row-wise}}$	0.852	0.697	0.750
	$P_{\text{Entry-wise}}$	0.990	0.947	0.888
2000	$P_{\text{Row-wise}}$	0.878	0.757	0.885
	$P_{\text{Entry-wise}}$	0.993	0.969	0.954

Table 4: Proportion of Correctly Recovered Rows ($P_{\text{Row-wise}}$) and Entries ($P_{\text{Entry-wise}}$) for the All-effect DeepCDM

4.3 Simulation Studies for DINA-Effect DeepCDMs

Denote the true coefficients as $\beta^{(d)}_{k_{d-1},S_d}$ for all $S_d \subseteq [K_d] \setminus \emptyset$, with $\beta^{(d)}_{k_{d-1},0} = c_0^{(d)}$. In DINA DeepCDMs, the first layer ($d = 1$) is modeled using the DINA formulation where $\beta_{k_{d-1},S_d} = c_1^{(d)}$ if $S_d = \mathcal{K}_{k_{d-1}}$, and zero otherwise, with $\mathcal{K}_{k_{d-1}}$ defined in Section 4.2. The DINA model

can be viewed as a special case of the all-effect model, in which non-zero coefficients are assigned only to the interaction of all attributes required by the $(d-1)$-th unit in each d-th layer model. This formulation allows the same estimation framework used for the all-effect model to be applied to the DINA model. The two deeper layers ($d = 2, 3$) are modeled using the main-effect specification, with parameters defined as in Equation (27). The constants $(c_0^{(d)}, c_1^{(d)})$ are specified as $(6, -3)$, $(3, -1.5)$, and $(3, -1.5)$ for $d = 1, 2, 3$, respectively.

The RMSE and aBias results are reported in Table 5. Again, these values exhibit a clear decreasing trend with increasing sample size, indicating improved estimation accuracy. To assess the recovery of the $\mathbf{Q}$-matrices, Table 6 presents the proportions of correctly estimated rows and entries for each $\mathbf{Q}^{(d)}$, $d = 1, 2, 3$. Overall, for all layers, the estimation accuracy of $\mathbf{Q}^{(d)}$ improves as the sample size increases.

Measurement Model	N	Layer (d)	RMSE			aBias		
			$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P_{CR}^{(d)}$	$\boldsymbol{\beta}^{(d)}$	$\pi^{(d)}$	$P_{CR}^{(d)}$
DINA	1000	1	0.436	0.0018	0.022	0.412	0.0015	0.017
		2	0.201	0.0280	0.048	0.136	0.0190	0.037
		3	0.636	0.0860	0.094	0.429	0.0650	0.070
			Computation time (min): 42.92			Iterations: 83.83		
	1500	1	0.428	0.0015	0.021	0.389	0.0012	0.016
		2	0.192	0.0220	0.044	0.128	0.0175	0.034
		3	0.451	0.0701	0.073	0.326	0.0520	0.059
			Computation time (min): 25.73			Iterations:80.14		
	2000	1	0.389	0.0013	0.017	0.353	0.0010	0.014
		2	0.144	0.0162	0.033	0.120	0.0130	0.026
		3	0.300	0.0437	0.051	0.210	0.0345	0.039
			Computation time (min): 62.37			Iterations: 68.76		

Table 5: RMSE and aBias for the DINA DeepCDM

5 Real Data Analysis

To demonstrate the applicability of the exploratory DeepCDM, we analyze student response data from the TIMSS 2019 assessment using a two-layer DeepCDM. Specifically, we examine responses from $N = 1595$ eighth-grade students in the United Arab Emirates who completed Booklet No.1, which includes both mathematics and science items. The dataset comprises responses to $J = 54$ items. Responses were preprocessed into binary indicators of correctness:

N	Proportion	Layer 1	Layer 2	Layer 3
1000	$P_{\text{Row-wise}}$	0.709	0.689	0.758
	$P_{\text{Entry-wise}}$	0.957	0.951	0.871
1500	$P_{\text{Row-wise}}$	0.852	0.697	0.760
	$P_{\text{Entry-wise}}$	0.990	0.956	0.888
2000	$P_{\text{Row-wise}}$	0.878	0.757	0.885
	$P_{\text{Entry-wise}}$	0.993	0.969	0.954

Table 6: Proportion of Correctly Recovered Rows ($P_{\text{Row-wise}}$) and Entries ($P_{\text{Entry-wise}}$) for the DINA DeepCDM

multiple-choice responses were coded as 1 if correct and 0 otherwise; constructed responses were coded as 1 only if they received the maximum score, and 0 otherwise. According to the *TIMSS 2019 Item Information - Grade 8*, items are classified into two primary domains: mathematics (items 1–28) and science (items 29–54). Each domain further includes four subdomains: mathematics encompasses Number, Algebra, Geometry, and Data & Probability; science includes Biology, Chemistry, Physics, and Earth Science. This hierarchical structure aligns naturally with the two-layer CDM, where the first layer consists of $K_1 = 8$ subdomain attributes, and the second layer consists of $K_2 = 2$ main domain attributes. The item-subdomain-domain assignment structure specified in the TIMSS documentation naturally gives rise to a set of provisional $\mathbf{Q}$ matrices, which are presented in the Supplementary Material. They indeed satisfy the strict identifiability conditions for General DeepCDMs.

To better understand the internal structure of this assessment, we applied a two-layer exploratory DeepCDM to fit the data. Given that our primary interest lies in uncovering the hierarchical structure of attributes rather than modeling complex attribute interactions, we selected the main-effect model as our measurement framework. Aligning the number of attributes with $(K_1, K_2) = (8, 2)$ to those specified in the provisional $\mathbf{Q}$-matrices serves two purposes. First, it allows the exploratory approach to empirically verify the provisional attribute-item mappings, providing evidence for the validity of the test's original design. Second, the exploratory model maintains sufficient flexibility to identify alternative attribute-item structures, potentially revealing subtle item-attribute relationships not fully anticipated during initial test construction. This dual functionality offers valuable insights for both test

validation and future item development.

The attributes in the first layer are numerically labeled from 1 to 8, whereas those in the second layer are labeled as A and B. After estimating the coefficient matrices, we re-ordered the first-layer attributes to better visualize the underlying block structures. Figure 2 presents heatmaps of the estimated parameters for both layers, from which a distinct structure emerges, aligning closely with the intended test design. Specifically, Figure 2 (left) reveals two clearly defined blocks of non-zero coefficients: the first block associates items 1–28 with Attributes 1, 4, 5, and 7, and the second block links items 29–54 to Attributes 2, 3, 6, and 8. This block structure mirrors TIMSS's explicit distinction between mathematics and science items. Based on this clear division, we infer that Attributes 1, 4, 5, and 7 represent subdomains belonging to a common domain, while Attributes 2, 3, 6, and 8 form subdomains within another domain. This hierarchical interpretation is further supported by the second-layer heatmap shown in Figure 2 (right), which exhibits sparsity: Attributes 1, 4, 5, and 7 exclusively load onto Meta-Attribute B; Attributes 3, 6, and 8 exclusively load onto Meta-Attribute A; and Attribute 2 uniquely cross-loads onto both Meta-Attributes A and B. This inferred hierarchical structure closely matches the provisional second-layer matrix $\mathbf{Q}^{(2)}$, except for the cross-loading behavior of Attribute 2. Given the item content and structure of $\mathbf{Q}^{(2)}$, we conclude that Meta-Attribute A corresponds to the science domain, and Meta-Attribute B to the mathematics domain. The cross-loading of Attribute 2 suggests that this subdomain, despite being associated with science, may also involve mathematical competence during the reasoning process.

Although the exploratory results align closely with the provisional test design, the estimated first-layer attribute structure exhibits some deviations from the provisional $\mathbf{Q}^{(1)}$-matrix. Before examining these deviations in detail, we first evaluate the plausibility of our exploratory findings by assessing model fit. Specifically, we compare the exploratory DeepCDM against a confirmatory DeepCDM that directly employs the provisional $\mathbf{Q}$-matrices. These two approaches represent fully data-driven and strictly design-driven modeling, respectively. Consistent with the procedure outlined in our simulation study, model fit is quantified using the BIC. The exploratory DeepCDM achieves an BIC of 89,656, markedly lower than the confirmatory model's BIC of 94,946. This improvement suggests that exploratory mod-

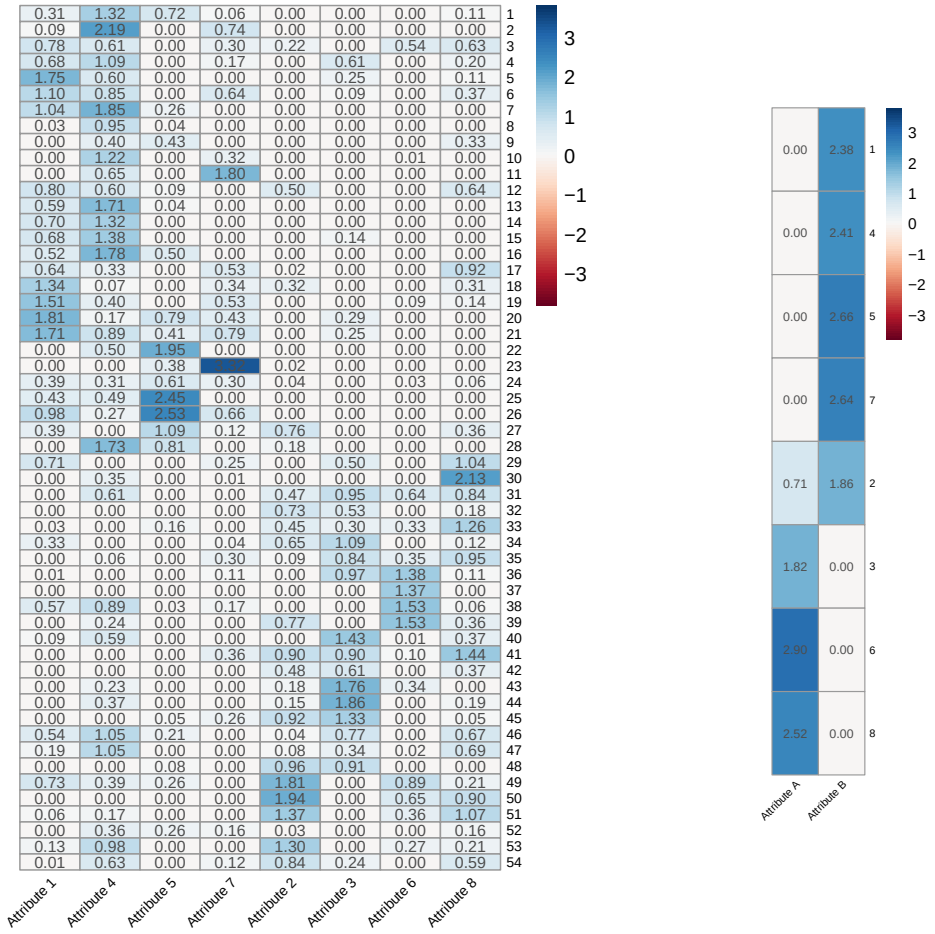

Figure 2: Heatmaps of Estimated Coefficients from Exploratory DeepCDM: First Layer (left) and Second Layer (right)

eling can be beneficial in uncovering item-attribute relationships not fully captured by the original test design.

Next, we investigate the potential item-attribute relationships. Although the secure item content is not publicly available, TIMSS provides detailed metadata for each item, including descriptive labels and associated topic areas. The item labels offer concise summaries of each item's content focus, while the topic areas reflect broader curricular domains defined by the TIMSS content framework. These metadata serve as valuable proxies, enabling us to infer the cognitive processes and skills required to answer each item. The complete metadata are provided in the Supplementary Material.

To facilitate interpretation, we focus on the five highest-loading items for each attribute,

Index	Extracted Attribute	Items	Cognitive Process
1	Algebraic Fluency	5, 18, 19, 20, 21	Manipulating algebraic expressions and solving equations in symbolic and applied contexts
2	Scientific Reasoning in Physical Contexts	48, 49, 50, 51, 53	Interpreting experimental conditions and reasoning about physical processes
3	Scientific Classification and Structure Reasoning	34, 40, 43, 44, 45	Categorizing scientific entities based on physical, chemical, or biological properties
4	Applied Quantitative Modeling	2, 7, 13, 16, 28	Applying mathematical concepts to real-world or semi-structured quantitative scenarios
5	Visual Quantitative Reasoning	1, 22, 25, 26, 27	Interpreting quantitative relationships through visual formats such as graphs, coordinates, and shaded figures
6	Environmental Systems Reasoning	35, 36, 37, 38, 39	Understanding interactions within environmental, planetary, and ecological systems
7	Spatial and Measurement Reasoning	6, 11, 17, 23, 10	Reasoning about shapes, spatial configurations, and geometric measurements
8	Biological and Ecological Reasoning	29, 30, 31, 33, 41	Inferring biological relationships and reasoning through cause-effect processes in ecological systems

Table 7: Summary of Extracted Attributes, Representative Items, and Cognitive Processes

balancing between representativeness and clarity. Each item is assigned to only one attribute group—specifically, the one for which it has the highest loading value among all attributes—ensuring that item groupings are mutually exclusive and reflect their most salient associations. For each attribute group, we carefully review the content of its assigned items, identify shared cognitive processes, and distill the latent ability the attribute is likely to capture. The distilled attribute names, along with their representative item groups and corresponding cognitive processes, are presented in Table 7.

To clarify and illustrate our interpretive process, we present two representative examples: Attribute 1 from the mathematics domain and Attribute 2 from the science domain. The detailed interpretation for all eight attributes is provided in the Supplementary Material. Attribute 1 is primarily associated with Items 5, 18, 19, 20, and 21. Based on TIMSS metadata, these items appear to involve tasks such as expressing the area of a rectangle algebraically, evaluating expressions by substituting values, identifying equivalent algebraic expressions, deriving a formula for stopping distance, and solving for an unknown variable

given the perimeter of a triangle. Although these items vary in surface content, they seem to share a common cognitive emphasis on algebraic manipulation and symbolic reasoning. This pattern suggests procedural fluency in algebra, which includes mastering algebraic structures, applying operations accurately, and recognizing equivalent mathematical forms. Accordingly, we interpret Attribute 1 as *Algebraic Fluency*, reflecting the ability to manipulate algebraic expressions and apply fundamental algebraic procedures.

Attribute 2 is primarily associated with Items 48, 49, 50, 51, and 53. According to TIMSS metadata, these items are likely to involve tasks such as explaining the behavior of gas molecules in an expanding balloon, evaluating appropriate conditions in a heat conduction experiment, reasoning about the effects of planetary gravity on vehicle weight, predicting the behavior of sound in a vacuum, and interpreting evidence related to global warming. While these items span different scientific topics, they appear to share a cognitive focus on reasoning through empirical or hypothetical scenarios, interpreting observations, and evaluating experimental setups. Based on this pattern, we interpret Attribute 2 as *Scientific Reasoning in Physical Contexts*, reflecting systematic reasoning about physical phenomena, empirical data, and conditions relevant to scientific inquiry. As scientific reasoning often draws on mathematical competence, this also supports the observed cross-loading of Attribute 2 onto both science and mathematics domains.

It is important to emphasize that the attribute structure identified through our analysis does not represent the only possible or ideal solution, as the interpretation relies on available metadata rather than direct access to detailed item content. Nevertheless, this empirical analysis illustrates how test data can be examined using an exploratory approach and demonstrates how the resulting attribute structure can be interpreted using accessible metadata. This reflects a common practical scenario, where exploratory results may not fully align with the original test design and detailed item content may be unavailable. By analyzing the derived results through metadata or with expert input, practitioners may uncover findings that offer new perspectives or supplementary insights into the test design.

6 Discussion

This paper builds a conceptual and methodological bridge between deep generative modeling and cognitive diagnosis. By significantly generalizing the DeepCDMs proposed by Gu (2024) to the challenging exploratory settings, we introduce a new class of models—*exploratory DeepCDMs*—that retain the expressive capacity of DGMs while incorporating the structural constraints, interpretability, and identifiability essential for diagnostic assessment. To enable estimation in this more complex, multi-layer setting with multiple unknown $\mathbf{Q}$-matrices, we proposed a novel *layer-wise EM algorithm* for regularized maximum likelihood estimation. This algorithm advances the literature by offering a principled, modular framework for learning complex hierarchical latent structures in diagnostic models. Both the algorithm derivation and the identifiability theory of DeepCDMs support a bottom-up, layer-by-layer estimation strategy, making the procedure not only efficient but also theoretically grounded.

A promising direction for future work is to relax the assumption of known latent dimensions across layers. One approach is to incorporate layer-wise dimension selection into the USVT-based initialization, using the largest spectral ratio of singular values, as proposed by Lee and Gu (2025). This procedure can be applied recursively, using estimated or sampled latent attributes from one layer as input to the next, enabling automatic, data-driven dimension selection. Alternative methods, such as the extended BIC (EBIC; Chen and Chen 2008) and the method of sieves (Shen and Wong 1994), may also support layer-wise model selection. It would also be valuable to extend the model to accommodate polytomous responses and attributes (Chen and de la Torre, 2013; Gao et al., 2021). A similar layer-wise EM algorithm could be developed by applying a one-layer EM procedure for polytomous data at each layer, with corresponding identifiability conditions established for the between-layer $\mathbf{Q}$-matrices. Another useful extension is to develop a stochastic version of the layer-wise EM algorithm. Such variants may offer computational advantages for large-scale, high-dimensional data and serve as a flexible alternative when full E-step computations are costly. More broadly, this work is motivated by the goal of integrating DGMs—and machine learning methods more generally—into cognitive diagnostic modeling. The simulation and empirical results suggest that this integration is a fruitful direction. Moving forward, exploring identifiability

in existing DGMs and adapting their algorithms to promote sparsity could benefit not only psychometrics but also other domains where interpretability is crucial.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

Bradshaw, L. and Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79(3):403–425.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chen, J. and de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37(6):419–437.

Chen, Y., Culpepper, S. A., and Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, 85(1):121–153.

Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, 110(510):850–866.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76:179–199.

de la Torre, J. and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353.

de la Torre, J. and Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order irt model approach. *Applied Psychological Measurement*, 33(8):620–639.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Gao, X., Ma, W., Wang, D., Cai, Y., and Tu, D. (2021). A class of cognitive diagnosis models for polytomous data. *Journal of Educational and Behavioral Statistics*, 46(3):297–322.

Gu, Y. (2024). Going deep in diagnostic modeling: Deep cognitive diagnostic models (deepcdms). *Psychometrika*, 89(1):118–150.

Gu, Y. and Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, 84(2):468–483.

Gu, Y. and Xu, G. (2020). Partial identifiability of restricted latent class models. *Annals of Statistics*, 48(4):2082–2107.

Gu, Y. and Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*, 31:449–472.

Hastie, T., Qian, J., and Tay, K. (2021). An introduction to glmnet. *CRAN R Repositary*, 5:1–35.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., and Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics*, 25(4):490–496.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25:258–272.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Le Roux, N. and Bengio, Y. (2008). Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649.

Lee, S. and Gu, Y. (2025). Deep discrete encoders: Identifiable deep generative models for rich data with discrete latent layers. *arXiv preprint arXiv:2501.01414*.

Liu, J., Lee, S., and Gu, Y. (2025). Exploratory general-response cognitive diagnostic models with higher-order structures. *Psychometrika*, pages 1–42. Published online.

Ma, W. (2022). A higher-order cognitive diagnosis model with ordinal attributes for dichotomous response data. *Multivariate behavioral research*, 57(2-3):408–421.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64(2):187–212.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771. PMLR.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Rohe, K. and Zeng, M. (2023). Vintage factor analysis with varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4).

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.

Salakhutdinov, R. and Hinton, G. (2009). Deep Boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455. PMLR.

Segal, E., Pe'er, D., Regev, A., Koller, D., Friedman, N., and Jaakkola, T. (2005). Learning module networks. *Journal of Machine Learning Research*, 6(4).

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20:345–354.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of statistical software*, 106.

Templin, J. L., Henson, R. A., Templin, S. E., and Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, 32(7):559–574.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61:287–307.

von Davier, M. and Lee, Y.-S. (2019). Handbook of diagnostic classification models. *Cham: Springer International Publishing*.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *Annals of Statistics*, 45:675–707.

Zhang, H., Chen, Y., and Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372.

Supplementary Material

This Supplementary Material is organized as follows. Supplement A outlines the identifiability results of DeepCDMs. Supplement B presents the sequences of regularization parameters used in the simulation study. Supplement C presents additional information for the real data analysis, including the provisional $\mathbf{Q}$-matrices, complete metadata and detailed interpretations of all eight extracted attributes presented in Section 5 of the main text.

A Theoretical Identifiability Conditions

This appendix outlines the identifiability results of DeepCDMs. For the technical proofs of these theoretical results, see Gu (2024).

A.1 Sharp Strict Identifiability Result for DeepDINA

In this subsection, we summarize the sharp necessary and sufficient conditions for the strict identifiability of the DeepDINA model, as established in prior work Gu (2024). The parameter space for the deep-layer population proportions $\boldsymbol{\pi}^{(D)}$ is defined as $\Delta^{2^{K_D}-1} = \left\{\pi^{(D)}_{\boldsymbol{\alpha}_\ell} : \sum_{\ell=1}^{2^{K_D}} \pi^{(D)}_{\boldsymbol{\alpha}_\ell} = 1,\, \pi^{(D)}_{\boldsymbol{\alpha}_\ell} > 0\right\}$. It is assumed that $\pi^{(D)}_{\boldsymbol{\alpha}_\ell} > 0$ for each deep latent profile $\boldsymbol{\alpha}_\ell \in \{0,1\}^{K_D}$—a standard condition consistent with those commonly imposed in single-layer CDMs. We now briefly review the definition of strict identifiability relevant to this setting.

Definition 1 (Strict Identifiability)**.** *An exploratory DeepCDM is said to be strictly identifiable, if the distribution of the observed vector* $\mathbf{R}$ *in* (5) *uniquely determines all of the following: all continuous parameters in the layerwise conditional distributions, the deepest proportion parameters* $\boldsymbol{\pi}^{(D)}$*, and all* $\mathbf{Q}$*-matrices at different depths* $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$*, up to some column/row permutation.*

A key assumption for DeepDINA's identifiability is the C-R-D conditions. In the traditional DINA model with a saturated attribute framework, these conditions are necessary and sufficient for identifiability, holding in both confirmatory (Gu and Xu, 2019) and exploratory settings (Gu and Xu, 2021). We summarize them below.

(C) **C**ompleteness. A $\mathbf{Q}$-matrix with K columns contains an identity submatrix $\mathbf{I}_K$ after some row permutation. That is, the $\mathbf{Q}$ can be row-permuted to be $\mathbf{Q} = [\mathbf{I}_K, (\mathbf{Q}^*)^\top]^\top$.

(R) **R**epeated-Measurement. Each of the K attributes is measured by at least three items.

(D) **D**istinctness. Assuming Condition (C) holds, after removing the identity submatrix $\mathbf{I}_K$ from $\mathbf{Q}$, the remaining submatrix $\mathbf{Q}^*$ contains K distinct column vectors.

Theorem 1 provides a sharp identifiability result for exploratory DeepDINA with arbitrary depth D, offering the necessary and sufficient conditions on the multiple $\mathbf{Q}$-matrices.

Theorem 1 (DeepDINA)**.** *Consider a ladder-shaped exploratory DeepDINA model with D latent layers and D between-layer $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$. The model is strictly identifiable if and only if each $\mathbf{Q}^{(d)}$, $d = 1, \ldots, D$, satisfies the C-R-D conditions.*

The sharp identifiability conditions in Theorem 1 impose transparent constraints on the $\mathbf{Q}$-matrices, which are also necessary and sufficient for identifying the DeepDINO model due to the duality between DINA and DINO (Chen et al., 2015). These conditions imply that in an identifiable DeepDINA, the layer sizes must satisfy $J > K_1 + \lceil \log_2(K_1) \rceil$ and $K_{d-1} > K_d + \lceil \log_2(K_d) \rceil$ for $d = 2, \ldots, D$ (Gu and Xu, 2021; Gu, 2024). This suggests a progressively shrinking ladder-like sparse architecture for the latent layers as depth increases.

A.2 Strict Identifiability Result for General DeepCDMs

This subsection outlines general strict identifiability conditions for any DeepCDM, including Hybrid DeepCDMs introduced in Section 2.2.

Theorem 2 (General DeepCDM)**.** *Consider an exploratory general DeepCDM with D latent layers and D between-layer $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$.* ***Either*** *Condition (S)* ***or*** *Condition (S^*) below is sufficient for strict identifiability of the model.*

(S) Each $\mathbf{Q}^{(d)}$ can be written as $\mathbf{Q}^{(d)} = [\mathbf{I}_{K_d}, \mathbf{I}_{K_d}, \mathbf{I}_{K_d}, (\mathbf{Q}^{(d)*})^\top]^\top$ after some column/row permutation, where $\mathbf{Q}^{(d)*}$ is an arbitrary $(K_{d-1} - 3K_d) \times K_d$ matrix (potentially empty).

(S) This condition is the combination of both (S1*) and (S2*) below.*

(S1) Each* $\mathbf{Q}^{(d)}$ *can be written as* $\mathbf{Q}^{(d)} = [\mathbf{I}_{K_d}, \mathbf{I}_{K_d}, (\mathbf{Q}^{(d)*})^\top]^\top$ *after some column/row permutation, where* $\mathbf{Q}^{(d)*}$ *is an arbitrary matrix (potentially empty).*

(S2) For any two different* K_d*-dimensional latent patterns* $\boldsymbol{\alpha}_c$, $\boldsymbol{\alpha}_\ell \in \{0,1\}^{K_d}$, *there exists some* $j > 2K_d$ *such that* $\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha}_c,\ \mathbf{Q}^{(d)},\ \boldsymbol{\theta}^{(d)}) \neq \mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha}_\ell,\ \mathbf{Q}^{(d)},\ \boldsymbol{\theta}^{(d)})$, *where* $\boldsymbol{\theta}^{(d)}$ *generically denotes continuous parameters required to fully specify the conditional distribution.*

Theorem 2 is broadly applicable to any DeepCDM, regardless of the diagnostic model used in each layer. Based on the theorem's conditions, the layer sizes must satisfy $J > 2K_1$ and $K_{d-1} > 2K_d$ for $d = 2, \ldots, D$, indicating a progressively shrinking, sparse latent structure as depth increases.

By comparing Theorems 1 and 2, we observe that the sufficient conditions for arbitrary DeepCDMs are stricter than those required for identifying DeepDINA. The next proposition confirms that when a DeepCDM includes a combination of DINA layers and main-effect/all-effect layers, the $\mathbf{Q}$-matrices for the DINA layers only need to satisfy the weaker C-R-D conditions, instead of the stronger Conditions (S) or (S*) in Theorem 2.

Proposition 1 (Hybrid DeepCDM)**.** *Consider a Hybrid DeepCDM with* D *latent layers and* D *between-layer* $\mathbf{Q}$*-matrices* $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$. *If each* $\mathbf{Q}^{(d)}$ *satisfies the identifiability conditions for the specific diagnostic model that* $\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}$ *follows (i.e., C-R-D for DINA, (S) or (S*) for main-effect or all-effect model), then the entire DeepCDM is strictly identifiable.*

A.3 Generic Identifiability of Main-effect and All-effect DeepCDMs

Strict identifiability is the strongest notion of identifiability, requiring that parameters be identifiable across the entire parameter space $\mathcal{T}$. A slightly weaker notion, *generic identifiability* (Allman et al., 2009), only requires identifiability almost everywhere in $\mathcal{T}$, allowing non-identifiability on a measure-zero subset $\mathcal{N} \subset \mathcal{T}$. As noted by Allman et al. (2009),

generic identifiability is often sufficient for real data analysis and is widely useful in practice. In what follows, we outline the conditions under which *main-effect* and *all-effect* DeepCDMs achieve generic identifiability. We begin by defining *main-effect-based* DeepCDMs.

Definition 2 (Main-effect-based DeepCDMs)**.** *A DeepCDM is said to be "main-effect-based", if the layerwise conditional distribution can be written as:*

$$\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha},\ \mathbf{Q}^{(d)},\ \boldsymbol{\beta}^{(d)}) = f\Big(\sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(d)} \Big\{ q_{j,k}^{(d)} \alpha_k \Big\} + \cdots \Big).$$

where $f(\cdot)$ is a link function, and the " $\cdots$ " refers to potentially more terms such as the interaction-effects of the α_k's and the intercept.

Note that Main-effect-based DeepCDMs also covers All-effect DeepCDMs, because the latter also incorporate the main effects of attributes. DeepDINA is not a main-effect-based DeepCDM since it lacks the main-effect coefficients, like $\beta_{j,k}^{(d)}$, outlined in Definition 2. These coefficients are key to achieving generic identifiability and allow relaxing the condition that each $\mathbf{Q}^{(d)}$ must contain a submatrix $\mathbf{I}_{K_d}$ (Gu and Xu, 2020; Chen et al., 2020). Next, we formally define and establish the generic identifiability of main-effect-based DeepCDMs.

Definition 3. *Define the allowable constrained parameter space for $\boldsymbol{\beta}^{(d)}$ in Definition 2 under the binary matrix $\mathbf{Q}^{(d)}$ as*

$$\Omega_{\text{main}}(\boldsymbol{\beta}^{(d)};\, \mathbf{Q}^{(d)}) = \{\beta_{j,k}^{(d)} \neq 0 \text{ if } q_{j,k}^{(d)} = 1;\ \text{and } \beta_{j,k}^{(d)} = 0 \text{ if } q_{j,k}^{(d)} = 0\}. \tag{28}$$

The continuous parameters and the $\mathbf{Q}$-matrices are said to be generically identifiable if the set of unidentifiable continuous parameters has measure zero with respect to the Lebesgue measure on their parameter space $\cup_{d=1}^{D} \Omega_{\text{main}}(\boldsymbol{\beta}^{(d)};\, \mathbf{Q}^{(d)}) \cup \Delta^{2^{K_D}-1}$.

Theorem 3. *Consider a main-effect-based DeepCDM. Suppose each $\mathbf{Q}^{(d)}$ can be written as $\mathbf{Q}^{(d)} = [(\mathbf{Q}_1^{(d)})^\top, (\mathbf{Q}_2^{(d)})^\top, (\mathbf{Q}^{(d)*})^\top]^\top$ after some column/row permutation and satisfies the following conditions. Then the main-effect-based DeepCDM is generically identifiable.*

(G1) Each $\mathbf{Q}_m^{(d)}$ *(*$m = 1, 2$*) has size* $K_d \times K_d$ *and takes the following form:*

$$\mathbf{Q}_m^{(d)} = \begin{pmatrix} 1 & * & \cdots & * \\ * & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 1 \end{pmatrix}, \quad m = 1, 2; \quad d = 1, \ldots, D.$$

That is, $\mathbf{Q}_1^{(d)}$ *and* $\mathbf{Q}_2^{(d)}$ *each has all the diagonal entries equal to one, whereas any off-diagonal entry is free to be either one or zero.*

(G2) The $(K_{d-1} - 2K_d) \times K_d$ *submatrix* $\mathbf{Q}^{(d)*}$ *in* $\mathbf{Q}^{(d)}$*,* $d = 1, \ldots, D$*, satisfies that each column contains at least one entry of "1".*

Theorem 3 relaxes the strict identifiability conditions from Theorem 2 by removing the requirement for any $\mathbf{Q}^{(d)}$ to contain an identity submatrix $\mathbf{I}_{K_d}$. Moreover, these generic identifiability conditions suggest a shrinking latent structure as depth increases, since (G1) and (G2) imply that $J > 2K_1$ and $K_d > 2K_{d+1}$ for $d = 1, \ldots, D-1$.

B Sequences of Regularization Parameters Used in the Simulation Study

Table B.1 presents the sequences of scaled regularization parameters $N \cdot s_d$ used in the simulation study across different sample sizes and layers. These sequences are applied consistently across all three measurement model settings: main-effect, all-effect, and DINA. In general, the tuning parameters are specified to decrease with increasing sample size, following theoretical guidance for regularization parameter selection Chen et al. (2015).

Sample size	Layer		
	1	2	3
1000	$(0.010, 0.011, 0.012)$	$(0.010, 0.011, 0.012)$	$(0.015, 0.016, 0.017)$
1500	$(0.009, 0.010, 0.011)$	$(0.009, 0.010, 0.011)$	$(0.014, 0.015, 0.016)$
2000	$(0.008, 0.009, 0.010)$	$(0.008, 0.009, 0.010)$	$(0.013, 0.014, 0.015)$

Table B.1: Sequences of scaled regularization parameters $N \cdot s_d$ used in the simulation study, across different sample sizes and layers.

C Supplement for Data Analysis

In this appendix, we provide supplementary information for the real data analysis in Section 5. Tables C.1 and C.2 present the provisional $\mathbf{Q}$-matrices derived from the TIMSS assessment design. Table C.3 lists metadata for each item, including descriptive labels and associated topic areas. In addition, detailed interpretations of the eight extracted attributes in Table 7 of the main text are provided below.

Item ID	Number	Algebra	Geometry	Data & Prob.	Biology	Chemistry	Physics	Earth Science
1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0
10	0	0	1	0	0	0	0	0
11	0	0	1	0	0	0	0	0
12	0	0	0	1	0	0	0	0
13	0	0	0	1	0	0	0	0
14	1	0	0	0	0	0	0	0
15	1	0	0	0	0	0	0	0
16	1	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0
18	0	1	0	0	0	0	0	0
19	0	1	0	0	0	0	0	0
20	0	1	0	0	0	0	0	0
21	0	1	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0
23	0	0	1	0	0	0	0	0
24	0	0	1	0	0	0	0	0
25	0	0	1	0	0	0	0	0
26	0	0	0	1	0	0	0	0
27	0	0	0	1	0	0	0	0
28	0	0	0	1	0	0	0	0
29	0	0	0	0	1	0	0	0
30	0	0	0	0	1	0	0	0
31	0	0	0	0	1	0	0	0
32	0	0	0	0	1	0	0	0
33	0	0	0	0	0	1	0	0
34	0	0	0	0	0	0	1	0
35	0	0	0	0	0	0	1	0
36	0	0	0	0	0	0	1	0
37	0	0	0	0	0	0	0	1
38	0	0	0	0	0	0	0	1
39	0	0	0	0	1	0	0	0
40	0	0	0	0	1	0	0	0
41	0	0	0	0	1	0	0	0
42	0	0	0	0	1	0	0	0
43	0	0	0	0	1	0	0	0
44	0	0	0	0	0	1	0	0
45	0	0	0	0	0	1	0	0
46	0	0	0	0	0	1	0	0
47	0	0	0	0	0	1	0	0
48	0	0	0	0	0	0	1	0
49	0	0	0	0	0	0	1	0
50	0	0	0	0	0	0	1	0
51	0	0	0	0	0	0	1	0
52	0	0	0	0	0	0	0	1
53	0	0	0	0	0	0	0	1
54	0	0	0	0	0	0	0	1

Table C.1: First-layer provisional $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)}_{54\times 8}$ for item booklet No.1 in TIMSS 2019 eighth grade assessment.

Subdomains \ Main Domains	Mathematics	Science
Number	1	0
Algebra	1	0
Geometry	1	0
Data and Probability	1	0
Biology	0	1
Chemistry	0	1
Physics	0	1
Earth Science	0	1

Table C.2: Second-layer $\mathbf{Q}$-matrix $\mathbf{Q}^{(2)}_{8\times 2}$ for TIMSS 2019 eighth grade assessment.

Item ID	Topic Area	Label
1	Fractions and Decimals	Octagon with equivalent shading
2	Integers	Time when Pat finishes last lap; Percentage of laps finished
3	Integers	Multiples of 3
4	Fractions and Decimals	Convert decimal to a fraction
5	Expressions, Operations, and Equations	Expression for area of rectangle
6	Expressions, Operations, and Equations	Expression with exponents of y
7	Relationships and Functions	Number of matches for figure 10; Rule for number of matches
8	Relationships and Functions	Graph of y = 2x
9	Geometric Shapes and Measurements	Rotation and reflection
10	Geometric Shapes and Measurements	Surface area of the prism
11	Geometric Shapes and Measurements	Value of angle x outside triangle
12	Probability	Number of balls in a bag
13	Data	Liv's smartphone use; Smartphone use listening to music
14	Integers	Statements for all values of integer a (DERIVED)
15	Fractions and Decimals	Arrow to show 5/12 on number line
16	Fractions and Decimals	Value of fraction X in square
17	Ratio, Proportion, and Percent	Number of blue beads on bracelet
18	Expressions, Operations, and Equations	Value of 2(6x - 3y)
19	Expressions, Operations, and Equations	Expression equivalent to 2y + 6xy2
20	Expressions, Operations, and Equations	Formula for stopping distance
21	Expressions, Operations, and Equations	Value of x given perimeter of triangle ABC
22	Relationships and Functions	Additional point on a straight line
23	Geometric Shapes and Measurements	Value of angle x in a quadrilateral
24	Geometric Shapes and Measurements	Methods of folding paper- height, diameter, surface area
25	Geometric Shapes and Measurements	Coordinates to complete KLMN- x coordinate, - y coordinate
26	Data	Mean temperature for 5 days
27	Data	Best graph for town information - jobs, boys and girls, population
28	Data	Bar graph of newspaper sales
29	Cells and Their Functions	Organism with cell walls
30	Ecosystems	How decomposers get energy
31	Ecosystems	Organism that competes with humans
32	Ecosystems	Garden with bird feeder: cat+birds, cat+birds, cat+mouse
33	Properties of Matter	Why Solution 2 is paler than 1

Item ID	Topic Area	Label
34	Physical States and Changes in Matter	Which is a physical change
35	Electricity and Magnetism	Model flashlight: Bulb won't light; 2 parallel bulbs; Comparison
36	Electricity and Magnetism	Two bar magnets repelling
37	Earth in the Solar System and the Universe	Planets: Shortest day length; Distance from Sun
38	Earth's Structure and Physical Features	Temperature outside an airplane
39	Ecosystems	Relationship between insects and flowering plants
40	Cells and Their Functions	Where in a cell DNA replication occurs
41	Ecosystems	Increase green space as carbon dioxide increases
42	Ecosystems	Why leaves' masses decreased
43	Characteristics and Life Processes of Organisms	Classify animals based on a single characteristic, Identify the characteristic used to classify animals
44	Composition of Matter	Location of subatomic particles
45	Composition of Matter	Order elements from smallest to largest atomic num
46	Properties of Matter	Acidic, basic, or neutral solution
47	Properties of Matter	Mixing an acid and base solution
48	Physical States and Changes in Matter	Gas molecules in an expanding balloon
49	Energy Transformation and Transfer	Things Tom should do (DERIVED): same type of wax on both rods, higher flame for the copper rod, paperclips from different materials, etc.
50	Motion and Forces	Vehicle with different weights on different planets
51	Light and Sound	Cell phone in a vacuum
52	Earth's Structure and Physical Features	Why balloon gets bigger as it rises
53	Earth's Processes, Cycles, and History	Evidence of global warming
54	Earth's Processes, Cycles, and History	Natural resource formation shown in diagrams

Table C.3: Metadata for TIMSS Items in Booklet 1

Attribute 1 is primarily associated with Items 5, 18, 19, 20, and 21. Based on TIMSS metadata, these items appear to involve tasks such as expressing the area of a rectangle algebraically, evaluating expressions by substituting values, identifying equivalent algebraic expressions, deriving a formula for stopping distance, and solving for an unknown variable given the perimeter of a triangle. Although these items vary in content, they share a common cognitive emphasis on algebraic manipulation and symbolic reasoning. This pattern suggests procedural fluency in algebra, which includes mastering algebraic structures, applying oper-

ations accurately, and recognizing equivalent mathematical forms. Accordingly, we interpret Attribute 1 as *Algebraic Fluency*, reflecting the ability to manipulate algebraic expressions and apply fundamental algebraic procedures.

Attribute 2 is primarily associated with Items 48, 49, 50, 51, and 53. According to TIMSS metadata, these items are likely to involve tasks such as explaining the behavior of gas molecules in an expanding balloon, evaluating appropriate conditions in a heat conduction experiment, reasoning about the effects of planetary gravity on vehicle weight, predicting the behavior of sound in a vacuum, and interpreting evidence related to global warming. While these items span different scientific topics, they share a cognitive focus on reasoning through empirical or hypothetical scenarios, interpreting observations, and evaluating experimental setups. Based on this pattern, we interpret Attribute 2 as *Scientific Reasoning in Physical Contexts*, reflecting systematic reasoning about physical phenomena, empirical data, and conditions relevant to scientific inquiry.

Attribute 3 corresponds to Items 34, 40, 43, 44, and 45, which, based on their metadata, appear to involve identifying physical changes, locating cellular processes, classifying organisms, recognizing subatomic structures, and ordering elements by atomic number. These items seem to require categorization and structural understanding of scientific entities across biology, chemistry, and physics. The common cognitive emphasis lies in classification and the organization of scientific knowledge. Accordingly, we interpret Attribute 3 as *Scientific Classification and Structure Reasoning*, reflecting the ability to sort and organize domain-specific information using scientific criteria.

Attribute 4 includes Items 2, 7, 13, 16, and 28. These items are likely to involve applying mathematical reasoning to contextualized or real-world situations, such as interpreting percentages and time, identifying numerical patterns, analyzing device usage, working with fractions, and interpreting a bar graph of newspaper sales. The shared emphasis appears to be on translating semi-structured scenarios into quantitative representations. We therefore interpret Attribute 4 as *Applied Quantitative Modeling*, referring to the ability to construct and use mathematical representations to understand and reason about contextualized quan-

titative information.

Attribute 5 is defined by Items 1, 22, 25, 26, and 27. According to metadata, these items appear to involve reasoning with shaded figures, identifying linear patterns, completing coordinate shapes, computing averages, and selecting appropriate graphs. The shared cognitive emphasis is on interpreting visual or spatial representations to extract quantitative meaning. As such, we interpret Attribute 5 as *Visual Quantitative Reasoning*, which highlights the ability to engage in quantitative thinking through visual cues and data structures.

Attribute 6 consists of Items 35, 36, 37, 38, and 39. These items are associated with topics such as electrical circuits, magnetic forces, planetary properties, atmospheric conditions, and ecological interactions. While varying in scientific content, they collectively seem to require reasoning about the dynamic relationships that govern natural or environmental systems. Therefore, we interpret Attribute 6 as *Environmental Systems Reasoning*, reflecting the process of analyzing complex physical and ecological interactions.

Attribute 7 corresponds to Items 6, 11, 17, 23, and 10. Based on their descriptions, these items likely require reasoning about spatial configurations, angle relationships, proportional reasoning, and surface area computation. The common cognitive demand appears to be spatial visualization integrated with quantitative reasoning. We interpret Attribute 7 as *Spatial and Measurement Reasoning*, denoting the ability to reason about shape, measurement, and geometric relationships. We note that Item 6, while involving algebraic expressions with exponents, may not directly reflect spatial or measurement reasoning. Its inclusion in this group may reflect empirical overlap rather than conceptual alignment and should therefore be interpreted with caution.

Attribute 8 is associated with Items 29, 30, 31, 33, and 41. These items involve topics such as cellular structure, energy flow in ecosystems, species interactions, substance concentration, and environmental impact. While the specific topics vary, they seem to require reasoning about biological mechanisms and ecological cause-effect patterns. Thus, we interpret Attribute 8 as *Biological and Ecological Reasoning*, reflecting the ability to understand and infer relationships and processes within living systems.