# Going Deep in Diagnostic Modeling: Deep Cognitive Diagnostic Models (DeepCDMs)

Yuqi Gu*

Department of Statistics, Columbia University

**Abstract**

Cognitive Diagnostic Models (CDMs) are discrete latent variable models popular in educational and psychological measurement. In this work, motivated by the advantages of deep generative modeling and by identifiability considerations, we propose a new family of *DeepCDMs*, to hunt for deep discrete diagnostic information. The new class of models enjoys nice properties of *identifiability*, *parsimony*, and *interpretability*. Mathematically, DeepCDMs are entirely identifiable, including even fully exploratory settings and allowing to uniquely identify the parameters and discrete loading structures (the "**Q**-matrics") at all different depths in the generative model. Statistically, DeepCDMs are parsimonious, because they can use a relatively small number of parameters to expressively model data thanks to the depth. Practically, DeepCDMs are interpretable, because the shrinking-ladder-shaped deep architecture can capture cognitive concepts and provide multi-granularity skill diagnoses from coarse- to fine-grained and from high-level to detailed. For identifiability, we establish transparent identifiability conditions for various DeepCDMs. Our conditions impose intuitive constraints on the structures of the multiple **Q**-matrices, and inspire a generative graph with increasingly smaller latent layers when going deeper. For estimation and computation, we focus on the confirmatory setting with known **Q**-matrices and develop Bayesian formulations and efficient Gibbs sampling algorithms. Simulation studies and an application to the TIMSS 2019 math assessment data demonstrate the usefulness of the proposed methodology.

**Keywords**: Bayesian inference; Bayesian network; Cognitive Diagnostic Model; DeepCDM; Deep generative model; Deep learning; Directed graphical model; Identifiability; **Q**-matrix.

## 1 Introduction

Cognitive Diagnostic Models (CDMs), or Diagnostic Classification Models (Rupp et al., 2010; von Davier and Lee, 2019), are powerful and popular discrete latent variable models in edu-

*Email: yuqi.gu@columbia.edu. Address: Room 928 SSW, 1255 Amsterdam Avenue, New York, NY 10027. This work is partially supported by NSF grant DMS-2210796.

cational and psychological measurement. Based on subjects' item responses, a CDM enables fine-grained diagnostic inference on multiple discrete latent attributes. Usually, each attribute is assumed to be binary and carries a specific meaning such as the mastery/deficiency of a skill, or the presence/absence of a mental disorder. In educational settings, the diagnostic feedback on the skill attributes provides details about students' weaknesses and strengths, and can facilitate targeted instructions. In the past two decades, CDMs have attracted increasing research attention (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).

In the early years after the inception of CDMs, they were mostly applied to settings specifically designed for a diagnostic purpose, such as the celebrated fraction-subtraction data first collected and analyzed by Tatsuoka (Tatsuoka, 1983). Recently, it is increasingly attractive to gear the diagnostic modeling methodology to large-scale modern educational assessments, such as the Trends in Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA) (e.g., see von Davier, 2008; Chen and de la Torre, 2014; George and Robitzsch, 2015; Gu and Xu, 2023). These applications create new opportunities and also bring about new challenges. For example, in the TIMSS 2019 eighth-grade math assessment, each item measures multiple granularities of skills: *Content* / *Cognitive* as the general ability domains, *Number* / *Algebra* / *Geometry* / *Data and Probability* as more specific skills under the *Content* domain, *Knowing* / *Applying* / *Reasoning* as more specific skills under the *Cognitive* domain, etc. These large-scale complex assessments call for new statistical and computational methods.

Reflecting on the current CDM (i.e., diagnostic modeling) literature, many studies adopt the *saturated model* for the latent attributes, in which every configuration of the attributes has a separate proportion parameter (e.g., Chen et al., 2015; Xu and Zhang, 2016; Chen et al., 2018; Xu and Shang, 2018; Fang et al., 2019; Gu and Xu, 2019; Chen et al., 2020; Balamuta and Culpepper, 2022). Though being fully flexible, the saturated attribute model is not parsimonious, because it requires exponentially many parameters to describe the attribute distribution ($2^K - 1$ ones for $K$ binary attributes). This lack of parsimony makes applying CDMs to modern high-dimensional-attribute settings very challenging, both statistically and computationally. There exist a few important exceptions to the saturated modeling practice,

including the log-linear attribute model in Xu and von Davier (2008), the higher-order IRT-based model in de la Torre and Douglas (2004), and the multivariate probit model with one continuous factor in Templin et al. (2008). These models either include parameters that are not straightforward to interpret (log-linear parameters in Xu and von Davier, 2008), or employ only a small number of continuous latent variables to model the attributes (de la Torre and Douglas, 2004; Templin et al., 2008).

The questions motivating this work are: Is there an even more flexible, yet still parsimonious and interpretable way, to model the high-dimensional latent attributes? Is it possible to fully retain the power and goal of diagnostic modeling, and provide discrete diagnoses in multiple latent granularities (as desired in the aforementioned TIMSS application)? Is it possible to establish identifiability guarantees for such models with complex latent structures? To address these questions, we propose a deep generative modeling framework for cognitive diagnosis, which features multiple, potentially deep, entirely discrete latent layers. We name the new family of models *Deep Cognitive Diagnostic Models* (DeepCDMs), to reflect that they can serve as tools to hunt for deep diagnostic information. DeepCDMs enjoy several desirable properties simultaneously: *parsimony and richness*, *interpretability*, and *identifiability*. We elaborate on these advantages in the following.

First, DeepCDMs are statistically parsimonious yet have rich representational power. On the one hand, the parsimony comes from that a DeepCDM avoid the exponential complexity of parameters in the saturated attribute model. In fact, a DeepCDM requires only a quadratic or even linear number of parameters with respect to the number of latent variables. Such a reduction of parameter complexity does not come at the cost of a less suitable model. On the contrary, our model is well-motivated by the fact that the fine-grained latent attributes often have structured dependence on each other due to some hidden mechanisms, for which the deep architecture is well-suited to model. Indeed, the TIMSS assessment in which each item targets multiple skill granularities provides practical evidence for this argument. On the other hand, introducing multiple, potentially deep, latent layers can greatly enhance the expressive and representational power of a model, as widely recognized in the deep learning community (Bengio et al., 2013; Goodfellow et al., 2016; Ranganath et al., 2015).

Second, DeepCDMs are mathematically identifiable under intuitive conditions on the

deep generative structure. Identifiability means that the parameters can be uniquely determined from the observed distribution. It is a highly desirable property and a prerequisite for valid statistical estimation. Recently, there have been an emerging literature addressing the identifiability issues of CDMs (Xu and Zhang, 2016; Xu, 2017; Culpepper, 2019b; Fang et al., 2019; Chen et al., 2020; Gu and Xu, 2019, 2020). However, all of these works focus on the saturated attribute model. It is unknown what conditions can ensure identifiability when higher order latent structures are present in a CDM. We establish identifiability for various DeepCDMs with an arbitrary number of latent layers. Our identifiability conditions impose intuitive constraints on the between-layer graph structures captured by multiple "**Q**-matrices". These conditions directly inform how to design a DeepCDM – a ladder/pyramid shaped sparse graphical model, with the observed item responses occupying the bottom layer, and increasingly smaller sizes of latent layers when going deeper (see Figure 1).

Third, DeepCDMs are practically interpretable. The shrinking-ladder-shaped probabilistic graphical model can capture cognitive concepts and provide diagnostics from coarse- to fine-grained, and from high-level to detailed. In a DeepCDM, when climbing up the ladder and going deeper, concepts become increasingly abstract and general, capturing the big picture of knowledge; when stepping down the ladder and going shallower, concepts become increasingly concrete and specific, capturing the fine-grained details of knowledge. Therefore, the proposed DeepCDM framework can characterize a complete picture of one's knowledge structure and provide diagnostic feedback in multiple different resolutions, with each layer offering one particular resolution. Such diagnostic information can facilitate more effective multi-resolution interventions than traditional CDMs with a saturated attribute model.

In summary, this paper makes the following contributions in theory, methodology, and computation. *First*, we introduce a deep generative modeling framework for cognitive diagnosis for the first time, and propose a general class of interpretable and parsimonious DeepCDMs. *Second*, we develop identifiability theory for various DeepCDMs, applicable to both confirmatory and fully exploratory settings. Our identifiability conditions provide insights into what deep generative graph one can fundamentally uncover in a DeepCDM: a shrinking latent ladder when going deeper. *Third*, we propose Bayesian formulations and Gibbs sampling algorithms for various DeepCDMs. In this initial paper, our Bayesian infer-

ence methods are developed for the confirmatory setting with known and fixed $\mathbf{Q}$-matrices. Our algorithms enforce certain monotonicity constraints on parameters and produce interpretable estimation results.

The rest of this paper is organized as follows. Section 2 reviews existing modeling approaches, proposes the general DeepCDM framework, and gives various specific examples. Section 3 proposes transparent identifiability conditions for various DeepCDMs and discusses their practical implications. Section 4 develops the Bayesian formulations of various DeepCDMs and their corresponding Gibbs sampling algorithms. Section 5 conducts simulations studies that corroborate the identifiability theory and demonstrate the performance of the proposed algorithms. Section 6 applies the DeepCDM methodology to data extracted from the TIMSS 2019 math assessment. Finally, Section 7 provides concluding remarks. The proofs of theorems and Gibbs sampling details are included in the Supplementary Material.

# 2 Deep Discrete Latent Variable Modeling for Diagnostic Purposes

## 2.1 Existing Approaches to Latent Attribute Modeling

A traditional CDM consists of two parts in the model: the measurement part and the latent part. The measurement part describes how the observed responses measure the latent attributes, and is closely related to the concept of the $\mathbf{Q}$-matrix (Tatsuoka, 1983). Various diagnostic goals have led to different specific measurement models, including the Deterministic Input Noisy output "And" gate model (DINA; Junker and Sijtsma, 2001), the Deterministic Input Noisy output "Or" gate model (DINO; Templin and Henson, 2006), the main-effect diagnostic models (DiBello et al., 1995; Maris, 1999; de la Torre, 2011), and the all-effect general diagnostic models (von Davier, 2008; Henson et al., 2009; de la Torre, 2011). We defer introducing the details of these measurement models to Section 2.3. Next, we briefly review existing models for the latent part in a CDM; that is, models for the latent attributes.

We focus on the the commonly considered case of binary attributes. Denote the $i$th subject's latent attribute profile by $\mathbf{A}_i = (A_{i,1}, \ldots, A_{i,K})$, then each $\mathbf{A}_i$ takes one of the

$|\{0,1\}^K| = 2^K$ possible configurations. In the current literature of CDMs, the most widely used model for the latent attributes is the saturated model (Chen et al., 2015; Xu and Zhang, 2016; Chen et al., 2018; Fang et al., 2019; Gu and Xu, 2019; Chen et al., 2020), which assumes that each binary pattern $\boldsymbol{\alpha} \in \{0,1\}^K$ has its separate proportion parameter $p_{\boldsymbol{\alpha}}$ with $\mathbb{P}(\mathbf{A}_i = \boldsymbol{\alpha}) = p_{\boldsymbol{\alpha}}$. These proportion parameters satisfy that $p_{\boldsymbol{\alpha}} \geq 0$ and $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$. Though being fully flexible and general, the saturated attribute model is not parsimonious, because it requires $2^K$ proportion parameters in $\boldsymbol{\pi}$, an exponential parameter complexity.

There exist two important approaches for modeling the binary attributes through a higher-order model. One approach is the higher-order latent trait model (HO-CDM) proposed by de la Torre and Douglas (2004), which uses one or more continuous latent variables to explain the binary attributes through an IRT-type model. In the unidimensional case, each student is assumed to have a higher-order continuous ability $\theta_i$, conditioned on which the attributes $A_{i1}, \ldots, A_{iK}$ are independently generated through a Rasch, 1PL, or 2PL model (also see the `GDINA` R package and Ma and de la Torre, 2020). See more discussions on the connections and differences between the HO-CDM and DeepCDMs in Section 5. Another approach proposed by Templin et al. (2008) employs the multivariate probit model with a one-dimensional continuous factor. This approach assumes that each binary attribute $A_{i,k}$ is obtained via dichotomizing a Normal random variable $\eta_{i,k}$ by a cut-off point, and the $K$ Normal variables $(\eta_{i,1}, \ldots, \eta_{i,K})$ are generated via a factor analysis model. Both of these two approaches use a small number of continuous latent variables to model the binary attributes.

Other than the higher-order latent variable models, the independence model and the log-linear model have also been considered for modeling the attributes (Maris, 1999; Xu and von Davier, 2008). The independence attribute model is often overly simplistic in practice. The log-linear model in Xu and von Davier (2008) is flexible, but employs parameters that are not straightforward to interpret. Another different model for the latent attributes is the attribute hierarchy method (AHM; Gierl et al., 2007; Templin and Bradshaw, 2014). The AHM assumes that the mastery of certain skill attributes is a prerequisite for that of others. As pointed out by Rupp et al. (2010), the existing AHMs are pattern classification approaches rather than probabilistic measurement models.

## 2.2 The New DeepCDM Framework

Motivated by the appeal to perform diagnostic modeling at multiple granularities, we propose the deep cognitive diagnostic modeling framework. We adopt the probabilistic graphical model (Wainwright et al., 2008; Koller and Friedman, 2009) terminology, specifically, a *directed* graphical model, to rigorously define a DeepCDM. Graphical models use a graph as the basis for compactly encoding a complex joint distribution of high-dimensional random variables. In the graphical representation, the nodes correspond to the random variables, and the edges correspond to direct probabilistic interactions between them.

A general *Directed Acyclic Graph* (DAG; also called a Bayesian network as in Pearl (1988)), is defined as follows. In a DAG, every edge has a direction, and there are no directed cycles. DAGs are well-suited to model the generative mechanism and causal relations involving latent variables; see the book Almond et al. (2015) for using Bayesian networks in educational assessment. Consider $M$ random variables $X_1, \ldots, X_M$ as $M$ nodes in a DAG. If there is a directed edge from $X_\ell$ to $X_m$, then $X_\ell$ is said to be a *parent* of $X_m$ and $X_m$ a *child* of $X_\ell$. Let $\mathrm{pa}(m) \subseteq \{1, \ldots, M\}$ be the set of indices of all parents of $X_m$. Then according to the general definition of a DAG, the joint distribution of the $X_1, \ldots, X_M$ factorizes as:

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

where $\mathbb{P}(X_m \mid X_{\mathrm{pa}(m)})$ is the conditional distribution of $X_m$ given its parent variables $X_{\mathrm{pa}(m)}$. The graph structure of a DAG encodes rich conditional dependence and independence relations among the node variables, as can be checked by examining (1). If a DAG consists of latent variables, then these latent variables need to be marginalized out in the joint distribution (1) in order to obtain the marginal distribution of the observed variables.

We next introduce the formulation and notation of a general DeepCDM. At the bottom layer of a DeepCDM are the observed response variables to the $J$ items, $\mathbf{R} = (R_1, \ldots, R_J)$. The first (i.e., shallowest) latent layer adjacent to the bottom layer collects the most fine-grained latent attributes, $\mathbf{A}^{(1)} = (A_1^{(1)}, \ldots, A_{K_1}^{(1)})$. Note that a CDM with a saturated attribute model stops here and assumes the $K_1$ attributes can be arbitrarily dependent on each

other. In contrast, we model the generating mechanism of the attributes through deeper latent layers. In a $D$-latent-layer DeepCDM, denote the $d$th latent layer (counting from the bottom) by $\mathbf{A}^{(d)} = (A_1^{(d)}, \ldots, A_{K_d}^{(d)})$ for each $d = 1, 2, \ldots, D$. All edges in a DeepCDM are pointing in the top-down direction, only potentially between two adjacent layers. See Figure 1 for an example of a DeepCDM with $D = 3$. The definition in (1) also implies that all the variables in any specific layer of a DeepCDM are conditionally independent given the variables in the above layer. Such a graphical model intuitively describes how the more specific latent skills are successively generated by the more general higher-level latent "meta-skills". To fully realize the diagnostic goal, a DeepCDM assumes all latent variables to be discrete. Later, our identifiability theory will reveal that there should be smaller and smaller latent layers when going deeper; that is, $K_1 > K_2 > \cdots > K_D$, another intuitive constraint.

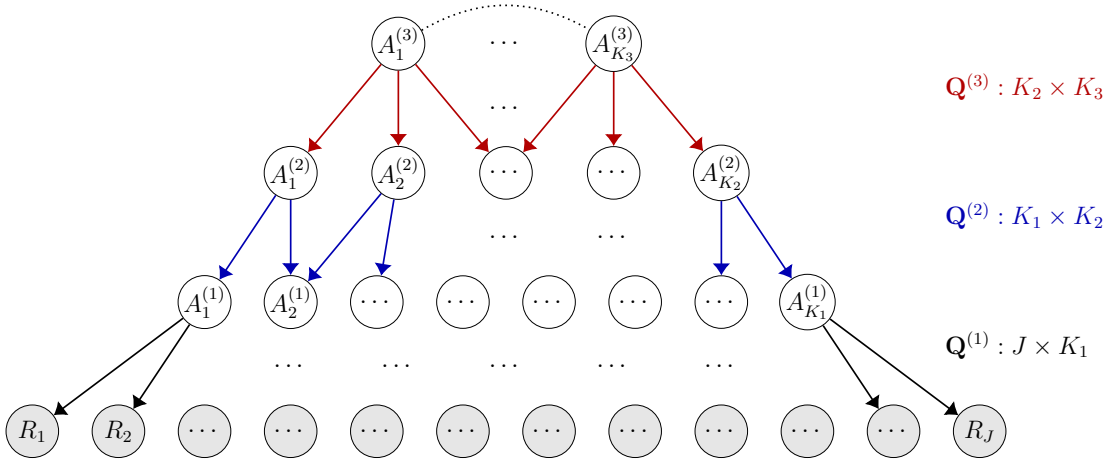



Figure 1: A ladder-shaped three-latent-layer DeepCDM. Gray nodes are observed variables, and white nodes are latent ones. Multiple layers of binary latent variables $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$, and $\mathbf{A}^{(3)}$ successively generate the observed binary responses $\mathbf{R}$. Binary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

A key feature of a DeepCDM is the *multiple* "$\mathbf{Q}$-matrices" at different depths of the graphical model, as in Figure 1. In traditional cognitive diagnosis, the $\mathbf{Q}$-matrix (Tatsuoka, 1983) is an important object that describes how the items measure the latent attributes. For example, if $J$ items are designed to measure $K$ latent attributes, then the $\mathbf{Q}$-matrix $\mathbf{Q} = (q_{j,k})$ has size $J \times K$, in which $q_{j,k} = 1$ or 0 indicates whether or not the $j$th item measures (i.e., directly depends on) the $k$th latent attribute. Recall that the edges in a graphical model exactly captures the direct dependence between variables, so $q_{j,k} = 1$ or

0 also reflects whether or not the $k$th latent node is a parent of the $j$th observed node in the graph. In other words, the traditional $\mathbf{Q}$-matrix summarize the sparse bipartite graph pattern between the latent attribute layer and the observed layer. This graphical perspective implies that a DeepCDM with $D$ latent layers should require $D$ matrices, $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, ..., $\mathbf{Q}^{(D)}$, to summarize the graph structure. In particular, $\mathbf{Q}^{(1)} = \left(q_{j,k}^{(1)}\right)$ has size $J \times K_1$ and resembles the traditional $\mathbf{Q}$-matrix; whereas for each $d = 2, \ldots, D$, the $K_{d-1} \times K_d$ matrix $\mathbf{Q}^{(d)} = \left(q_{k,\ell}^{(d)}\right)$ is similar in spirit to $\mathbf{Q}^{(1)}$, but describes how the variables in the $(d-1)$th latent layer depend on those in the layer above, the $d$th latent layer. Graphically, the entry $q_{k,\ell}^{(d)} = 1$ or 0 indicates whether or not latent variable $A_\ell^{(d)}$ is a parent of latent variable $A_k^{(d-1)}$. In this work, we will focus on developing estimation methods for the *confirmatory* DeepCDMs, where the $\mathbf{Q}$-matrices are assumed to be fixed and known.

According to the general definition of DAGs in (1) and the DeepCDM setting specified in the last paragraph, the *joint distribution* of all the variables, including the latent ones, is

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

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

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

where we make explicit how the different $\mathbf{Q}$-matrices appear in different factors in the joint distribution. The generic superscript "CDM" in the conditional distributions in (3) and (4) means that the conditional distribution conforms to a Cognitive Diagnostic Model, in each layer of the potentially deep generative process. Marginalizing out all the latent variables $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)}$ in (2) gives the marginal distribution of the observed response vector $\mathbf{R}$:

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

where $\boldsymbol{r}$ is an observed response pattern, and $\boldsymbol{\alpha}^{(d)}$ is a latent pattern for the $d$th latent layer. This work focuses on binary observed and latent variables with $\boldsymbol{r} \in \{0,1\}^J$ and

$\boldsymbol{\alpha}^{(d)} \in \{0,1\}^{K_d}$, where each observed variable denotes the correct/wrong response and each latent variable denotes the presence/absence of a skill or a meta-skill.

We model the latent variables $\mathbf{A}^{(D)}$ in the deepest latent layer of a DeepCDM using a categorical distribution, similar to traditional CDMs. Specifically, we allow for two possible generating mechanisms for $\mathbf{A}^{(D)}$ and $\mathbf{A}^{(D-1)} \mid \mathbf{A}^{(D)}$: the pyramid mechanism and the ladder mechanism. In the pyramid case, $\mathbf{A}^{(D-1)}$ follows a latent class model (Goodman, 1974) with $\mathbf{A}^{(D)}$ serving as the latent class variable; in this case $K_D = 1$ and $\mathbf{A}^{(D)}$ ranges in $\{1, \ldots, B\}$ for some integer $B$. In the ladder case, $\mathbf{A}^{(D-1)}$ follows yet another CDM with $\mathbf{A}^{(D)}$ serving as the highest order latent traits; in this case $K_D > 1$ and $\mathbf{A}^{(D)} \in \{0,1\}^{K_D}$. Both mechanisms still use fully discrete latent variables and their corresponding distributions are:

$$\mathbb{P}(\mathbf{A}^{(D)} = \boldsymbol{\alpha}) = \begin{cases} \pi_{\boldsymbol{\alpha}}^{\text{ladder}}, \quad \forall \boldsymbol{\alpha} \in \{0,1\}^{K_D}, & \text{in a ladder-shaped DeepCDM;} \\ \pi_{\boldsymbol{\alpha}}^{\text{pyramid}}, \quad \forall \boldsymbol{\alpha} \in \{1, \ldots, B\}, & \text{in a pyramid-shaped DeepCDM.} \end{cases} \tag{6}$$

The proportion parameters satisfy $\sum_{\boldsymbol{\alpha} \in \{0,1\}^{K_D}} \pi_{\boldsymbol{\alpha}}^{\text{ladder}} = 1$ or $\sum_{b=1}^{B} \pi_b^{\text{pyramid}} = 1$. Now we have completed specifying a general DeepCDM.

It is worth noting that in the literature of factor analysis of continuous data, hierarchical factor models (Schmid and Leiman, 1957) or higher-order factor models (Yung et al., 1999) are important and popular models that also contain multiple layers of factors. These models belong to the family of using continuous linear latent factors to model continuous responses, in which the statistical dependence among variables can be just summarized as *covariance or correlation* matrices. By contrast, the proposed DeepCDMs are a family of higher-order discrete latent variable models for discrete data. DeepCDMs can model various *nonlinear* and *non-additive* relationships among variables, e.g., DeepDINA with the interaction term of higher-order attributes and DeepLLM with the logistic link. These complex dependencies cannot be simply summarized by covariance or correlation matrices as in hierarchical continuous linear factor models in Schmid and Leiman (1957) and Yung et al. (1999).

## 2.3 Specific Examples of DeepCDMs

This subsection provides various specific examples of DeepCDMs under the general framework put forth in Section 2.2. Recall Equation (2) states that the joint distribution of all variables factorizes into the product of layerwise conditional distributions. As the superscript "CDM" in the conditional distributions in (3)–(4) implies, each conditional distribution conforms to a CDM. With a slight abuse of notation, we next also write the observed layer $\mathbf{R}$ as $\mathbf{A}^{(0)}$, so that all of the layerwise conditionals can be written uniformly as $\mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}, \mathbf{Q}^{(d)})$, for $d = 1, \ldots, D$. In the following, we define specific DeepCDMs based on which diagnostic model the layerwise conditionals follow.

**Example 1** (DeepDINA)**.** The DINA model proposed by Junker and Sijtsma (2001) is a popular and fundamental model that adopts the conjunctive assumption. DINA assumes that students are expected to answer an item correctly only when they possess all required attributes of the item (i.e., the item's parent attributes in the graphical model). Our DeepDINA model adopts the conjunctive assumption *for each layer*'s conditional distribution. In particular, the conditional distribution of $A_j^{(d-1)}$ given its parent variables is

$$\begin{aligned} &\mathbb{P}^{\text{DINA}}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha},\ \mathbf{Q}^{(d)},\ \boldsymbol{c}^{(d)},\ \boldsymbol{g}^{(d)}) \\ &= (1 - s_j^{(d)}) \cdot \mathbb{1}\left(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j^{(d)}\right) + g_j^{(d)} \cdot \mathbb{1}\left(\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j^{(d)}\right) \end{aligned} \tag{7}$$

where the notation "$\succeq$" means "elementwisely greater than or equal to", and "$\not\succeq$" means otherwise. The $\mathbb{1}(\cdot)$ denotes a binary indicator function. The parameters $\boldsymbol{s}^{(d)} = (s_1^{(d)}, \ldots, s_{K_{d-1}}^{(d)})$ and $\boldsymbol{g}^{(d)} = (g_1^{(d)}, \ldots, g_{K_{d-1}}^{(d)})$ can be thought of as "quasi" slipping and guessing parameters, respectively. The interpretation of DeepDINA in an educational context is that, students are expected to master a skill (or a meta-skill) only when they possess all its higher-order parent skills in the probabilistic graphical model. Similar to Junker and Sijtsma (2001), we assume $g_j^{(d)} < 1 - s_j^{(d)}$ for each $j$ and $d$. This constraint can be interpreted as: comparing the subjects who master all the parent skills of an attribute $A_j^{(d-1)}$ and the subjects who do not, the former ones have higher probability of mastering this skill $A_j^{(d-1)}$ itself.

The interpretation of DeepDINA in Example 1 that students are expected to master a

skill when possessing all its higher-order parent skills may appear similar to the attribute hierarchy method (AHM; Gierl et al., 2007; Templin and Bradshaw, 2014). However, we point out that the AHM and DeepCDMs are not directly comparable, because the former assumes that the attributes can be directly connected to items whereas the latter assume high-order latent structures organized in multiple layers. Another modeling difference is that DeepDINA does not impose hard constraints on which attribute patterns are permissible as in AHMs. The quasi-guessing parameters $\boldsymbol{g}^{(d)}$ in DeepDINA the probabilities that a student masters lower-level skills even when lacking their parent meta-skills.

**Example 2** (DeepDINO)**.** The DINO model proposed by Templin and Henson (2006) adopts a disjunctive assumption and assumes that subjects are expected to provide a positive response to an item as long as they possess at least one parent attribute. The DeepDINO model adopts the layerwise disjunctive assumption and has the following conditional:

$$\begin{aligned} &\mathbb{P}^{\text{DINO}}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha},\ \mathbf{Q}^{(d)},\ \boldsymbol{c}^{(d)},\ \boldsymbol{g}^{(d)}) \\ &= (1 - s_j^{(d)}) \cdot \mathbb{1}\left(\alpha_k = 1 \text{ for some } k \text{ for which } q_{j,k}^{(d)} = 1\right) \\ &\quad + g_j^{(d)} \cdot \mathbb{1}\left(\alpha_k = 0 \text{ for all } k \text{ for which } q_{j,k}^{(d)} = 1\right). \end{aligned} \tag{8}$$

As DINO is often applied to psychiatric diagnosis, the new DeepDINO can also be interpreted in this context as follows: patients are expected to exhibit a symptom (or meta-symptom) as long as they possess one of its higher-level "parent" symptoms or mental disorders.

**Example 3** (Main-effect DeepCDMs)**.** We use "Main-effect DeepCDMs" to generically refer to DeepCDMs in which the layerwise conditionals follow a main-effect diagnostic model. Specifically, a main-effect diagnostic model assumes that the probability of $A_j^{(d-1)} = 1$ depends on the main effects of those parent attributes through a link function $f(\cdot)$:

$$\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha},\ \mathbf{Q}^{(d)},\ \boldsymbol{\beta}^{(d)}) = f\left(\beta_{j,0}^{(d)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(d)} \left\{q_{j,k}^{(d)} \alpha_k\right\}\right). \tag{9}$$

Note that not all the $\beta_{j,k}^{(d)}$ in the above equation are needed in the model specification. Only if $q_{j,k}^{(d)} = 1$ will the corresponding $\beta_{j,k}^{(d)}$ be incorporated in the model. When the link function $f$ is the identity, (9) gives the Additive Cognitive Diagnosis Model (ACDM; de la Torre,

2011); when $f$ is the inverse logit function, (9) gives the Logistic Linear Model (LLM; Maris, 1999); yet another parametrization of (9) gives rise to the Reduced Reparameterized Unified Model (R-RUM; DiBello et al., 1995).

**Example 4** (All-effect DeepCDMs)**.** We use "All-effect DeepCDMs" to refer to DeepCDMs in which the layerwise conditionals follow an all-effect diagnostic model. An all-effect diagnostic model assumes that the probability of $A_j^{(d-1)} = 1$ depends on all of the possible main effects and interaction effects of the parent attributes:

$$\begin{aligned}\mathbb{P}(A_j^{(d-1)} = 1 \mid \mathbf{A}^{(d)} = \boldsymbol{\alpha},\ \mathbf{Q}^{(d)},\ \boldsymbol{\beta}^{(d)}) = f\Big(\beta_{j,0}^{(d)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(d)} \left\{q_{j,k}^{(d)} \alpha_k\right\} \\ + \sum\nolimits_{1\le k_1<k_2\le K_d} \beta_{j,k_1k_2}^{(d)} \left\{q_{j,k_1}^{(d)} \alpha_{k_1}\right\}\left\{q_{j,k_2}^{(d)} \alpha_{k_2}\right\} + \cdots + \beta_{j,12\cdots K_d}^{(d)} \prod\nolimits_{k=1}^{K_d} \left\{q_{j,k}^{(d)} \alpha_k\right\}\Big).\end{aligned} \tag{10}$$

Similar to Example (3), not all the $\beta$-coefficients in the above equation are needed to specify the model. In particular, if $\boldsymbol{q}_j^{(d)}$ contains $K_j$ ones, then $2^{K_j}$ parameters are needed in (10). When the link function $f$ is the identity, (9) gives the Generalized DINA model (GDINA; de la Torre, 2011); when $f$ is the inverse logit, (9) gives the Log-linear CDM (LCDM; Henson et al., 2009); see the General Diagnostic Model (GDM) framework in von Davier (2008).

The parameters $\boldsymbol{c}^{(d)}$ and $\boldsymbol{g}^{(d)}$, $d = 1, \ldots, D$ in Examples 1–2 and $\boldsymbol{\beta}^{(d)}$, $d = 1, \ldots, D$ in Examples 3–4 are continuous parameters that help specify the conditional distribution of the binary variables in a DeepCDM. When $d = 1$, these parameters just resemble the item parameters in a traditional CDM. In a DeepDINA or DeepDINO, the number of continuous parameters required to model the latent attributes is $2\sum_{d=1}^{D-1} K_d + 2^{K_D} - 1$, while in a main-effect DeepCDM, this number is at most $\sum_{d=1}^{D-1} K_d(K_{d+1} + 1) + 2^{K_D} - 1$. We will discuss more about the remarkable reduction of parameter complexity in a DeepCDM in the end of Section 3, after our identifiability conditions imply upper bounds for $K_1, \ldots, K_D$.

We emphasize that the most flexible feature of the DeepCDM framework is that, different diagnostic models (including DINA, DINO, main-effect, and all-effect) can be flexibly combined in different layers of a DeepCDM. For example, in some practical applications, it may be desirable to adopt the most general all-effect diagnostic model for the bottom data layer for its flexibility in modeling the effects of the fine-grained attributes, whereas adopt

the simpler main-effect or DINA model in the deeper latent layers for their parsimony and interpretability. We call such DeepCDMs the *Hybrid DeepCDMs*. Hybrid DeepCDMs allow to balance the expressivity and parsimony of a model, and offer a wide range of possibilities to construct a specific diagnostic model based on substantive considerations.

The proposed DeepCDMs cover the latent tree models (Mourad et al., 2013) as a special case. In a latent tree model, each variable has at most one parent in a tree graph; whereas a DeepCDM allows for a general DAG, in which each variable can have multiple parents (e.g., variable $\mathbf{A}_2^{(1)}$ in Figure 1). In terms of the generative model, a pyramid-shaped DeepCDM is closely related to the Bayesian Pyramid proposed in Gu and Dunson (2023) and can be viewed as the latter adapted for diagnostic modeling goals. While the Bayesian Pyramid was implemented under the main-effect model and applied to extract genetic latent traits from DNA nucleotide sequences (Gu and Dunson, 2023), the DeepCDM framework is motivated by the need to hunt for deep diagnostic information and provides useful psychometric tools to this end. To better serve this goal, we develop a suite of methods and algorithms applicable to various layerwise diagnostic modeling assumptions; see Section 4 for details.

# 3 Identifiability Theory of DeepCDMs

Recently, there has been an emerging literature addressing the identifiability issues of CDMs (Xu and Zhang, 2016; Xu, 2017; Culpepper, 2019b; Fang et al., 2019; Chen et al., 2020; Gu and Xu, 2019, 2020, 2021). However, all of the above works focus on the saturated attribute model. The only exception in the CDM literature is Gu and Xu (2022), which establishes identifiability of hierarchical CDMs under attribute hierarchies; but as aforementioned, a CDM with an attribute hierarchy is a not a fully probabilistic measurement model, so their corresponding identifiability conditions do not apply to DeepCDMs. In this section, we propose transparent identifiability conditions for various DeepCDMs. In the most general exploratory model settings, our theory guarantees the identifiability of all $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$ and all continuous parameters in the model. When the $\mathbf{Q}$-matrices are known as in the confirmatory settings, all of our identifiability conclusions still directly apply.

## 3.1 Sharp Strict Identifiability Result for DeepDINA

DINA is one of the most basic and popular models in cognitive diagnosis. We establish sharp necessary and sufficient conditions for identifying the exploratory DeepDINA. Here "exploratory" means that the $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$ are not assumed to be known and fixed. Such an identifiability notion will be the most flexible and useful one in practice; see identifiability results for exploratory diagnostic models with a saturated attribute model in Chen et al. (2015), Xu and Shang (2018), Culpepper (2019b), Chen et al. (2020), and Gu and Xu (2021). Denote the parameter space for the deep proportion parameters $\boldsymbol{\pi}^{\text{deep}}$ by $\Delta^{2^{K_D}-1} = \{\pi^{\text{deep}}_{\boldsymbol{\alpha}_\ell} : \sum_{\ell=1}^{2^{K_D}-1} \pi^{\text{deep}}_{\boldsymbol{\alpha}_\ell} = 1, \pi^{\text{deep}}_{\boldsymbol{\alpha}_\ell} > 0\}$; throughout this work, we assume $\pi^{\text{deep}}_{\boldsymbol{\alpha}_\ell} > 0$ holds for every deep latent pattern $\boldsymbol{\alpha}_\ell \in \{0,1\}^{K_D}$. This is a common assumption also adopted for single-latent-layer CDMs. We next define the strict identifiability.

**Definition 1** (Strict Identifiability)**.** *An exploratory DeepCDM model 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* $\boldsymbol{\pi}^{\text{Deep}}$*, and all* $\mathbf{Q}$*-matrices at different depths* $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(D)}$*, up to some column/row permutation.*

The identifiability notion in Definition 1 that each $\mathbf{Q}$-matrix is identifiable up to some column/row permutation is a trivial and inevitable phenomenon when there exist multiple latent variables; see Chen et al. (2015) and Xu and Shang (2018).

Next, we summarize the existing necessary and sufficient identifiability conditions for the traditional DINA model with a saturated attribute model. These conditions will also play important roles in the identifiability of DeepDINA. Specifically, the following conditions (C), (R), and (D) are known to be necessary and sufficient for strict identifiability of DINA, both in the confirmatory case with a known $\mathbf{Q}$-matrix (Gu and Xu, 2019) and in the exploratory case with an unknown $\mathbf{Q}$-matrix (Gu and Xu, 2021):

(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.

We will call the above three conditions the C-R-D conditions for short. Our next theorem establishes sharp identifiability result for the exploratory DeepDINA with an arbitrary depth $D$, by providing 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 conditions in Theorem 1 are also necessary and sufficient for identifying the DeepDINO model introduced in Example 2, because of the duality between DINA and DINO (Chen et al., 2015). The sharp identifiability conditions in Theorem 1 put transparent constraints on the $\mathbf{Q}$-matrices, and equivalently, transparent constraints on the between-layer graphical structures. In a graphical model, define $X_m$ to be an *exclusive* child of $X_\ell$ if the former has the latter has its *only* parent. The deep C-R-D conditions in Theorem 1 can be translated into graphical language as follows: each latent variable in the deep graphical model should have at least one exclusive child (Condition (C)) and at least three children in total (not necessarily all exclusive; Condition (R)) in the layer below; and after removing one exclusive child for each latent variable, the remaining sets of children of the $K_d$ latent variables in the $d$th latent layer should be mutually distinct (Condition (D)) for $d = 1, \ldots, D$.

The following Example 5 illustrates the theoretical result in Theorem 1.

**Example 5.** *Consider a DeepDINA model with $D = 2$, and two $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$:*

$$\mathbf{Q}^{(1)} = \left(\begin{array}{c} \mathbf{I}_5 \\ \hline 0\ 0\ 0\ 0\ 1 \\ 1\ 1\ 0\ 1\ 0 \\ 1\ 0\ 1\ 1\ 0 \\ 0\ 1\ 1\ 1\ 1 \end{array}\right)_{9\times 5}, \quad \mathbf{Q}^{(2)} = \left(\begin{array}{c} 1\ 0 \\ 0\ 1 \\ \hline 1\ 0 \\ 0\ 1 \\ 1\ 1 \end{array}\right)_{5\times 2}.$$

*It is easy to verify that both $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ satisfy the C-R-D conditions. Therefore, a ladder-shaped DeepDINA model with $J = 9$ observed response variables, $K_1 = 5$ finest-grained latent attributes, and $K_2 = 2$ meta latent attributes in the deepest layer, is strictly identifiable. The*

*identifiable quantities include* $\mathbf{Q}$*-matrices* $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, *deepest proportion parameters* $\boldsymbol{\pi}^{\text{deep}}_{4\times 1}$, *(quasi-)slipping and guessing parameters at both layers* $(\boldsymbol{s}^{(1)}_{9\times 1}, \boldsymbol{g}^{(1)}_{9\times 1})$ *and* $(\boldsymbol{s}^{(2)}_{5\times 1}, \boldsymbol{g}^{(2)}_{5\times 1})$.

As can be seen from the toy example in Example 5, we have $J > K_1 > K_2$ under the identifiable DeepDINA there. In general, if a $\mathbf{Q}$-matrix of size $J \times K$ satisfies the C-R-D conditions, then there is a natural constraint on how large $K$ can be with respect to $J$: $J > K + \lceil \log_2(K) \rceil$ (Gu and Xu, 2021). This means in an identifiable DeepDINA, the sizes of the layers in the graphical model should satisfy $J > K_1 + \lceil \log_2(K_1) \rceil$, and $K_{d-1} > K_d + \lceil \log_2(K_d) \rceil$ for $d = 2, \ldots, D$. This suggests an increasingly shrinking ladder architecture of the latent layers when going deeper.

## 3.2 Strict Identifiability Result for General DeepCDMs

This subsection provides fully general strict identifiability conditions for a arbitrary DeepCDM. These conditions are also applicable to Hybrid DeepCDMs introduced in Section 2.3. From the identifiability result for DeepDINA in Theorem 1, one can see that it is those between-layer $\mathbf{Q}$-matrices that drive and deliver identifiability. In fact, this is correct intuition that applies much more broadly. Next, we formalize this intuition by establishing a general identifiability result for an arbitrary DeepCDM.

**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 suffices 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* $\mathbf{Q}^{(d)} = [\mathbf{I}_{K_d}, \mathbf{I}_{K_d}, \mathbf{I}_{K_d}, (\mathbf{Q}^{(d)*})^\top]^\top$ *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.*

**Remark 1.** Condition (S) in Theorem 2 is similar to the conditions in Theorem 4 in Gu and Dunson (2023) for identifying the Bayesian Pyramid model there. Condition (S$^*$) in Theorem 2 relaxes the requirement on $\mathbf{Q}$-matrices compared to Condition (S), and impose an additional requirement on the conditional probabilities to establish identifiability. Condition (S$^*$) is similar to conditions (C1) and (C2) in Culpepper (2019b) imposed on the traditional $\mathbf{Q}$-matrix, which were proposed to identify an exploratory diagnostic model for ordinal responses with a one-latent-layer saturated attribute model.

Theorem 2 is fully general, and is applicable regardless of which specific diagnostic model each layer in a DeepCDM follows. According to the conditions in Theorem 2, the sizes of the layers in the graphical model should satisfy $J > 2K_1$, and $K_{d-1} > 2K_d$ for $d = 2, \ldots, D$, which also suggests an increasingly shrinking sparse latent ladder when going deeper.

Comparing the conditions in Theorems 1 and 2, one can see that the general sufficient conditions for an arbitrary DeepCDM are stronger than those needed for identifying the DeepDINA. The next proposition further guarantees that if a DeepCDM consists of a mix of DINA-layers and main-effect/all-effect layers, then those $\mathbf{Q}$-matrices corresponding to 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.*

Proposition 1 reveals a key technical insight that our identifiability proofs leverage. That is, identifiability of DeepCDMs can be examined and established in a layer-by-layer manner, from the bottom up. This seemingly intuitive argument is rigorously true thanks to the probabilistic formulation of the directed graphical model and the discreteness nature of all the latent variables. See the proof of Theorem 1 in the Supplementary Material for details.

### 3.3 Generic Identifiability of Main-effect and All-effect DeepCDMs

Strict identifiability is the strongest possible identifiability notion, requiring parameters to be everywhere identifiable in their parameter space $\mathcal{T}$. A slightly weaker notion called *generic identifiability* (Allman et al., 2009), instead requires parameters to be identifiable almost everywhere in $\mathcal{T}$, allowing identifiability to fail on a measure-zero subset $\mathcal{N}$ of $\mathcal{T}$. As pointed out by Allman et al. (2009), generic identifiability often suffices for real data analyses purposes and is a very useful identifiability notion in practice. In the CDM literature, Gu and Xu (2020) and Chen et al. (2020) proposed generic identifiability conditions for variants of CDMs with a saturated attribute model. Next, we build on the existing generic identifiability conditions to establish generic identifiability of main-effect and all-effect DeepCDMs. We define the *main-effect-based* DeepCDMs as follows.

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

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

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

Note that DeepDINA and DeepDINO are not main-effect-based DeepCDMs, because they do not contain the main-effect coefficients such as those $\beta_{j,k}^{(d)}$ in Definition 2. These main-effect coefficients are essential to generic identifiability and allow for relaxing the condition that each $\mathbf{Q}^{(d)}$ should contain a submatrix $\mathbf{I}_{K_d}$ (Gu and Xu, 2020; Chen et al., 2020). We next formally define and establish 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*

$$\Omega_{\text{main}}(\boldsymbol{\beta}^{(d)};\, \mathbf{Q}^{(d)}) = \{\beta_{j,k}^{(d)} \neq 0 \text{ if } q_{j,k}^{(d)} = 1;\ \text{and } \beta_{j,k}^{(d)} = 0 \text{ if } q_{j,k}^{(d)} = 0\}. \tag{11}$$

*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 significantly relaxes the strict identifiability conditions in Theorem 2, by not requiring any $\mathbf{Q}^{(d)}$ to contain an identity submatrix $\mathbf{I}_{K_d}$. Note that these generic identifiability conditions in Theorem 3 also imply a shrinking latent ladder when going deeper, because (G1) and (G2) implicitly requires $J>2K_1$ and $K_d>2K_{d+1}$ for $d=1,\ldots,D-1$.

The natural upper bounds on the values of $K_1$, $K_2$, ... given by all of our identifiability conditions further confirms the statistical parsimony of DeepCDMs. For example, in a two-latent-layer DeepCDM with $K_1=7$ latent variables in the shallower latent layer and $K_2=2$ ones in the deeper layer (which is the scenario in the real data analysis in Section 6), the number of parameters required by DeepLLM is $\sum_{k=1}^{K_1}(\sum_{\ell=1}^{K_2}q_{k,\ell}^{(2)}+1)+2^{K_2}-1$, which is at most 24, and that required by DeepDINA is $2K_1+2^{K_2}-1=17$; whereas the number of parameters required in a saturated attribute model would be $2^{K_1}-1=127$. Such a remarkable reduction of parameter complexity facilitates applying DeepCDMs when there is a large number of fine-grained latent attributes but a relatively small sample size.

The easily understandable and intuitively interpretable identifiability conditions presented in this section are an appealing property of DeepCDMs. We next provide some insights into our proof strategy. The reason why we can establish identifiability in a layer-by-layer manner is two-fold. *First*, in a mutilayer directed graphical model, when arrows are all top-down and only occur between adjacent layers, marginalizing out all the latent variables deeper than the shallowest layer result in a marginal restricted latent class model (RLCM; Xu, 2017; Gu and Xu, 2020). Once the proportion parameters for this RLCM are identifiable, this shallowest latent layer's distribution is uniquely identified and can be theoretically treated as if observed when investigating identifiability of deeper layers. *Second*, we exploit one key property of existing identifiability theory of RLCMs – identifiability holds under conditions on the $Q$-matrix for *arbitrary marginal distributions* of the latent attributes. This property allows us to extend the identifiability conclusion to very flexible deep models since deeper layers could induce quite complex marginal dependencies among the latent attributes. Although proving identifiability is not technically very challenging upon realizing the above two key facts, we believe that uncovering these two facts to rigorously show identifiability still contributes to our understanding about CDMs and their potential.

On a related note, the HO-CDM proposed by de la Torre and Douglas (2004) is a very popular and widely used high-order CDM. However, whether and when parameters in a general HO-CDM with multiple higher-order continuous latent traits are fully identifiable is still unknown. So there currently lacks a rigorous statistical justification for valid parameter estimation in that model. To our best knowledge, DeepCDMs are the first higher-order CDMs that are shown to be fully identifiable.

# 4 Bayesian Inference for DeepCDMs

Recently, Bayesian formulation and estimation of CDMs have gained increasing interest; see Culpepper (2015), Chen et al. (2018), Fang et al. (2019), Chen et al. (2020), and Liu et al. (2020), among others. Bayesian approaches can incorporate prior beliefs into the model formulation, and quantify the statistical uncertainty through the posterior distributions. Moreover, in the CDM context, Bayesian estimation algorithms can conveniently incorpo-

rate meaningful constraints into the posterior sampling process, including the monotonicity constraints on the model parameters (Culpepper, 2015) and the identifiability constraints on the $\mathbf{Q}$-matrix (Chen et al., 2018).

In this section, we propose Bayesian formulations for several DeepCDMs and develop their corresponding efficient Gibbs sampling algorithms. As mentioned earlier, in this work we focus on developing Bayesian inference methods for the confirmatory setting with fixed and known $\mathbf{Q}$-matrices. For simplicity of presentation but without loss of generality, this section focuses on two-latent-layer DeepCDMs. We point out that all of our Bayesian inference procedures can be extended to a DeepCDM with more latent layers; this is the case thanks to both the conditional independence of non-adjacent layers in a DeepCDM and our layerwise Gibbs sampling steps. Now consider a two-latent-layer DeepCDM with $K_1$ fine-grained attributes and $K_2$ deeper meta attributes. With a sample of size $N$, denote the $N \times K_1$ first-layer latent attribute matrix by $(a_{ij}^{(1)})$, and denote the $N \times K_2$ second-layer latent variable matrix by $(a_{ij}^{(2)})$. Denote the $i$th row of these two matrices by $\boldsymbol{a}_i^{(1)}$ and $\boldsymbol{a}_i^{(2)}$, respectively. Let $\boldsymbol{\theta}^{(d)}$ generically denote the continuous parameters needed to specify the conditional distribution $\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}$.

## 4.1 Bayesian Inference for DeepDINA

For any positive integer $M$, we denote $[M] = \{1, \ldots, M\}$. The following continuous parameters are needed to specify a two-latent-layer DeepDINA: item parameters $\boldsymbol{\theta}^{(1)} = (\boldsymbol{s}_{J\times 1}^{(1)}, \boldsymbol{g}_{J\times 1}^{(1)})$, quasi-item parameters $\boldsymbol{\theta}^{(2)} = (\boldsymbol{s}_{K_1\times 1}^{(2)}, \boldsymbol{g}_{K_1\times 1}^{(2)})$, and deep proportion parameters $\boldsymbol{\pi}^{\text{deep}} = (\pi_1, \ldots, \pi_{2^{K_2}})$. Consider a sample of size $N$ and denote the observed $N \times J$ data matrix by $\mathcal{R} = (r_{ij})$. Define a $K_2$-dimensional vector $\boldsymbol{v}^{(2)} = (2^{K_2-1}, 2^{K_2-2}, \ldots, 2^0)^\top$, then $\boldsymbol{v}^{(2)}$ induces a bijection between the binary patterns and integers (Culpepper, 2019a), and we define binary patterns $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{2^{K_2}} \in \{0,1\}^{K_2}$ such that $\boldsymbol{\alpha}_\ell^\top \boldsymbol{v}^{(2)} = \ell - 1$, for $\ell = 1, \ldots, 2^{K_2}$.

When $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ are fixed, DeepDINA has the following model formulation,

$$r_{ij} \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{q}_j^{(1)}, \boldsymbol{\theta}^{(1)} \sim \text{Bernoulli}\left(\left(1 - s_j^{(1)}\right)^{\xi_{1,ij}} \left(g_j^{(1)}\right)^{1-\xi_{1,ij}}\right), \; \xi_{1,ij} = \mathbb{1}(\boldsymbol{a}_i^{(1)} \succeq \boldsymbol{q}_j^{(1)}); \quad (12)$$

$$a_{ik}^{(1)} \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{q}_k^{(2)}, \boldsymbol{\theta}^{(2)} \sim \text{Bernoulli}\left(\left(1-s_k^{(2)}\right)^{\xi_{2,ik}}\left(g_k^{(2)}\right)^{1-\xi_{2,ik}}\right),\ \xi_{2,ik} = \mathbb{1}(\boldsymbol{a}_i^{(2)} \succeq \boldsymbol{q}_k^{(2)}); \quad (13)$$

$$p(s_j^{(d)}, g_j^{(d)}) \propto (s_j^{(d)})^{a_s-1}(1-s_j^{(d)})^{b_s-1}(g_j^{(d)})^{a_g-1}(1-g_j^{(d)})^{b_g-1} \cdot \mathbb{1}(g_j^{(d)} + s_j^{(d)} < 1),$$
$$j \in [J] \text{ for } d=1, \text{ and } j \in [K_1] \text{ for } d=2; \quad (14)$$

$$p(\boldsymbol{a}_i^{(2)} \mid \boldsymbol{\pi}^{\text{deep}}) \propto \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell)},\ 0 \le \pi_\ell \le 1,\ \sum_{\ell=1}^{2^{K_2}} \pi_\ell = 1; \quad p(\boldsymbol{\pi}^{\text{deep}}) = \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\delta_\ell - 1}. \quad (15)$$

The prior for $\boldsymbol{\pi}^{\text{deep}} = (\pi_1, \ldots, \pi_{2^{K_2}})$ in (15) is the Dirichlet distribution with parameters $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_L)$. The prior for $s_j^{(d)}, g_j^{(d)}$ in (14) is a product of two truncated Beta densities with hyperparameters $(a_s, b_s)$ and $(a_g, b_g)$, respectively, similar to that in Culpepper (2015). The monotonicity constraint $g_j^{(d)} < 1 - s_j^{(d)}$ in (14) ensures each item or attribute provides information to differentiate the capable and incapable subjects (Junker and Sijtsma, 2001).

The above Bayesian formulation of DeepDINA facilitates convenient posterior inference via a Gibbs sampler. Specifically, we sample each entry $a_{i,k}^{(1)}$ individually to better leverage the multilayer generative process and to boost computational efficiency; this is different from sampling the entire latent vector $\boldsymbol{a}_i^{(1)}$ as in many previous Bayesian estimation approaches for CDMs. Define $\boldsymbol{a}_{i,-k}^{(1)}$ to be the $(K_1-1)$-dimensional subvector of $\boldsymbol{a}_i^{(1)}$ containing entries other than $a_{i,k}^{(1)}$. The full conditional distribution of $a_{i,k}^{(1)}$ is:

$$k = 1, \ldots, K_1: \quad \mathbb{P}(a_{i,k}^{(1)} = 1 \mid -) = \mathbb{P}(a_{i,k}^{(1)} = 1 \mid \boldsymbol{r}_i, \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$$
$$= \frac{\mathbb{P}(a_{i,k}^{(1)} = 1 \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)} = 1, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})}{\sum_{x=0,1}\mathbb{P}(a_{i,k}^{(1)} = x \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)} = x, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})};$$

In the above display, the "$-$" in the conditioning set for $a_{i,k}^{(1)}$ generically summarizes all of the other quantities in the posterior, and the first equality is derived from the conditional independence properties of the graphical model. As for the second latent layer $\boldsymbol{a}_i^{(2)}$, we sample it from the categorical posterior with $2^{K_2}$ components. The full conditional distribution of each element in $\boldsymbol{s}^{(1)}$, $\boldsymbol{g}^{(1)}$, $\boldsymbol{s}^{(2)}$, and $\boldsymbol{g}^{(2)}$ is a truncated Beta, and that of $\boldsymbol{\pi}^{\text{deep}}$ is a Dirichlet; we provide the detailed forms of these conditional distributions in the Supplementary Material.

## 4.2 Bayesian Inference for Hybrid GDINA-DINA

A two-latent-layer Hybrid GDINA-DINA model features a GDINA layer for modeling $\mathbf{R} \mid \mathbf{A}^{(1)}$ and a DINA layer for modeling $\mathbf{A}^{(1)} \mid \mathbf{A}^{(2)}$. Such a model may be useful in practical scenarios when it is desirable to adopt the general diagnostic model in the bottom layer for its flexibility and adopt a simpler DINA model in the deeper layer for its parsimony. The Hybrid GDINA-DINA model has the following generative process,

$$r_{ij} \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{q}_j^{(1)}, \boldsymbol{\theta}^{(1)} \sim \text{Bernoulli}\Big(\beta_{j,0}^{(1)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(1)} \left\{q_{j,k}^{(1)} a_{i,k}^{(1)}\right\} + \sum_{1 \le k_1 < k_2 \le K_d} \beta_{j,k_1k_2}^{(1)} \left\{q_{j,k_1}^{(1)} a_{i,k_1}^{(1)}\right\} \left\{q_{j,k_2}^{(1)} a_{i,k_1}^{(1)}\right\} + \cdots + \beta_{j,12\cdots K_d}^{(1)} \prod_{k=1}^{K_d} \left\{q_{j,k}^{(1)} a_{i,k}^{(1)}\right\}\Big); \tag{16}$$

$$a_{ik}^{(1)} \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{q}_k^{(2)}, \boldsymbol{\theta}^{(2)} \sim \text{Bernoulli}\left(\left(1 - s_k^{(2)}\right)^{\xi_{2,ik}} \left(g_k^{(2)}\right)^{1-\xi_{2,ik}}\right), \ \xi_{2,ik} = \mathbb{1}\left(\boldsymbol{a}_i^{(2)} \succeq \boldsymbol{q}_k^{(2)}\right). \tag{17}$$

Since $\mathbf{A}^{(1)} \mid \mathbf{A}^{(2)}$ follows the DINA model, we adopt the same truncated Beta priors as that in (14) for the quasi-item parameters and enforce $g_k^{(2)} < 1 - s_k^{(2)}$. As for the model for $\mathbf{R} \mid \mathbf{A}^{(1)}$, we adopt the GDINA formulation proposed by de la Torre (2011) in (16) by using the identity link function $f(\cdot)$ in the all-effect general diagnostic model. A general diagnostic model with an identity link facilitates Gibbs sampling steps without data augmentation. Note that in order to perform Gibbs sampling directly, it is not convenient to directly work with the $\beta$-coefficients in (16) and sample from their posteriors. Instead, similar to the existing GDINA EM algorithm in the literature, we adopt an invertible reparameterization of the $\beta$-coefficients and define a set of $\theta$-coefficients that directly correspond to conditional correct response probabilities and are easy to sample from. Define $\mathcal{K}_j = \{k \in [K] : \ q_{j,k}^{(1)} = 1\}$, which is the set of indices of the latent attributes that item $j$ measures. Then each $\beta$-coefficient in the GDINA layer in (16) can be equivalently written as $\beta_{j,S}^{(1)}$, where $S$ is a subset of $\mathcal{K}_j$; for example, $\beta_{j,\varnothing}^{(1)} = \beta_{j,0}^{(1)}$, $\beta_{j,\{k\}}^{(1)} = \beta_{j,k}^{(1)}$, and $\beta_{j,\mathcal{K}_j}^{(1)}$ corresponds to the parameter for highest order interaction effect of the required attributes. For any subset $S \subseteq \mathcal{K}_j$, denote by $\boldsymbol{q}_{j,S}^{(1)} := (q_{j,k}^{(1)};\ k \in S)$ the subvector of $\boldsymbol{q}_j$. We now define the $\theta$-parameters as follows,

$$\theta_{j,S}^{(1)} = \sum_{S' \subseteq S} \beta_{j,S'}^{(1)} \overset{(\star)}{=} \mathbb{P}(r_{i,j} = 1 \mid \boldsymbol{a}_i^{(1)\top} \boldsymbol{q}_{j,S}^{(1)} = \boldsymbol{q}_{j,S}^{(1)\top} \boldsymbol{q}_{j,S}^{(1)}), \tag{18}$$

$$\forall S \subseteq \mathcal{K}_j = \{k \in [K] : \ q_{j,k}^{(1)} = 1\},$$

where the equality indexed by "$(\star)$" can be verified by simply following the definition of the $\beta$-parameters. For example, $\theta_{j,\{k\}}^{(1)} = \beta_{j,\varnothing}^{(1)} + \beta_{j,\{k\}}^{(1)}$ represents the probability of providing positive response to item $j$ given that the subject only masters the $k$th latent attribute $A_k^{(1)}$. With the above reparametrization and equality "$(\star)$", the $\theta$-parameters directly represent positive response probabilities of certain clearly defined latent classes in the population. This structure implies that we can endow $\theta_{j,S}$ with a Beta prior and then have a Beta posterior. In particular, let the prior for $\theta_{j,S}^{(1)}$ be $\text{Beta}(a_\theta, b_\theta)$, then its posterior distribution is

$$\text{Beta}\left(a_\theta + \sum_{i=1}^{N} r_{i,j} \mathbb{1}\left(\boldsymbol{a}_i^{(1)\top} \boldsymbol{q}_{j,S}^{(1)} = \boldsymbol{q}_{j,S}^{(1)\top} \boldsymbol{q}_{j,S}^{(1)}\right),\ b_\theta + \sum_{i=1}^{N} (1 - r_{i,j}) \mathbb{1}\left(\boldsymbol{a}_i^{(1)\top} \boldsymbol{q}_{j,S}^{(1)} = \boldsymbol{q}_{j,S}^{(1)\top} \boldsymbol{q}_{j,S}^{(1)}\right)\right),$$

where $S$ ranges in all the possible subsets of $\mathcal{K}_j$. This completes the description on how to sample the continuous parameters for the GDINA layer.

Interpretable monotonicity constraints can also be incorporated into the posterior sampling of the $\theta_{j,S}^{(1)}$ parameters. For example, it may be reasonable to impose the constraint that the main-effect parameters of the attributes, i.e, $\beta_{j,k}^{(1)}$ in (9), are positive (Culpepper, 2019b). In our parametrization of $\theta_{j,S}^{(1)}$, this constraint is equivalent to requiring $\theta_{j,\{k\}}^{(1)} > \theta_{j,\varnothing}^{(1)}$ for each $k = 1, \ldots, K_1$. Such a constraint can be easily enforced by sampling $\theta_{j,\{k\}}^{(1)}$ from a truncated Beta posterior as follows:

$$\text{Beta}\left(a_\theta + \sum_{i=1}^{N} r_{i,j} \mathbb{1}\left(a_{i,k}^{(1)} q_{j,k}^{(1)} = q_{j,k}^{(1)}\right), b_\theta + \sum_{i=1}^{N} (1 - r_{i,j}) \mathbb{1}\left(a_{i,k}^{(1)} q_{j,k}^{(1)} = q_{j,k}^{(1)}\right)\right) \cdot \mathbb{1}(\theta_{j,\{k\}}^{(1)} > \theta_{j,\varnothing}^{(1)}).$$

we provide the details of the Gibbs sampler for the Hybrid GDINA-DINA in the Supplementary Material.

## 4.3 Bayesian Inference for DeepLLM

In this subsection, we consider the two-latent-layer Deep Logistic Linear Model (DeepLLM). Let $\sigma(x) = 1/(1+e^{-x})$ denote the inverse logit function (i.e., sigmoid function). For $\boldsymbol{a}_2^{(i)}$ and

$\boldsymbol{\pi}^{\text{deep}}$, we adopt the same formulation and prior as (15). As for the additional parameters in a DeepLLM, we adopt the following formulation,

$$r_{ij} \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{q}_j^{(1)}, \boldsymbol{\theta}^{(1)} \sim \text{Bernoulli}\left(\sigma\left(\beta_{j,0}^{(1)} + \sum\nolimits_{k=1}^{K_d} \beta_{j,k}^{(1)} q_{j,k}^{(1)} a_{i,k}\right)\right); \tag{19}$$

$$a_{ik}^{(1)} \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{q}_k^{(2)}, \boldsymbol{\theta}^{(2)} \sim \text{Bernoulli}\left(\sigma\left(\beta_{k,0}^{(2)} + \sum\nolimits_{m=1}^{K_2} \beta_{k,m}^{(2)} q_{k,m}^{(1)} a_{i,m}\right)\right); \tag{20}$$

$$\beta_{j,k}^{(1)} \mid q_{j,k}^{(1)} = 1 \sim N(0, \sigma_\beta^2) \cdot \mathbb{1}(\beta_{j,k}^{(1)} > 0), \ \beta_{k,m}^{(2)} \mid q_{k,m}^{(2)} = 1 \sim N(0, \sigma_\beta^2) \cdot \mathbb{1}(\beta_{k,m}^{(2)} > 0). \tag{21}$$

The natural constraints imposed by the $\mathbf{Q}$-matrices $(\beta_{j,k}^{(1)} \mid q_{j,k}^{(1)} = 0) \equiv 0$ and $(\beta_{k,m}^{(2)} \mid q_{k,m}^{(2)} = 0) \equiv 0$ can be readily enforced throughout the sampling process. In order to facilitate efficient Gibbs sampling steps based on full conditional distributions of all the parameters, we propose to use the Polya-Gamma data augmentation in Polson et al. (2013). This data augmentation strategy was also recently adopted for Bayesian Pyramids for multivariate categorical data in Gu and Dunson (2023) and for saturated CDMs in Balamuta and Culpepper (2022). Different from these existing works, we apply Polya-Gamma augmentation not only for observed data layer $\mathbf{R}$, but also for the latent layer $\mathbf{A}^{(1)}$, due to our multilayer logistic linear model assumption. Specifically, we introduce auxiliary variables $w_{i,j}^{(1)}$ for $j \in [J]$, $w_{i,k}^{(2)}$ for $k \in [K_1]$ that follow the Polya-Gamma prior $\text{PG}(1, 0)$. Introduce the following notation:

$$\phi_{i,j}^{(1)} = \beta_{j,0}^{(1)} + \sum\nolimits_{k=1}^{K_1} \beta_{j,k}^{(1)} q_{j,k}^{(1)} a_{i,k}^{(1)}, \quad \phi_{i,k}^{(2)} = \beta_{k,0}^{(2)} + \sum\nolimits_{m=1}^{K_2} \beta_{k,m}^{(2)} q_{k,m}^{(2)} a_{i,m}^{(2)}.$$

Denote the probability density function of $\text{PG}(1, 0)$ by $p^{\text{PG}}(w \mid 1, 0)$. By the property of the Polya-Gamma variables in Polson et al. (2013), we have the following identity for $\phi_{i,j}^{(1)}$:

$$\frac{\exp(\phi_{i,j}^{(1)} r_{i,j})}{1 + \exp(\phi_{i,j}^{(1)})} = 2 \exp\left\{(r_{i,j} - 1/2)\phi_{i,j}^{(1)}\right\} \int_0^\infty \exp\left\{-w_{i,j}^{(1)} (\phi_{i,j}^{(1)})^2 / 2\right\} p^{\text{PG}}(w_{i,j}^{(1)} \mid 1, 0) \mathrm{d} w_{i,j}^{(1)};$$

and there is a similar identity for $\phi_{i,k}^{(2)}$. A nice consequence of the above equality is that the conditional posterior distributions for all the $\beta_{j,0}^{(1)}$ and $\beta_{j,k}^{(1)}$ are still Gaussian, and the conditional posterior distribution of each $w_{i,j}^{(1)}$ is still Polya-Gamma, with $(w_{i,j}^{(1)} \mid -) \sim \text{PG}(1, \phi_{i,j}^{(1)})$. Similar posterior forms can be derived for $\beta_{k,m}^{(2)}$ and $w_{i,k}^{(2)}$, which are also Gaussian

and Poyla-Gamma, respectively. Such posterior distributions are easy to sample from and are the building blocks of our efficient Gibbs sampler for a DeepLLM. We provide the details of this Gibbs sampler for DeepLLM in the Supplementary Material.

We point out that our Gibbs samplers described in Sections 4.1–4.3 can be readily extended to deeper models containing more than two latent layers. To see this, note that DeepCDMs have a nice property implied by the graphical model: given any layer $\mathbf{A}^{(d)}$, the layer above it $\mathbf{A}^{(d+1)}$ and the layer below it $\mathbf{A}^{(d-1)}$ are conditionally independent. This means in a DeepCDM with an arbitrary number of layers, when sampling parameters and latent structures for any specific layer, we only need to consider its two adjacent layers and derive the full conditional distributions based on these local model information. This fact allows straightforward extensions of our Gibbs sampling procedures to general hybrid DeepCDMs.

# 5 Simulation Studies

We conduct simulation studies for the three two-latent-layer DeepCDMs considered in Section 4: DeepDINA in Section 4.1, Hybrid GDINA-DINA in Section 4.2, and DeepLLM in Section 4.3. We also conduct two additional simulation studies, one comparing a DeepCDM to a traditional CDM with a saturated attribute model, and one evaluating a DeepCDM's robustness to deeper layer model misspecification. The following three different generative graphical structures (equivalently, forms of $\mathbf{Q}^{(1)}_{J\times K_1}$ and $\mathbf{Q}^{(2)}_{K_1\times K_2}$) are considered:

$$\text{structure (a):}\quad \mathbf{Q}^{(1)}_{30\times 6}=\begin{pmatrix} & & \mathbf{I}_6 & & & \\ & & \mathbf{I}_6 & & & \\ & & \mathbf{I}_6 & & & \\ & & \mathbf{I}_6 & & & \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix},\qquad \mathbf{Q}^{(2)}_{6\times 2}=\begin{pmatrix} \mathbf{I}_2 \\ \mathbf{I}_2 \\ \mathbf{I}_2 \end{pmatrix}; \tag{22}$$

$$\text{structure (b):}\quad \mathbf{Q}^{(1)}_{30\times 7} = \begin{pmatrix} & & & \mathbf{I}_7 & & & \\ & & & \mathbf{I}_7 & & & \\ & & & \mathbf{I}_7 & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \qquad \mathbf{Q}^{(2)}_{7\times 3} = \begin{pmatrix} & \mathbf{I}_3 & \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}; \tag{23}$$

$$\text{structure (c):}\quad \mathbf{Q}^{(1)}_{30\times 8} = \begin{pmatrix} & & & \mathbf{I}_8 & & & & \\ & & & \mathbf{I}_8 & & & & \\ & & & \mathbf{I}_8 & & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{Q}^{(2)}_{8\times 3} = \begin{pmatrix} & \mathbf{I}_3 & \\ & \mathbf{I}_3 & \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}. \tag{24}$$

Denote the above three pairs of $\mathbf{Q}$-matrices by $\{\mathbf{Q}^{(1)}_a, \mathbf{Q}^{(2)}_a\}$, $\{\mathbf{Q}^{(1)}_b, \mathbf{Q}^{(2)}_b\}$, and $\{\mathbf{Q}^{(1)}_c, \mathbf{Q}^{(2)}_c\}$, respectively. In all the simulation experiments, the Gibbs sampling algorithm is run for 15,000 iterations, with the first 10,000 iterations discarded as burn-in. Based on the last 5000 posterior samples, we calculate the posterior means of the continuous parameters as their point estimators. We observed sufficiently good convergence and mixing behaviors of all the Gibbs samplers through preliminary simulations.

**Simulation Study I: Two-latent-layer DeepDINA.** Under each of the three pairs of $\mathbf{Q}$-matrices in (22)–(24), we specify the true item/quasi-item parameters to be $s^{(1)}_j = g^{(1)}_j = 0.1$ for all $j \in [J]$, and $s^{(2)}_k = g^{(2)}_k = 0.25$ for all $k \in [K_1]$. We specify the true deep proportion parameters to be $\boldsymbol{\pi}^{\text{deep}} = (1/2^{K_2}, \ldots, 1/2^{K_2})$, that is, uniform over the $2^{K_2}$ deep latent patterns. We consider three sample sizes $N = 500, 1000, 2000$, and carry out 100 independent simulation replicates in each of the nine resulting simulation settings. The $\mathbf{Q}$-matrices $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ are fixed to the ground truths during estimation. We consider the

posterior means of the model parameters as their point estimators, and calculate the mean Root Mean Squared Errors (RMSE) and mean absolute biases (aBias), each averaged across the 100 simulation replicates. Here the mean absolute bias is a valid measure of the bias performance of an estimator, which is both broadly used in statistics (Morris et al., 2019) and also in previous studies about CDMs (Xu and Shang, 2018; Chen et al., 2020). Note that directly averaging the bias itself (instead of the absolute bias that we consider) across simulation replicates may give a misleading result, because positive and negative biases can cancel out each other. Table 1 presents the simulation results of the average RMSE and average aBias for the slipping and guessing parameters $\boldsymbol{\theta}^{(1)}_{\text{DINA}}$, for the quasi-slipping and quasi-guessing parameters $\boldsymbol{\theta}^{(2)}_{\text{DINA}}$, and the deep proportion parameters $\boldsymbol{\pi}^{\text{deep}}$.

| Structure | $(J,\ K_1,\ K_2)$ | $N$ | RMSE | | | aBias | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\boldsymbol{\theta}^{(1)}_{\text{DINA}}$ | $\boldsymbol{\theta}^{(2)}_{\text{DINA}}$ | $\boldsymbol{\pi}^{\text{deep}}$ | $\boldsymbol{\theta}^{(1)}_{\text{DINA}}$ | $\boldsymbol{\theta}^{(2)}_{\text{DINA}}$ | $\boldsymbol{\pi}^{\text{deep}}$ |
| (a) in (22) | (30, 6, 2) | 500 | 0.021 | 0.060 | 0.063 | 0.017 | 0.050 | 0.050 |
| | | 1000 | 0.015 | 0.046 | 0.049 | 0.012 | 0.038 | 0.040 |
| | | 2000 | 0.011 | 0.038 | 0.040 | 0.009 | 0.031 | 0.032 |
| (b) in (23) | (30, 7, 3) | 500 | 0.039 | 0.072 | 0.042 | 0.033 | 0.062 | 0.033 |
| | | 1000 | 0.033 | 0.070 | 0.047 | 0.029 | 0.061 | 0.038 |
| | | 2000 | 0.029 | 0.066 | 0.044 | 0.026 | 0.058 | 0.036 |
| (c) in (24) | (30, 8, 3) | 500 | 0.031 | 0.064 | 0.038 | 0.026 | 0.054 | 0.030 |
| | | 1000 | 0.026 | 0.060 | 0.037 | 0.022 | 0.051 | 0.029 |
| | | 2000 | 0.021 | 0.054 | 0.032 | 0.019 | 0.047 | 0.026 |

Table 1: Two-latent-layer DeepDINA simulation results.

Note that the three generative graph structures in (22)–(24) all satisfy the strict identifiability conditions for the DeepDINA model. Specifically, all the $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ satisfy the C-R-D conditions, therefore Theorem 1 guarantees the strict identifiability of the parameters $\boldsymbol{\theta}^{(1)}_{\text{DINA}}$, $\boldsymbol{\theta}^{(2)}_{\text{DINA}}$, and $\boldsymbol{\pi}^{\text{deep}}$. This identifiability conclusion is empirically confirmed by the simulation results in Table 1, where the estimation errors of these identifiable quantities measured through RMSE and aBias are all reasonably small.

**Simulation Study II: Two-latent-layer Hybrid GDINA-DINA.** Under the two-latent-layer Hybrid GDINA-DINA model, we specify the deeper DINA-layer's true parame-

ters to be the same as that in the DeepDINA case with $s_k^{(2)} = g_k^{(2)} = 0.25$ for all $k \in [K_1]$, and also specify the deep proportion parameters as $\boldsymbol{\pi}^{\text{deep}} = (1/2^{K_2}, \ldots, 1/2^{K_2})$. As for the GDINA-layer's parameters, we specify them in the same way as the simulations in Xu and Shang (2018) and Chen et al. (2020); that is, for each item $j \in [J]$, set the lowest correct response probability to 0.2 for all-zero attribute profiles, set the highest correct response probability to 0.8 for all-one attribute profiles, and set all the main-effect and interaction-effect parameters under the GDINA model to be equal. The above true parameter specification can be equivalently written in the following mathematical form,

$$\mathbb{P}^{\text{GDINA}}(R_j = 1 \mid \mathbf{A}^{(1)} = \boldsymbol{\alpha}, \boldsymbol{\beta}^{(1)}) = \theta_{j,S}^{(1)} = \sum_{S \subseteq \mathcal{K}_j} \beta_{j,S}^{(1)}, \quad \text{where } \mathcal{K}_j = \{k \in [K] : \ q_{j,k}^{(1)} = 1\};$$

$$\beta_{j,\varnothing}^{(1)} = 0.2, \quad \beta_{j,S}^{(1)} = (0.8 - 0.2)/(2^{|\mathcal{K}_j|} - 1) \ \text{ for } \ S \subseteq \mathcal{K}_j, S \neq \varnothing.$$

During the Bayesian posterior sampling process, we enforce the monotonicity constraint described in Section 4.2 by sampling the transformed parameters $\theta_{j,\{k\}} = \beta_{j,\varnothing}^{(1)} + \beta_{j,\{k\}}^{(1)}$ from the truncated Beta posteriors; this ensures the main-effect parameters $\beta_{j,\{k\}}^{(1)}$ to be positive. Table 2 presents the simulation results under the Hybrid GDINA-DINA model.

| Structure | $(J,\ K_1,\ K_2)$ | $N$ | RMSE | | | aBias | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\boldsymbol{\beta}_{\text{GDINA}}^{(1)}$ | $\boldsymbol{\theta}_{\text{DINA}}^{(2)}$ | $\boldsymbol{\pi}^{\text{deep}}$ | $\boldsymbol{\beta}_{\text{GDINA}}^{(1)}$ | $\boldsymbol{\theta}_{\text{DINA}}^{(2)}$ | $\boldsymbol{\pi}^{\text{deep}}$ |
| (a) in (22) | (30, 6, 2) | 500 | 0.046 | 0.064 | 0.059 | 0.037 | 0.052 | 0.047 |
| | | 1000 | 0.035 | 0.056 | 0.056 | 0.028 | 0.045 | 0.046 |
| | | 2000 | 0.025 | 0.042 | 0.044 | 0.020 | 0.033 | 0.036 |
| (b) in (23) | (30, 7, 3) | 500 | 0.056 | 0.073 | 0.045 | 0.045 | 0.058 | 0.036 |
| | | 1000 | 0.041 | 0.063 | 0.044 | 0.033 | 0.051 | 0.035 |
| | | 2000 | 0.030 | 0.052 | 0.039 | 0.024 | 0.042 | 0.031 |
| (c) in (24) | (30, 8, 3) | 500 | 0.052 | 0.069 | 0.043 | 0.042 | 0.056 | 0.034 |
| | | 1000 | 0.039 | 0.062 | 0.039 | 0.032 | 0.050 | 0.031 |
| | | 2000 | 0.028 | 0.050 | 0.033 | 0.023 | 0.040 | 0.027 |

Table 2: Two-latent-layer Hybrid GDINA-DINA simulation results.

Table 2 shows that our method can accurately estimate all the parameters under the Hybrid GDINA-DINA model and the estimation accuracy improves as sample size grows.

Indeed, all the $\mathbf{Q}_a^{(1)}$, $\mathbf{Q}_b^{(1)}$, and $\mathbf{Q}_c^{(1)}$ satisfy the identifiability conditions for general diagnostic models (condition S in Theorem 2), and all the $\mathbf{Q}_a^{(2)}$, $\mathbf{Q}_b^{(2)}$, and $\mathbf{Q}_c^{(2)}$ satisfy the C-R-D conditions for identifying the DINA model. Therefore, Proposition 1 guarantees that all the parameters $\boldsymbol{\beta}_{\text{GDINA}}^{(1)}$, $\boldsymbol{\theta}_{\text{DINA}}^{(1)}$, and $\boldsymbol{\pi}^{\text{deep}}$ in this Hybrid DeepCDM are fully identifiable, as supported by the numerical evidence in Table 2.

**Simulation Study III: Two-latent-layer DeepLLM.** We conduct simulations for the DeepLLM, using the Gibbs sampler with the multilayer Polya-Gamma data augmentation strategy developed in Section 4.3. The true parameters in the two-latent-layer DeepLLM are specified as follows. Inside the inverse logit function, the intercept parameters for the two layers are set to $\beta_{j,0}^{(1)} = -3$ for all $j \in [J]$ and $\beta_{k,0}^{(2)} = -2$ for all $k \in [K_1]$; the shallower layer's main-effect parameters are set to $\beta_{j,k}^{(1)} = 6/\left(\sum_{k'=1}^{K_1} q_{j,k'}^{(1)}\right)$ for which $q_{j,k}^{(1)} = 1$, and the deeper layer's main-effect parameters are set to $\beta_{k,m}^{(2)} = 4/\left(\sum_{m'=1}^{K_2} q_{k,m'}^{(2)}\right)$ for which $q_{k,m}^{(2)} = 1$. Note that these $\beta$-parameters in a DeepLLM are all inside the inverse logit function $f(x) = e^x/(1+e^x)$ to generate the correct response probability, so they are on a different scale than those probability parameters under the DINA or GDINA model. Table 3 presents the estimation accuracy results for the two-latent-layer DeepLLM model.

| Structure | $(J,\ K_1,\ K_2)$ | $N$ | RMSE | | | aBias | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\boldsymbol{\beta}_{\text{LLM}}^{(1)}$ | $\boldsymbol{\beta}_{\text{LLM}}^{(2)}$ | $\boldsymbol{\pi}^{\text{deep}}$ | $\boldsymbol{\beta}_{\text{LLM}}^{(1)}$ | $\boldsymbol{\beta}_{\text{LLM}}^{(2)}$ | $\boldsymbol{\pi}^{\text{deep}}$ |
| (a) in (22) | (30, 6, 2) | 500 | 0.360 | 0.339 | 0.026 | 0.284 | 0.262 | 0.021 |
| | | 1000 | 0.247 | 0.215 | 0.016 | 0.196 | 0.171 | 0.013 |
| | | 2000 | 0.175 | 0.161 | 0.011 | 0.139 | 0.129 | 0.009 |
| (b) in (24) | (30, 7, 3) | 500 | 0.362 | 0.514 | 0.031 | 0.284 | 0.402 | 0.025 |
| | | 1000 | 0.254 | 0.407 | 0.024 | 0.199 | 0.315 | 0.019 |
| | | 2000 | 0.181 | 0.303 | 0.018 | 0.144 | 0.228 | 0.015 |
| (c) in (24) | (30, 8, 3) | 500 | 0.376 | 0.491 | 0.023 | 0.294 | 0.379 | 0.018 |
| | | 1000 | 0.264 | 0.352 | 0.017 | 0.208 | 0.272 | 0.014 |
| | | 2000 | 0.187 | 0.212 | 0.012 | 0.149 | 0.166 | 0.010 |

Table 3: Two-latent-layer DeepLLM simulation results.

The simulation results in Table 2 also show decreasing estimation errors with growing sample sizes. We point out that the "RMSE" and "aBias" values in different tables are

not directly comparable, because the logistic-scale parameters $\boldsymbol{\beta}^{(2)}_{\text{LLM}}$ and $\boldsymbol{\beta}^{(2)}_{\text{LLM}}$ in Table 3 have larger magnitudes than DINA/GDINA parameters in the previous Tables 1–2. The three first-layer $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)}_a$, $\mathbf{Q}^{(1)}_b$, and $\mathbf{Q}^{(1)}_c$ all satisfy the identifiability conditions under general diagnostic models which cover the LLM as a special case, so the $\boldsymbol{\beta}^{(1)}_{\text{LLM}}$ are always identifiable across structures (a), (b), and (c) (see the layerwise identifiability argument in Proposition 1). As for the second-layer $\mathbf{Q}$-matrix in the three settings, $\mathbf{Q}^{(2)}_a$ and $\mathbf{Q}^{(2)}_c$ satisfy the strict identifiability conditions for LLM while $\mathbf{Q}^{(2)}_c$ satisfies the generic identifiability conditions for LLM. For quantities $\boldsymbol{\beta}^{(2)}_{\text{LLM}}$ and $\boldsymbol{\pi}^{\text{deep}}$ associated with $\mathbf{Q}^{(2)}$, Table 3 shows that their estimation errors in the generic identifiability case (b) are still reasonably small, though slightly worse than those in the strictly identifiable cases (a) and (c). Overall, all the above simulation results corroborate the identifiability conclusions about DeepCDMs, and also provide evidence that our Bayesian estimation algorithms have good empirical performance.

In addition to the estimation performance of the population parameters, we also present the attribute classification accuracy for different layers of attributes in Table 4. The numbers in this table are calculated as follows: in each simulation replicate, we obtain the posterior modes of each subject's each attribute entry in the shallower-layer $\mathbf{A}^{(1)}$ (similarly for the deeper-layer $\mathbf{A}^{(2)}$), and then average them across the 100 simulation replicates to get the attribute classification accuracy. For all three DeepCDMs and all three $\mathbf{Q}$-matrices structures (a), (b), and (c), the attribute classification accuracy numbers remain reasonably high, basically exceeding 90% for the shallower $\mathbf{A}^{(1)}$ and exceeding 70% for the deeper $\mathbf{A}^{(2)}$. The classification accuracy for deeper attributes is lower than that for shallower ones, which is an inevitable characteristic shared by all higher-order latent variable models widely used in statistics. Despite this, the fact that the deeper attributes still have classification accuracies beyond 70%, and even beyond 90% for DeepLLM, demonstrates that the estimation quality of deeper attributes in our model does not degrade too much and is still acceptable. Furthermore, Table 4 indicates that the DeepLLM has the best performance in classifying the deeper $\mathbf{A}^{(2)}$ and the smallest gap between the classification accuracies of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. This observation suggests that in the considered settings, DeepLLM may be a more preferrable model among the DeepCDM family in terms of estimating the deeper latent attributes.

| Structure | $(J,\ K_1,\ K_2)$ | $N$ | DeepDINA | | Hybrid G-D | | DeepLLM | |
|---|---|---|---|---|---|---|---|---|
| | | | $\mathbf{A}^{(1)}$ | $\mathbf{A}^{(2)}$ | $\mathbf{A}^{(1)}$ | $\mathbf{A}^{(2)}$ | $\mathbf{A}^{(1)}$ | $\mathbf{A}^{(2)}$ |
| (a) in (22) | (30, 6, 2) | 500 | 0.985 | 0.805 | 0.926 | 0.783 | 0.998 | 0.956 |
| | | 1000 | 0.984 | 0.822 | 0.928 | 0.786 | 0.998 | 0.960 |
| | | 2000 | 0.984 | 0.833 | 0.929 | 0.795 | 0.998 | 0.959 |
| (b) in (22) | (30, 7, 3) | 500 | 0.969 | 0.738 | 0.903 | 0.706 | 0.994 | 0.873 |
| | | 1000 | 0.969 | 0.737 | 0.905 | 0.705 | 0.994 | 0.878 |
| | | 2000 | 0.969 | 0.737 | 0.906 | 0.712 | 0.995 | 0.881 |
| (c) in (22) | (30, 8, 3) | 500 | 0.970 | 0.774 | 0.896 | 0.739 | 0.994 | 0.909 |
| | | 1000 | 0.971 | 0.774 | 0.898 | 0.740 | 0.994 | 0.913 |
| | | 2000 | 0.971 | 0.779 | 0.899 | 0.743 | 0.994 | 0.917 |

Table 4: Attribute classification accuracy across all of the simulation settings.

| Structure | $(J,\ K_1,\ K_2)$ | $N$ | RMSE of $\boldsymbol{\pi}^{(1)}$ | | | Computation time (min) | | |
|---|---|---|---|---|---|---|---|---|
| | | | Deep | Saturated | Ratio | Deep | Saturated | Ratio |
| (a) in (22) | (30, 6, 2) | 500 | 0.004 | 0.012 | 34.6% | 1.3 | 5.5 | 24.1% |
| | | 1000 | 0.004 | 0.012 | 29.2% | 2.5 | 10.5 | 24.2% |
| | | 2000 | 0.003 | 0.012 | 22.0% | 5.3 | 21.4 | 24.6% |
| (b) in (23) | (30, 7, 3) | 500 | 0.003 | 0.005 | 49.7% | 1.7 | 10.3 | 16.5% |
| | | 1000 | 0.003 | 0.005 | 54.6% | 3.4 | 19.4 | 17.4% |
| | | 2000 | 0.003 | 0.005 | 59.4% | 6.6 | 37.4 | 17.6% |
| (c) in (24) | (30, 8, 3) | 500 | 0.001 | 0.004 | 26.2% | 2.0 | 20.0 | 10.0% |
| | | 1000 | 0.001 | 0.004 | 24.6% | 4.6 | 48.5 | 9.5% |
| | | 2000 | 0.001 | 0.004 | 21.5% | 8.5 | 76.0 | 11.2% |

Table 5: Comparisons between the two-latent-layer DeepDINA and the saturated DINA model in terms of the RMSE of the proportions $\boldsymbol{\pi}^{(1)}$ of the fine-grained latent attributes $\mathbf{A}^{(1)}$ and the computation time.

**Simulation Study IV: Comparison to the saturated attribute model.** In this simulation study, we generate data using a DeepCDM (DeepDINA here) but estimate parameters using both the DeepCDM and the traditional one-layer CDM (DINA here) with a saturated attribute model. We compare (a) the computation time of the two models, and also (b) their accuracy in recovering the proportions $\boldsymbol{\pi}^{(1)}$ of the latent attributes $\mathbf{A}^{(1)}$. The distribution of $\mathbf{A}^{(1)}$ can be parameterized by $\boldsymbol{\pi}^{(1)} = (\pi^{(1)}_{\boldsymbol{\alpha}}; \boldsymbol{\alpha} \in \{0,1\}^{K_1})$ where $\pi^{(1)}_{\boldsymbol{\alpha}} = \mathbb{P}(\mathbf{A}^{(1)} = \boldsymbol{\alpha})$. Under DINA with a traditional saturated attribute model, $\boldsymbol{\pi}^{(1)}$ are directly treated as parameters and estimated, while in the DeepDINA model, $\boldsymbol{\pi}^{(1)}$ follows another higher-order DINA

model and can be calculated after estimating these higher-order parameters. Here we focus on comparing the accuracy of recovering the distribution of $\mathbf{A}^{(1)}$ via $\boldsymbol{\pi}^{(1)}$ because this is the key difference between the two models. Table 5 displays the average RMSEs of $\boldsymbol{\pi}^{(1)}$ and the average computation time under the two models. In particular, the 6th column "Ratio" in Table 5 displays the ratios of RMSEs under the deep and the saturated model (i.e., ratios of numbers in the 4th and 5th columns in the table), and the 9th column "Ratio" displays the ratio of computation time under the two models (i.e., ratios of numbers in the 7th and 8th columns in the table). Compared to the traditional estimation method for the one-layer DINA model, our DeepDINA method yields 20%–60% of the RMSE in estimating $\boldsymbol{\pi}^{(1)}$ and takes 9%–25% of the computation time. These comparisons imply that appropriately taking into account higher-order discrete structures will lead to both more accurate estimation and more efficient computation. Here, more accurate estimation is thanks to the suitable modeling of the latent attribute dependence, and more efficient computation is thanks to the statistical parsimony and our efficient Gibbs sampling steps of a fewer number of parameters.

**Simulation Study V: Robustness of DeepCDM to deep layer misspecification.** We perform a simulation study to evaluate our method's performance under a misspecified higher-order model. Here we generate data from the HO-CDM in de la Torre and Douglas (2004) that have higher-order continuous latent traits behind the binary latent attributes. Consider structure (c) in (24) with $J = 30$ items, $K_1 = 8$ attributes, and $K_2 = 3$ higher-order continuous latent traits $(\theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}) =: \boldsymbol{\theta}^{(2)}$. Let $\theta_1^{(2)}, \theta_2^{(2)}, \theta_3^{(2)}$ follow independent standard normal distributions. Given $\boldsymbol{\theta}^{(2)}$, the first-layer CDM parameters are set to be the same in the previous DeepLLM simulation setting. Then we fit the data using our Gibbs sampler developed for DeepLLM, and then examine the estimated shallower-layer item parameters $\boldsymbol{\beta}^{(1)}$ under this misspecified model. For better visualization, for a randomly generated dataset, in Figure 2 we plot the heatmap of the estimated $\boldsymbol{\beta}^{(1)}$ in the form of $J \times K_1$ matrix whose sparsity pattern is given by the $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)} \in \{0,1\}^{J \times K_1}$. We can see that the estimated coefficients $\widehat{\boldsymbol{\beta}}^{(1)}$ under a misspecified higher-order model is still close to the ground truth, even for a relatively small sample size $N = 500$. For a larger sample size $N = 2000$, the estimated $\widehat{\boldsymbol{\beta}}^{(1)}$ matrix becomes closer to the truth.

Furthermore, we also look beyond a single simulation trial and carry out 100 independent simulation replicates to assess our method's average performance under model misspecification. Figure 3 presents the boxplots of root mean squared errors (RMSEs) of the estimated shallower-layer $\boldsymbol{\beta}^{(1)}$ parameters based on the 100 replicates. This figure clearly shows a decreasing trend of estimation errors of $\boldsymbol{\beta}^{(1)}$ as sample size increases. Together with the previous Figure 2, we have empirically demonstrated that our DeepCDM methodology has some robustness to model misspecification of the deeper-layers.

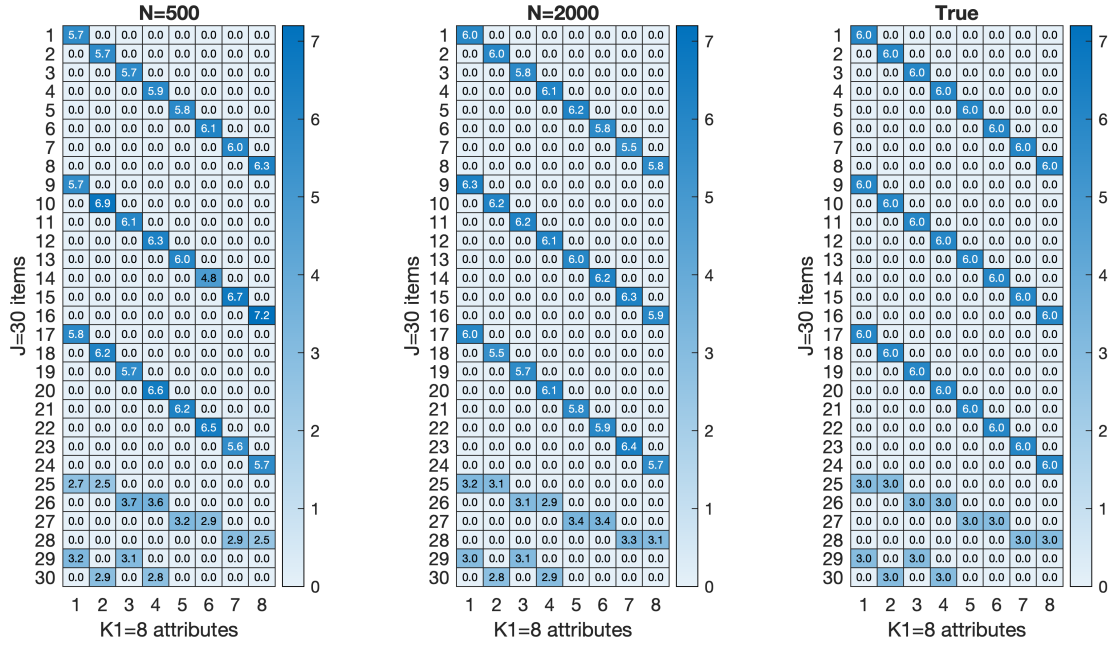



Figure 2: Estimated first-layer parameters $\boldsymbol{\beta}^{(1)}$ under a misspecified latent attribute model. The data are generated from a continuous higher-order latent trait model but estimated using our DeepLLM method.

We next offer more discussions between the connections and differences between the very popular HO-CDM and the proposed DeepCDMs. As described in de la Torre and Douglas (2004), the motivation for proposing the HO-CDM includes parsimony and interpretability. For the HO-CDM, the parsimony comes from using an IRT model with continuous latent traits to model the binary attributes, and the interpretability comes from defining a plausible model for the relationship between general ability and specific knowledge. On one hand, as mentioned in Section 1, DeepCDMs also similarly have the advantages of parsimony

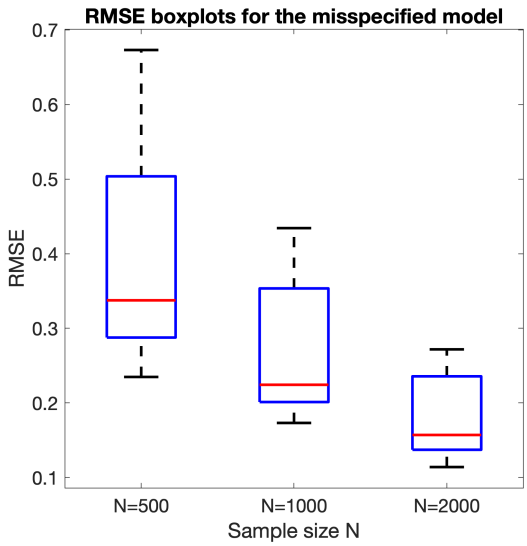



Figure 3: RMSE boxplots for the estimated first-layer parameters $\boldsymbol{\beta}^{(1)}$ under a misspecified latent attribute model. Results are based on 100 independent simulation replications for each sample size.

and interpretability. On the other hand, there are also several key differences between the HO-CDM and DeepCDMs. *First*, DeepCDMs use fully discrete latent layers, which offer a different interpretation of multi-granularity skill diagnosis. *Second*, the above simulation study implies that a special member in the DeepCDM family – DeepLLM – can serve as an approximation to HO-CDM; our DeepLLM method can robustly estimate the item parameters for data generated from HO-CDM. It is then worth emphasizing that DeepLLM is just a special member of the DeepCDM family, and that other members in this family can flexibly model structures well beyond the logistic linear form used in DeepLLM and HO-CDM. For example, DeepDINA or Hybrid GDINA-DINA can model the nonlinear conjunctive relationship or interaction effects of higher-order discrete attributes, and they are still identifiable and easy to estimate via Gibbs sampling (see Section 4). However, there currently do not exist extensions of HO-CDM to nonlinear higher-order latent variable settings.

# 6 Application to the TIMSS Assessment Data

We demonstrate the DeepCDM methodology by applying it to data extracted from the TIMSS 2019 math assessment mentioned in Section 1; the data are accessed from the TIMSS

2019 International Database (Fishbein et al., 2021). We use two-latent-layer DeepCDMs to analyze the US student response data to item block No.2 in the eighth grade math assessment. Prior to our analysis, the original student response data are converted into binary correct/wrong responses as follows, based on the *TIMSS 2019 Item Information* available in the online database (Fishbein et al., 2021). For multiple-choice items, a student response is coded as one if the response matches the correct answer key, and coded as zero otherwise; for constructed response items, a student response is coded as one if the number of scores received is equal to the maximal score of the item, and coded as zero otherwise.

Among the US eighth grade participants, we consider students that took the math item block No.2 and give responses to all the $J = 28$ items in this block. This results in a binary observed data matrix containing responses from $N = 972$ students. The online *TIMSS 2019 Item Information - Grade 8* provides details about which specific skills each test item is measuring, and we use these information to construct the $\mathbf{Q}$-matrices. There are four *content* skills: $\alpha_1^{(1)}$: Number; $\alpha_2^{(1)}$: Algebra; $\alpha_3^{(1)}$: Geometry; and $\alpha_4^{(1)}$: Data and Probability; and three *cognitive* skills: $\alpha_5^{(1)}$: Knowing; $\alpha_6^{(1)}$: Applying; and $\alpha_7^{(1)}$: Reasoning. These content and cognitive skills can be viewed as subcompetences for which it is desirable to provide fine-grained diagnoses. Therefore, we model these seven skills as $K_1 = 7$ fine-grained attributes in the shallower latent layer in a DeepCDM. In fact, each test item is listed as measuring one content skill and one cognitive skill; for example, the first item in block No.2 measures $\alpha_1^{(1)}$: Number, and $\alpha_5^{(1)}$: Knowing. We use such available item information to obtain the first-layer $J \times K_1$ $\mathbf{Q}$-matrix $\mathbf{Q}_{28\times 7}^{(1)}$ in Table 6. Further, as already implied by the above skill descriptions, the seven specific skills naturally belong to two general domains: the content domain and the cognitive domain. Here, the wordings of naming "content" and "cognitive" as two "domains" are official terms defined by and provided in the online TIMSS 2019 Assessment Frameworks. Diagnosing a student's states on these latent domains can reflect their general strengths/weaknesses on these two broad aspects. So the deeper latent layer in our DeepCDM has two *domain attributes*: $\alpha_1^{(2)}$: Content and $\alpha_2^{(2)}$: Cognitive. According to the equivalence between the direct dependencies among variables and the $\mathbf{Q}$-matrix entries, we can use the above attribute information to construct a $K_1 \times K_2$ matrix $\mathbf{Q}_{7\times 2}^{(2)} = (q_{k,m}^{(2)})$, shown in Table 7.

| Item ID | $\alpha_1^{(2)}$ Number | $\alpha_2^{(2)}$ Algebra | $\alpha_3^{(2)}$ Geometry | $\alpha_4^{(2)}$ Data Prob. | $\alpha_5^{(2)}$ Knowing | $\alpha_6^{(2)}$ Applying | $\alpha_7^{(2)}$ Reasoning |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 7 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 8 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 9 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 10 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 11 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 13 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 14 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 15 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 16 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 17 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 18 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 20 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 21 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 22 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 23 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 24 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 25 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 26 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 27 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 28 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |

Table 6: First-layer $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)}_{28\times 7}$ for item block No.2 in TIMSS 2019 eighth grade math assessment.

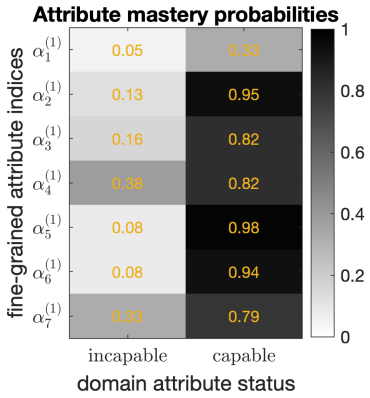
| | | $\alpha_1^{(1)}$ Content Domain | $\alpha_2^{(1)}$ Cognitive Domain |
|---|---|---|---|
| $\alpha_1^{(2)}$ | Number | 1 | 0 |
| $\alpha_2^{(2)}$ | Algebra | 1 | 0 |
| $\alpha_3^{(2)}$ | Geometry | 1 | 0 |
| $\alpha_4^{(2)}$ | Data and Probability | 1 | 0 |
| $\alpha_5^{(2)}$ | Knowing | 0 | 1 |
| $\alpha_6^{(2)}$ | Applying | 0 | 1 |
| $\alpha_7^{(2)}$ | Reasoning | 0 | 1 |

Table 7: Second-layer $\mathbf{Q}$-matrix $\mathbf{Q}^{(2)}_{7\times 2}$ for TIMSS 2019 eighth grade math assessment.

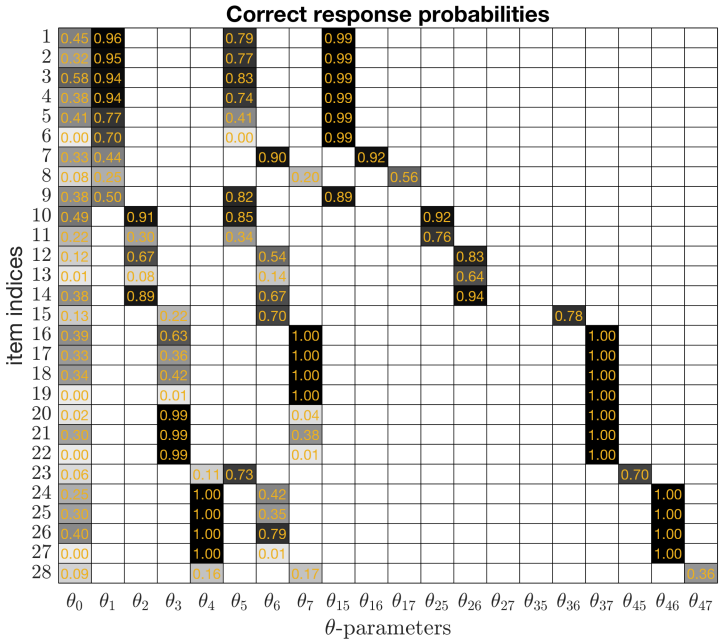
We then apply our Bayesian estimation method to the TIMSS data. DeepDINA is not used here because $\mathbf{Q}^{(1)}$ does not satisfy the C-R-D conditions (i.e., does not contain an identity submatrix $\mathbf{I}_{K_1}$), and hence does not give an identifiable DeepDINA model. As for DeepLLM and Hybrid GDINA-DINA (abbreviated as Hybrid G-D hereafter), it is not difficult to verify that $\mathbf{Q}^{(1)}_{28\times 7}$ in Table 6 satisfies the generic identifiability conditions (G1) and (G2) in Theorem 3 for main-effect-based models, and that $\mathbf{Q}^{(2)}_{7\times 2}$ in Table 7 satisfies the strict identifiability condition (S) in Theorem 2 for general diagnostic models. This means all the parameters in DeepLLM and Hybrid G-D are all strictly or generically identifiable. Note that $\mathbf{Q}^{(2)}$ has all the rows each being either $(1,0)$ or $(0,1)$, in which case the Hybrid G-D model in fact covers both DeepDINA and DeepLLM as special cases and offers a more general alternative. Therefore we focus on the more general Hybrid G-D model next.

We run the Gibbs sampler for Hybrid G-D for 15,000 iterations and retain the last 5000 as our posterior samples, the same as in the simulation studies. Based on these samples, the posterior means are calculated for all the continuous parameters in the model. The deep proportion parameters' posterior means are $\overline{\boldsymbol{\pi}}^{\text{deep}} = (0.477,\ 0.033,\ 0.059,\ 0.430)$, which correspond to deep latent patterns $\mathbf{A}^{(2)} = (0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, respectively. This estimated $\overline{\boldsymbol{\pi}}^{\text{deep}}$ implies that the two domain attributes exhibit a relatively high correlation. As for the quasi-item parameters characterizing $\mathbb{P}(A^{(1)}_k \mid \mathbf{A}^{(2)}, \mathbf{Q}^{(2)})$ and item parameters characterizing $\mathbb{P}(R^{(1)}_j \mid \mathbf{A}^{(1)}, \mathbf{Q}^{(1)})$, we plot their posterior means in Figure 4. Specifically, Figure 4(a) shows the conditional attribute mastery probabilities given the domain attributes, with its left column showing the quasi-guessing parameters $\boldsymbol{g}^{(2)} = (g^{(2)}_1, \ldots, g^{(2)}_7)^\top$, and right column showing one minus the quasi-slipping parameters $\mathbf{1}_{7\times 1} - \boldsymbol{s}^{(2)} = (1-s^{(2)}_1, \ldots, 1-s^{(2)}_7)^\top$. Figure 4(b) shows the conditional correct response probabilities given the fine-grained attributes, that is, the $\theta$-parameters in (18). For each item $j$, the column $\theta_0$ refers to $\theta^{(1)}_{j,\varnothing}$; column $\theta_k$ refers to $\theta^{(1)}_{j,\{k\}}$ for $k = 1, \ldots, 7$; column $\theta_{15}$ refers to the $\theta^{(1)}_{j,\{1,5\}}$, etc. For a item $j \in \{1, \ldots, 28\}$, only those "effective" $\theta$-parameters are ploted in Figure 4. For example, the first item requires the first and the fifth attributes (i.e., Number and Knowing), so only four $\theta$-parameters are "effective" and shown in the first line in Figure 4(b): $\theta_0$, $\theta_1$, $\theta_5$, and $\theta_{15}$.

To further inspect the latent attributes' mutual dependence, we calculate the element-wise posterior modes of the discrete latent profiles and obtain the $N \times K_1$ binary matrix




(a) based on $\boldsymbol{\theta}^{(2)}_{\text{DINA}}$




(b) based on $\boldsymbol{\theta}^{(1)}_{\text{GDINA}}$

Figure 4: TIMSS 2019 eighth-grade math assessment US data, item block No.2, estimated parameters from the Hybrid-GDINA-DINA model. Plot (a): deeper DINA-layer parameters, with the left column being $\boldsymbol{g}^{(2)}$ and the right column being $\mathbf{1}_{7\times 1} - \boldsymbol{s}^{(2)}$; plot (b): conditional correct response probabilities under GDINA.

$\bar{\mathbf{A}}^{(1)} = (\bar{a}^{(1)}_{i,k})$ and the $N \times K_2$ binary matrix $\bar{\mathbf{A}}^{(2)} = (\bar{a}^{(1)}_{i,m})$. Specifically, each binary entry $\bar{a}^{(1)}_{i,k}$ is the posterior mode of $a^{(1)}_{i,k}$ based on the retained posterior samples, and $\bar{a}^{(2)}_{i,m}$ is similarly obtained. Based on the $K_1 = 7$ columns of $\bar{\mathbf{A}}^{(1)}$ and $K_2 = 2$ columns of $\bar{\mathbf{A}}^{(2)}$, we generate the scatterplot matrices in Figure 5. In this figure, the two plots on the left show the correlation between the second-layer domain attributes (Figure 5(a)) and those between pairs of the first-layer fine-grained attributes (Figure 5(c)). The two plots on the right panel of Figure 5 show the jittered versions of the scatterplot matrices, which more explicitly visualize the pairwise joint distributions of latent variables. As expected, the seven fine-grained latent skills show relatively high positive dependencies on each other, which supports using the DeepCDM modeling framework. Moreover, the estimated posterior mode matrices $\bar{\mathbf{A}}^{(1)}$ and $\bar{\mathbf{A}}^{(2)}$ provide multi-granularity diagnoses of students' strengths/weaknesses on both the two broader domain attributes and the seven more fine-grained attributes.

Next, we also perform a comparative analysis of a TIMSS 2019 fourth-grade math assess-

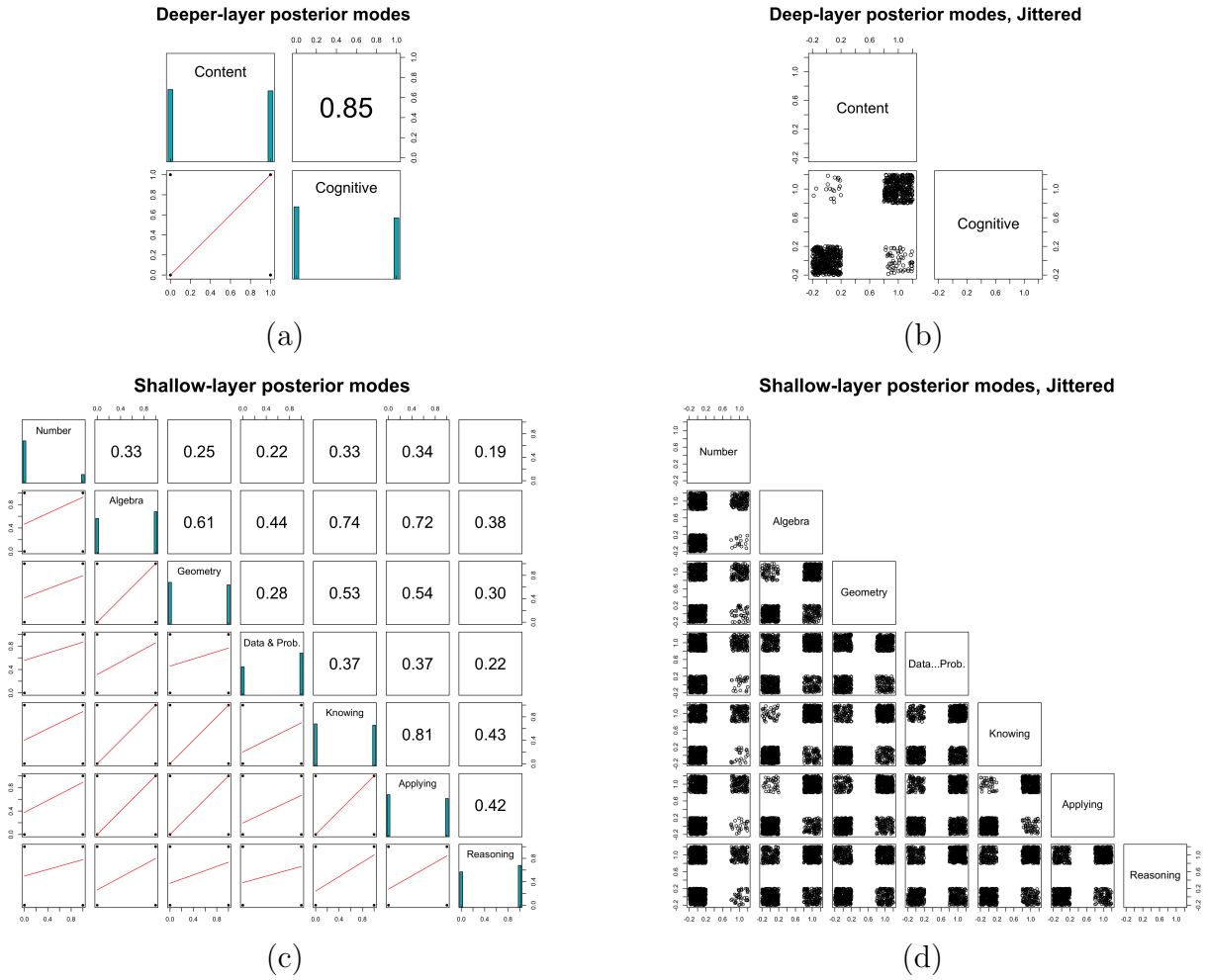
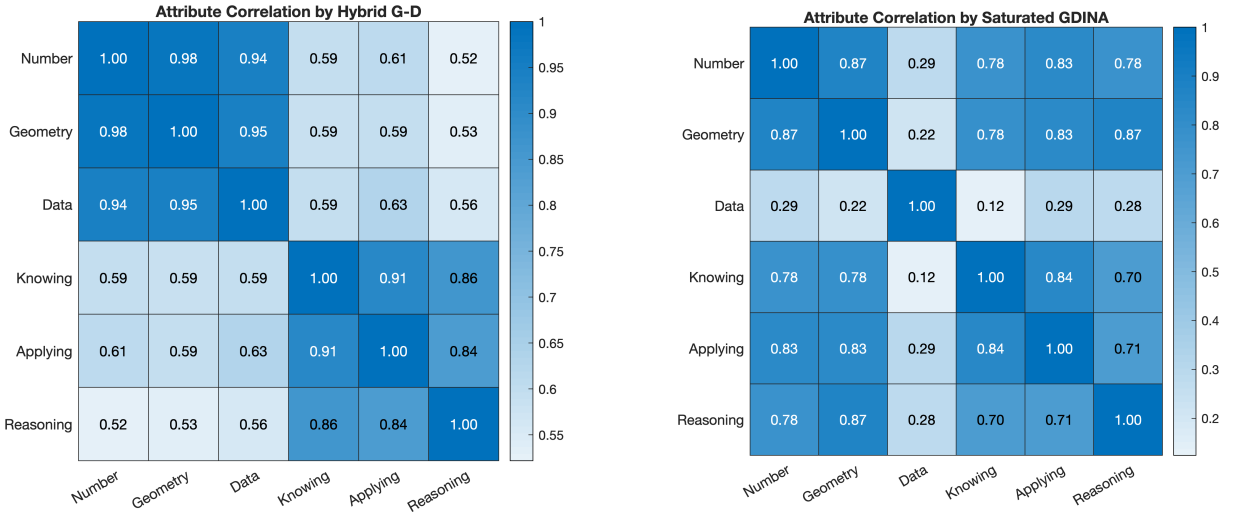



Figure 5: TIMSS 2019 eighth-grade math assessment US data, item block No.2, estimated latent profiles. In plots (b) and (d), the sample data points are jittered from zero/one.

ment dataset (item block No. 7) using both a DeepCDM and a traditional CDM to see their difference. Specifically, we consider both the Hybrid G-D model (which is GDINA with a higher-order DINA layer), and GDINA with a saturated latent attribute model. In terms of *statistical parsimony*, our Hybrid G-D requires much fewer parameters than GDINA with a saturated latent layer. In particular, to model $K_1 = 6$ fine-grained latent attributes, the Hybrid G-D model uses only $3+6\times 2 = 15$ parameters while the traditional saturated attribute model uses a large number of $2^6 - 1 = 63$ parameters. Such statistical parsimony implies that our DeepCDM would require a smaller sample size to reach the same level of parameter estimation precision. In terms of *substantive interpretations*, the correlation plots in Figure 6 show that the Hybrid G-D model gives a much more interpretable correlation structure

among the fine-grained latent attributes (in the left panel) than GDINA with a saturated attribute model (in the right panel). Specifically, recall that the first three attributes fall in the "Content" domain and the last three attributes fall in the "Cognitive" domain. The nearly block diagonal heatmap in Figure 6(a) shows that our DeepCDM induces much higher correlations among attributes within a same domain than those across two different domains. On the other hand, for the GDINA model with a saturated attribute model, Figure 6(b) shows a somewhat counter-intuitive pattern: "Data" has a relatively small correlation with all other attributes and there are no clear separation between the content-related attributes and the cognitive-related ones.




(a) Hybrid GDINA-DINA model.

(b) GDINA with a saturated attribute model.

Figure 6: Estimated attribute correlation plots given by the proposed Hybrid GDINA-DINA model (i.e., GDINA model with a higher-order DINA layer) in (a) and GDINA with a saturated attribute model in (b) for the TIMSS 2019 4th grade math booklet 7 dataset.

# 7 Discussion

In this work, we have proposed a new family of interpretable diagnostic models called DeepCDMs, established transparent identifiability conditions and general identifiability theory, and developed Bayesian estimation methods for them. On one hand, DeepCDMs are well motivated by the applied goal of uncovering rich and structured diagnostic information from

educational and behavioral data. Through the estimated multilayer latent profiles, DeepCDMs enable multi-granularity diagnoses of latent attributes from coarse to fine-grained and from high-level to detailed. On the other hand, in terms of discrete latent structures, DeepCDMs share similarities with powerful deep learning models such as deep belief networks (Hinton et al., 2006) and deep Boltzmann machines (Salakhutdinov and Larochelle, 2010), and are expressive modeling tools. Distinctively, DeepCDMs are fully identifiable under our conditions, which is a desirable property lacked by most deep learning models. In a nutshell, our identifiability conditions can be summarized as: as long as each $\mathbf{Q}^{(d)}$ satisfies the identifiability condition under the CDM to which the shallower layer $\mathbf{A}^{(d-1)}$ (or $\mathbf{R}$ if $d = 1$) conforms, then the entire DeepCDM is identifiable. Our identifiability guarantees form the very foundation for deriving interpretable and reliable insights in practical applications, and offer the very guidelines on adopting a shrinking-ladder-shaped generative graph structure. Simulation results empirically corroborate the identifiability conclusions, and also demonstrate the good practical performance of our Bayesian estimation algorithms.

In our real data example in Section 6 and other potential future applications, the deeper-layer binary variables are not used in order to capture the person's *continuous variability* in the coarse-grained higher-order skills as in the HO-CDM in de la Torre and Douglas (2004). Instead, the higher-order meta attributes provide an additional layer of *discrete diagnoses* of the persons' higher-order skills. Such a diagnostic modeling goal shares a similar motivation with originally using CDMs as an alternative modeling tool to the classical (multidimensional) IRT models with continuous latent traits. Historically, IRT has been the dominating modeling methodology in educational and psychological measurement, thanks to their excellent ability of capturing subjects' latent variability. Nonetheless, in the recent two decades, CDMs have also emerged as powerful alternative tools that provide fine-grained discrete diagnoses of skills, instead of capturing the continuous variability. In this sense, we view the proposed DeepCDMs as going further down the road of diagnostic classification, by providing skill diagnoses with multiple layers of granularity. To fully realize the applied potential of the proposed new framework, our far-reaching goal is for practitioners to design new cognitive diagnostic assessments directly inspired by the DeepCDM identifiability theory.

DeepCDMs suppose that the latent variables follow a multilayer generative structure.

In practice, admittedly, it may not always be the case that attributes follow multiple neat layers as in a DeepCDM. On the other hand, however, we believe that in a number of CDM modeling and application scenarios, the advantages of DeepCDMs in terms of statistical parsimony, practical interpretability, and identifiability outweigh the induced limitation. Our motivation for proposing DeepCDMs is not to replace, but to complement, other latent structural models (including attribute hierarchy methods, higher-order continuous latent trait models) in the CDM literature as an alternative family of interpretable and identifiable models. Specifically, we expect DeepCDMs will be suitable for those applications where multi-resolution discrete diagnoses of latent attributes are of interest. We hope this work contributes a useful first step towards a versatile toolbox of providing statistically justified multi-granularity diagnostic classification.

The proposed DeepCDM framework unlocks many interesting future research possibilities. First, this paper has focused on binary responses and binary latent variables in all the layers, but the DeepCDM framework can be readily extended to polytomous responses and polytomous attributes (Chen and de la Torre, 2013; Gao et al., 2021). Similar identifiability conditions on the between-layer $\mathbf{Q}$-matrices may be obtained, and corresponding Bayesian estimation methods can also be developed. To this end, the Bayesian Pyramid model and its corresponding Bayesian estimation method in Gu and Dunson (2023) is an example, which deals with multivariate unordered categorical data with binary latent layers. Second, this paper develops Markov Chain Monte Carlo algorithms for estimation. In the future, it would also be useful to develop more scalable variational Bayesian inference algorithms or EM algorithms for DeepCDMs to enhance computational efficiency.

Another interesting future direction is to perform *exploratory* DeepCDM analysis and estimate the $\mathbf{Q}$-matrices from data. This initial work has focused on *confirmatory* scenarios in which multi-granularity design information are available and can be directly translated into the $\mathbf{Q}$-matrices. Nevertheless, all of our identifiability results are fully general and applicable to the exploratory settings with unknown $\mathbf{Q}$-matrices. This means we have also obtained identifiability guarantees for directly estimating all the $\mathbf{Q}$-matrices in a DeepCDM. In recent years, there has been an increasing interest in exploratory estimation of CDMs, including those using Bayesian approaches (Culpepper, 2019b; Chen et al., 2020; Balamuta

and Culpepper, 2022) and those using frequentist ones (Chen et al., 2015; Xu and Shang, 2018; Gu and Xu, 2023). Developing efficient methods to estimate the multiple $\mathbf{Q}$-matrices in a DeepCDM is important future work. Furthermore, in an even more exploratory setting, it would also be interesting to study how to select the number of latent variables $K_1$, $K_2$, etc. in each layer in a DeepCDM. Nonparametric Bayesian approaches can be useful tools toward this end (e.g., Fang et al., 2019; Chen et al., 2021; Gu and Dunson, 2023).

On the application front, for modern large-scale educational assessments such as TIMSS and PISA, we believe there is a promising future potential of using the DeepCDM methodology to model and analyze high-dimensional response data, to generate new insights into student achievement, and to enhance multi-granularity instruction and intervention. Indeed, the TIMSS 2019 eighth grade math assessment offers more levels of item information than are used in our current data analysis. For example, under the "Number" skill, there are still four different topic areas: *Integers* / *Fractions and decimals* / *Ratio, proportion, and percent*, which are candidates for more fine-grained attributes. In the future, advancing and refining the computational techniques for DeepCDMs with more layers can help extract even more nuanced diagnoses about student subcompetences from large-scale assessment data.

On a final note, we would like to give a broader discussion on DeepCDMs' implications. In applied cognitive psychology, the concept of "higher order thinking skills" was put forward (Brookhart, 2010; Schraw and Robinson, 2011) which includes problem solving, critical thinking, creativity, and so on; in linguistics, the "ladder of abstraction" idea was proposed (Hayakawa, 1947; Munson et al., 2011) to describe the way humans think and communicate in varying degrees of abstraction through languages; and in deep learning, an influential review article Bengio et al. (2013) pointed out that using deep architectures can potentially lead to progressively more abstract features at higher layers of representations. Our shrinking-ladder-shaped DeepCDMs attempt to offer principled and identifiable statistical models to back up such substantive theory and deep learning heuristics. We hope the DeepCDM framework will be useful for practitioners, illuminating for theoreticians, and triggering fruitful future research on using rigorous statistical methods to cross-fertilize the fields of (deep) machine learning and psychometrics.

**Supplementary Material.** The Supplementary Material contains the proofs of the identifiability theorems and the details of the Gibbs sampling algorithms for posterior computation.

**Acknowledgements.** This work is partially supported by NSF Grant DMS-2210796. The author thanks the editor Prof. Matthias von Davier and two anonymous reviewers for their many helpful and constructive comments that helped improve this paper's quality.

# 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.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.

Balamuta, J. J. and Culpepper, S. A. (2022). Exploratory restricted latent class models with monotonicity requirements under PÒLYA–GAMMA data augmentation. *Psychometrika*, pages 1–43.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

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, J. and de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the Programme for International Student Aassessment in reading. *Psychology*, 5(18):1967.

Chen, Y., Culpepper, S. A., Chen, Y., and Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika*, 83(1):89–108.

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.

Chen, Y., Liu, Y., Culpepper, S. A., and Chen, Y. (2021). Inferring the number of attributes for the exploratory DINA model. *Psychometrika*, 86(1):30–64.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40(5):454–476.

Culpepper, S. A. (2019a). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, 84(2):333–357.

Culpepper, S. A. (2019b). An exploratory diagnostic model for ordinal responses with binary attributes: identifiability and estimation. *Psychometrika*, 84(4):921–940.

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.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment*, 361389.

Fang, G., Liu, J., and Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, 84(1):19–40.

Fishbein, B., Foy, P., and Yin, L. (2021). TIMSS 2019 User Guide for the International Database (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-database/.

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.

George, A. C. and Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology*, 11(3):189–205.

Gierl, M. J., Leighton, J. P., and Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills. *Cognitive diagnostic assessment for education: Theory and applications, Cambridge, UK: Cambridge University Press*, pages 242 – 274.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Gu, Y. and Dunson, D. B. (2023). Bayesian pyramids: identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426.

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.

Gu, Y. and Xu, G. (2022). Identifiability of hierarchical latent attribute models. *Statistica Sinica*, to appear.

Gu, Y. and Xu, G. (2023). A joint MLE approach to large-scale structured latent attribute analysis. *Journal of the American Statistical Association*, 118(541):746–760.

Hayakawa, S. I. (1947). *Language in action*. Harcourt, Brace and company.

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.

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.

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

Liu, C.-W., Andersson, B., and Skrondal, A. (2020). A constrained Metropolis–Hastings Robbins–Monro algorithm for Q matrix estimation in DINA models. *Psychometrika*, 85(2):322–357.

Ma, W. and de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93:1–26.

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

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Mourad, R., Sinoquet, C., Zhang, N. L., Liu, T., and Leray, P. (2013). A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47:157–203.

Munson, B., Edwards, J., Beckman, M. E., Cohn, A. C., Fougeron, C., and Huffman, M. K. (2011). Phonological representations in language acquisition: Climbing the ladder of abstraction. *The Oxford Handbook of Laboratory Phonology*, pages 288–309.

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

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

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

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

Salakhutdinov, R. and Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700. JMLR Workshop and Conference Proceedings.

Schmid, J. and Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61.

Schraw, G. and Robinson, D. H. (2011). *Assessment of higher order thinking skills*. IAP.

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

Templin, J. and Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2):317–339.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3):287.

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.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523):1284–1295.

Xu, G. and Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3):625–649.

Xu, X. and von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series*, 2008(1):i–18.

Yung, Y.-F., Thissen, D., and McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, 64:113–128.

# Supplement to "Diving Deep in Diagnostic Modeling: DeepCDMs"

In this Supplementary Material, Section S.1 presents the proofs of the identifiability results of DeepCDMs, and Section S.2 provides the posterior computation details of the Gibbs sampling algorithms for DeepCDMs.

## S.1 Proofs of the Identifiability Results

All of our identifiability proofs leverage a key technical insight about DeepCDMs – that is, identifiability can be examined and established in a layer-by-layer manner, from the bottom up, thanks to the probabilistic formulation of the directed graphical model. This insight was initially used in **?** to establish identifiability of the deep Bayesian Pyramid model for multivariate categorical data.

*Proof of Theorem 1.* Recall the joint distribution of all the random variables in a DeepCDM (including a DeepDINA model and a Hybrid DeepCDM) is

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

The marginal distribution of the observed vector $\mathbf{R}$ is obtained by marginalizing out all the latent variables $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(D)}$ in the above joint distribution. According to the definition of a general directed acyclic graph (DAG), the marginal distribution of each latent vector $\mathbf{A}^{(d)}$ for layer $d = 1, \ldots, D-1$ can be written as

$$\begin{aligned} &\mathbb{P}(\mathbf{A}^{(d)} = \boldsymbol{\alpha}^{(d)}) \qquad\qquad (S.1) \\ &= \sum_{\boldsymbol{\alpha}^{(d+1)} \in \{0,1\}^{K_{d+1}}} \cdots \sum_{\boldsymbol{\alpha}^{(D)} \in \{0,1\}^{K_D}} \prod_{m=d+1}^{D} \mathbb{P}(\mathbf{A}^{(m-1)} = \boldsymbol{\alpha}^{(m-1)} \mid \mathbf{A}^{(m)} = \boldsymbol{\alpha}^{(m)}) \cdot \mathbb{P}(\mathbf{A}^{(D)} = \boldsymbol{\alpha}^{(D)}). \end{aligned}$$

Now we specifically marginalize out all latent variables except the shallowest layer $\mathbf{A}^{(1)}$ in the joint distribution,

$$
\begin{aligned}
&\mathbb{P}(\mathbf{R}=\boldsymbol{r}) \\
=& \sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}\in\{0,1\}^{K_D}} \mathbb{P}(\mathbf{R}=\boldsymbol{r}, \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}, \ldots, \mathbf{A}^{(D)}=\boldsymbol{\alpha}^{(D)}) \\
=& \sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}) \times \\
& \underbrace{\sum_{\boldsymbol{\alpha}^{(2)}\in\{0,1\}^{K_2}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}\in\{0,1\}^{K_D}} \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)}=\boldsymbol{\alpha}^{(d-1)} \mid \mathbf{A}^{(d)}=\boldsymbol{\alpha}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}=\boldsymbol{\alpha}^{(D)})}_{\mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)})} \\
=& \sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}) \cdot \mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}),
\end{aligned} \tag{S.2}
$$

We introduce a notation $\boldsymbol{\pi}^{(1)} = \left(\pi^{(1)}_{\boldsymbol{\alpha}};\ \boldsymbol{\alpha}\in\{0,1\}^{K_1}\right)$ to collect the proportion parameters of the categorical distribution that $\mathbf{A}^{(1)}$ follows in (S.2):

$$
\mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}) = \pi^{(1)}_{\boldsymbol{\alpha}}, \quad \forall \boldsymbol{\alpha}\in\{0,1\}^{K_1}. \tag{S.3}
$$

Then $\boldsymbol{\pi}^{(1)}$ lives in the $(2^{K_1}-1)$-dimensional probability simplex. Then based solely on $\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}$, the probability mass function of the random vector $\mathbf{R}$ can be written as follows for each $\boldsymbol{r}\in\{0,1\}^J$,

$$
\mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \boldsymbol{\pi}^{(1)}, \boldsymbol{\theta}^{(1)}, \mathbf{Q}^{(1)}) = \sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \pi^{(1)}_{\boldsymbol{\alpha}^{(1)}} \prod_{j=1}^{J} \mathbb{P}(R_j=r_j \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}, \boldsymbol{\theta}^{(1)}, \mathbf{Q}^{(1)}), \tag{S.4}
$$

where the notation $\boldsymbol{\theta}^{(1)}$ collects all the continuous parameters needed to specify the conditional distribution of $\mathbf{R} \mid \mathbf{A}^{(1)}$ under $\mathbf{Q}^{(1)}$. For example, under the DeepDINA model, $\boldsymbol{\theta}^{(1)}$ denotes the collection of $\boldsymbol{s}^{(1)}$ and $\boldsymbol{g}^{(1)}$. Note that (S.4) gives a restricted latent class model (equivalently, a CDM) for $\mathbf{R}$ with $2^{K_1}$ latent classes, subject to the constraints induced by the $J\times K_1$ $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)}$. Similarly, according to the general marginal distribution of $\mathbf{A}^{(d)}$

in (S.1), we also have

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

which is another cognitive diagnostic model for the "response vector" being $\mathbf{A}^{(d)}$ and the "latent attribute vector" being $\mathbf{A}^{(d+1)}$ under the $\mathbf{Q}$-matrix $\mathbf{Q}^{(d+1)}$, where $d = 2, \ldots, D$.

Now consider the DeepDINA model setting in Theorem 1. When $\mathbf{R} \mid \mathbf{A}^{(1)}$ follows the DINA model, then as long as $\mathbf{Q}^{(1)}$ satisfies the C-R-D conditions in Gu and Xu (2021), then $\mathbf{Q}^{(1)}$ itself and the continuous parameters $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\pi}^{(1)}$ are identifiable. Note that the statement that $\boldsymbol{\pi}^{(1)}$ is identifiable means the marginal distribution of $\mathbf{A}^{(1)}$ is identifiable, which implies $\mathbf{A}^{(1)}$ can be treated as if it is observed when studying the identifiability of $\mathbf{Q}^{(2)}$, $\boldsymbol{\theta}^{(2)}$, and the marginal distribution of $\mathbf{A}^{(2)}$. Therefore, if $\mathbf{Q}^{(2)}$ also satisfies the C-R-D conditions, then $\mathbf{Q}^{(2)}$, $\boldsymbol{\theta}^{(2)}$, and the marginal distribution of $\mathbf{A}^{(2)}$ are identifiable. Now it is easy to see that we can proceed in a layerwise manner from bottom up, and examining whether $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, …, $\mathbf{Q}^{(D)}$ satisfy the identifiability conditions successively. Specifically, under a DeepDINA model, as long as all the $\mathbf{Q}^{(d)}$ satisfy the C-R-D conditions, then all the $\mathbf{Q}$-matrices and all the continuous parameters $(\boldsymbol{s}^{(d)}, \boldsymbol{g}^{(d)})$, $d = 1, \ldots, D$ and $\boldsymbol{\pi}^{\text{deep}}$ are strictly identifiable. This proves the sufficiency part in Theorem 1.

To show the necessity part in Theorem 1, we only need to note that if $\mathbf{Q}^{(d)}$ fails to satisfy the C-R-D conditions, then certain parameters in $\boldsymbol{\pi}^{(d)}$ and $\boldsymbol{\theta}^{(d)}$ will not identifiable, indicating the non-identifiability of the DeepDINA model. This proves the necessity of the proposed identifiability conditions and completes the proof of Theorem 1. □

*Proof of Theorem 2 and Proposition 1.* We use the same insight elaborated in the proof of Theorem 1: the layerwise proof argument of identifiability. Specifically, the marginal distribution of $\mathbf{R}$ in (S.2), the marginal distribution of $\mathbf{A}^{(1)}$ in (S.3), and the conditional distribution of $\mathbf{R}$ given $\mathbf{A}^{(1)}$ in (S.4) all hold generally for an arbitrary DeepCDM and a Hybrid DeepCDM. Therefore, we still start with the bottom two layers and examine whether $\mathbf{Q}^{(1)}$ satisfies the identifiability conditions for a general CDM; if so, we then examine $\mathbf{Q}^{(2)}$, so on and so forth. First, we consider the case that condition (S) holds; that is, 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. In this case, following a similar argument as the proof of Theorem 4 in **?** but constraining to considering binary responses, we obtain the strict identifiability of $(\boldsymbol{\theta}^{(d)}, \mathbf{Q}^{(d)})$ for $d = 1, \ldots, D$ and that of $\boldsymbol{\pi}^{\text{deep}}$. Second, we consider the case that condition (S$^*$) holds, then following a similar argument as the proof of Theorem 1 in Culpepper (2019b) but constraining to considering binary responses, we also obtain the strict identifiability of all the parameters and $\mathbf{Q}$-matrices in a general DeepCDM. This proves Theorem 2.

Further note that the above layerwise proof strategy does not require each layer in a DeepCDM to conform to the same diagnostic model. This means in a Hybrid CDM where some layers follow the DINA (or DINO) model and some layers follow the main-effect or all-effect diagnostic models, we can examine their corresponding $\mathbf{Q}$-matrices using the respective identifiability conditions in Theorems 1 or 2 to assess identifiability. For example, if the marginal distribution of $\mathbf{A}^{(d)}$ is already identified, then $\mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)}$ follows the DINA model, then $\mathbf{Q}^{(d+1)}$ only needs to satisfy the weaker C-R-D conditions to proceed to the deeper layer. This proves Proposition 1. □

*Proof of Theorem 3.* Similarly as the proofs of strict identifiability results, we still use the layerwise identifiability argument. In the literature, Theorem 4 in Gu and Xu (2021) established generic identifiability for single-latent-layer main-effect/all-effect CDMs (also see Gu and Xu (2020) and Chen et al. (2020)) under the considered conditions (G1) and (G2) in its single-layer form ($D = 1$); in that theorem, the Lebesgue measure-zero subset of the parameter space where identifiability may break down only concerns the item parameters. That means, in the context of a DeepCDM consisting of main-effect or all-effect layers, as long as the item parameters $\boldsymbol{\theta}^{(1)} \in \Omega_{\text{main}}(\boldsymbol{\beta}^{(1)}; \mathbf{Q}^{(1)})$ do not fall within the layer-specific unidentifiable subset $\mathcal{N}^{(1)}$ which has measure zero in $\Omega_{\text{main}}(\boldsymbol{\beta}^{(1)}; \mathbf{Q}^{(1)})$, then $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\pi}^{(1)}$, and $\mathbf{Q}^{(1)}$ are identifiable. This implies that as long as the between-layer continuous parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(D)}$ do not fall within the finite union of the measure-zero subsets of the parameter space $\cup_{d=1}^{D} \Omega_{\text{main}}(\boldsymbol{\beta}^{(d)}; \mathbf{Q}^{(d)})$, then the entire main-effect or all-effect DeepCDM is identifiable. This proves the generic identifiability conclusion in Theorem 3 under conditions (G1) and (G2). □

## S.2 Details for the Gibbs Sampling Algorithms

### S.2.1 Gibbs Sampler for Two-latent-layer DeepDINA

For $i \in [N]$, $j \in [J]$, and $k \in [K_1]$, introduce binary ideal response indicators $\xi_{1,ij}$ and $\xi_{2,ik}$:

$$\xi_{1,ij} = \prod_{k=1}^{K_1} \left(a_{i,k}^{(1)}\right)^{q_{j,k}^{(1)}}, \quad \xi_{2,ik} = \prod_{m=1}^{K_2} \left(a_{i,m}^{(2)}\right)^{q_{k,m}^{(2)}}. \tag{S.5}$$

Denote $s_j^{(1)}$, $g_j^{(1)}$, $s_k^{(2)}$, and $g_k^{(2)}$ by $s_{1,j}$, $g_{1,j}$, $s_{2,k}$, and $g_{2,k}$, respectively. Under the priors specified in the main text, the posterior distribution in the two-latent-layer DeepDINA can be written as

$$\begin{aligned}
&p(\boldsymbol{\theta}_{\text{DINA}}^{(1)}, \boldsymbol{\theta}_{\text{DINA}}^{(2)}, \boldsymbol{\pi}^{\text{deep}}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \mid \mathbf{R}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}) \\
\propto &\prod_{i=1}^{N}\prod_{j=1}^{J}\left[(1-s_{1,j})^{\xi_{1,ij}} g_{1,j}^{1-\xi_{1,ij}}\right]^{r_{i,j}} \left[s_{1,j}^{\xi_{1,ij}}(1-g_{1,j})^{1-\xi_{1,ij}}\right]^{1-r_{i,j}} \\
&\times \prod_{i=1}^{N}\prod_{\ell=1}^{2^{K_2}}\left\{\pi_\ell \prod_{k=1}^{K_1}\left[(1-s_{2,k})^{\xi_{2,ik}} g_{2,k}^{1-\xi_{2,ik}}\right]^{a_{i,k}^{(1)}} \left[s_{2,k}^{\xi_{2,ik}}(1-g_{2,k})^{1-\xi_{2,ik}}\right]^{1-a_{i,k}^{(1)}}\right\}^{\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell)} \\
&\times \prod_{j=1}^{J}[s_{1,j}^{a_s-1}(1-s_{1,j})^{b_s-1} g_{1,j}^{a_g-1}(1-g_{1,j})^{b_g-1}\mathbb{1}(g_{1,j} < 1-s_{1,j})] \\
&\times \prod_{k=1}^{K_1}[s_{2,k}^{a_s-1}(1-s_{2,k})^{b_s-1} g_{2,k}^{a_g-1}(1-g_{2,k})^{b_g-1}\mathbb{1}(g_{2,k} < 1-s_{2,k})] \times \prod_{\ell=1}^{2^{K_2}}\pi_\ell^{\delta-1}
\end{aligned}$$

Based on the above posterior, the full conditional distributions of the quantities $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\pi}^{\text{deep}}$, $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ are as follows.

(1) Sample $s_{1,j}^{(1)}$ and $g_{1,j}^{(1)}$ from truncated Beta distributions:

$$\begin{aligned}
s_j^{(1)} &\sim \text{Beta}\left(1+\sum\nolimits_{i=1}^{N}(1-r_{ij})\xi_{1,ij},\ 1+\sum\nolimits_{i=1}^{N} r_{ij}\xi_{1,ij}\right)\cdot \mathbb{1}(s_j^{(1)} < 1-g_j^{(1)}); \\
g_j^{(1)} &\sim \text{Beta}\left(1+\sum\nolimits_{i=1}^{N} r_{ij}(1-\xi_{1,ij}),\ 1+\sum\nolimits_{i=1}^{N}(1-r_{ij})(1-\xi_{1,ij})\right)\cdot \mathbb{1}(g_j^{(1)} < 1-s_j^{(1)}).
\end{aligned}$$

(2) Sample $s_{2,k}^{(2)}$ and $g_{2,k}^{(2)}$ from truncated Beta distributions:

$$s_k^{(2)} \sim \text{Beta}\left(1+\sum_{i=1}^N (1-a_{ik}^{(1)})\xi_{2,ik},\ 1+\sum_{i=1}^N a_{ik}^{(1)}\xi_{2,ik}\right)\cdot \mathbb{1}(s_k^{(2)} < 1-g_k^{(2)});$$
$$g_k^{(2)} \sim \text{Beta}\left(1+\sum_{i=1}^N a_{ik}^{(1)}(1-\xi_{2,ik}),\ 1+\sum_{i=1}^N (1-a_{ik}^{(1)})(1-\xi_{2,ik})\right)\cdot \mathbb{1}(g_k^{(2)} < 1-s_k^{(2)}).$$

(3) Sample $\boldsymbol{\pi}^{\text{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\text{deep}} \sim \text{Dirichlet}\left(\delta_1+\sum_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_1),\ldots,\delta_{2^{K_2}}+\sum_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{2^{K_2}})\right).$$

(4) Sample each entry $a_{i,k}^{(1)}$ from the Bernoulli distribution with the following probability:

$$\begin{aligned}\mathbb{P}(a_{i,k}^{(1)}=1 \mid -) &= \mathbb{P}(a_{i,k}^{(1)}=1 \mid \boldsymbol{r}_i, \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \\ &= \frac{\mathbb{P}(a_{i,k}^{(1)}=1 \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=1, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})}{\sum_{x=0,1}\mathbb{P}(a_{i,k}^{(1)}=x \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=x, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})},\end{aligned}$$

where the conditional distributions $\mathbb{P}(a_{i,k}^{(1)}=x \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=x, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})$ just directly follow the likelihood defined under the DeepDINA model in Section 4.1 of the main text, and they are both DINA.

(5) Sample each pattern $\boldsymbol{a}_i^{(2)}$ from the categorical distribution with $|\{0,1\}^{K_2}| = 2^{K_2}$ components with the following probabilities:

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid -) &= \mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\pi}^{\text{deep}}); \\ &= \frac{\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell, \boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})},\end{aligned}$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})$ also directly follow the definition of DeepDINA, with the former being a Dirichlet distribution and the latter following a DINA model conditional distribution.

Overall, our Gibbs sampler cycles through the above five steps iteratively to approximate the posterior distributions of all the quantities.

### S.2.2 Gibbs Sampler for Hybrid GDINA-DINA

Recall that we will focus on those $\theta_{j,S}^{(1)}$ parameters for the shallower GDINA layer during the Gibbs sampling, which denote conditional positive response probabilities:

$$\theta_{j,S}^{(1)} = \sum_{S' \subseteq S} \beta_{j,S'}^{(1)} = \mathbb{P}(r_{i,j} = 1 \mid \boldsymbol{a}_i^{(1)\top} \boldsymbol{q}_{j,S}^{(1)} = \boldsymbol{q}_{j,S}^{(1)\top} \boldsymbol{q}_{j,S}^{(1)}).$$

Introduce binary indicators for the GDINA layer as

$$\xi_{1,ij,S} = \mathbb{1}\left(\boldsymbol{a}_i^{(1)\top} \boldsymbol{q}_{j,S}^{(1)} = \boldsymbol{q}_{j,S}^{(1)\top} \boldsymbol{q}_{j,S}^{(1)}\right), \quad i \in [N],\ j \in [J],\ S \subseteq \mathcal{K}_j,$$

where the notation $\mathcal{K}_j = \{k \in [K_1] :\ q_{j,k}^{(1)} = 1\}$ was defined in the main text. For the deeper DINA layer, we still introduce binary ideal response indicators $\xi_{2,ik}$ for $k \in [K_1]$ similarly as the previous (S.5). Under the priors specified in the main text, the posterior distribution in the Hybrid GDINA-DINA can be written as

$$\begin{aligned}
&p(\boldsymbol{\theta}_{\text{GDINA}}^{(1)}, \boldsymbol{\theta}_{\text{DINA}}^{(2)}, \boldsymbol{\pi}^{\text{deep}}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \mid \mathbf{R}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}) \\
\propto &\prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{S \subseteq \mathcal{K}_j} \left[ \left(\theta_{j,S}^{(1)}\right)^{r_{i,j}\xi_{1,ij,S}} \left(1 - \theta_{j,S}^{(1)}\right)^{(1-r_{i,j})\xi_{1,ij,S}} \right] \\
\times &\prod_{i=1}^{N} \prod_{\ell=1}^{2^{K_2}} \left\{ \pi_\ell \prod_{k=1}^{K_1} \left[ (1 - s_{2,k})^{\xi_{2,ik}} g_{2,k}^{1-\xi_{2,ik}} \right]^{a_{i,k}^{(1)}} \left[ s_{2,k}^{\xi_{2,ik}} (1 - g_{2,k})^{1-\xi_{2,ik}} \right]^{1-a_{i,k}^{(1)}} \right\}^{\mathbb{1}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell)} \\
\times &\prod_{j=1}^{J} \prod_{S \subseteq \mathcal{K}_j} \left[ (\theta_{j,S}^{(1)})^{a_\theta - 1} (1 - \theta_{j,S}^{(1)})^{a_\theta - 1} \mathbb{1}(\theta_{j,S}^{(1)} > \theta_{j,\varnothing}^{(1)} \text{ if } S \text{ is a singleton set}) \right] \\
\times &\prod_{k=1}^{K_1} [s_{2,k}^{a_s - 1} (1 - s_{2,k})^{b_s - 1} g_{2,k}^{a_g - 1} (1 - g_{2,k})^{b_g - 1} \mathbb{1}(g_{2,k} < 1 - s_{2,k})] \times \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\delta - 1}.
\end{aligned}$$

Our Gibbs sampler will cycle through the following steps iteratively.

(1) Sample each $\theta_{j,S}^{(1)}$ from the (truncated) Beta distribution:

$$\theta_{j,S}^{(1)} \sim \text{Beta}\left(a_\theta + \sum_{i=1}^{N} r_{i,j}\xi_{1,ij,S},\ b_\theta + \sum_{i=1}^{N} (1 - r_{i,j})\xi_{1,ij,S}\right) \mathbb{1}(\theta_{j,S}^{(1)} > \theta_{j,\varnothing}^{(1)} \text{ if } S \text{ is a singleton set}).$$

(2) Sample $s_{2,k}^{(2)}$ and $g_{2,k}^{(2)}$ from truncated Beta distributions:

$$s_k^{(2)} \sim \text{Beta}\left(1+\sum_{i=1}^N (1-a_{ik}^{(1)})\xi_{2,ik},\ 1+\sum_{i=1}^N a_{ik}^{(1)}\xi_{2,ik}\right)\cdot \mathbb{1}(s_k^{(2)} < 1-g_k^{(2)});$$
$$g_k^{(2)} \sim \text{Beta}\left(1+\sum_{i=1}^N a_{ik}^{(1)}(1-\xi_{2,ik}),\ 1+\sum_{i=1}^N (1-a_{ik}^{(1)})(1-\xi_{2,ik})\right)\cdot \mathbb{1}(g_k^{(2)} < 1-s_k^{(2)}).$$

(3) Sample $\boldsymbol{\pi}^{\text{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\text{deep}} \sim \text{Dirichlet}\left(\delta_1+\sum_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_1),\ldots,\delta_{2^{K_2}}+\sum_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{2^{K_2}})\right).$$

(4) Sample each entry $a_{i,k}^{(1)}$ from the Bernoulli distribution with the following probability:

$$\mathbb{P}(a_{i,k}^{(1)}=1 \mid -) = \frac{\mathbb{P}(a_{i,k}^{(1)}=1 \mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=1,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})}{\sum_{x=0,1}\mathbb{P}(a_{i,k}^{(1)}=x \mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=x,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})},$$

where the conditional distributions $\mathbb{P}(a_{i,k}^{(1)}=x \mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)}=x,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})$ follow the likelihood under the DINA and GDINA, respectively.

(5) Sample each pattern $\boldsymbol{a}_i^{(2)}$ from the categorical distribution with $|\{0,1\}^{K_2}| = 2^{K_2}$ components with the following probabilities:

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid -) &= \mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid \boldsymbol{a}_i^{(1)},\boldsymbol{\theta}^{(2)},\boldsymbol{\pi}^{\text{deep}});\\ &= \frac{\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell,\boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'},\boldsymbol{\theta}^{(2)})},\end{aligned}$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'},\boldsymbol{\theta}^{(2)})$ also directly follow the definition of DeepDINA, with the former being a Dirichlet distribution and the latter following a DINA model conditional distribution.

### S.2.3 Gibbs Sampler for Two-latent-layer DeepLLM

The posterior distribution of the two-latent-layer DeepLLM can be written as

$$p(\boldsymbol{\beta}_{\text{LLM}}^{(1)},\boldsymbol{\beta}_{\text{LLM}}^{(2)},\boldsymbol{\pi}^{\text{deep}},\mathbf{A}^{(1)},\mathbf{A}^{(2)} \mid \mathbf{R},\mathbf{Q}^{(1)},\mathbf{Q}^{(2)})$$

$$\propto \prod_{i=1}^{N}\left\{\prod_{j=1}^{J}\frac{\exp\left(r_{i,j}\left(\beta_{j,0}^{(1)}+\sum_{k=1}^{K_1}q_{j,k}^{(1)}\beta_{j,k}^{(1)}a_{i,k}^{(1)}\right)\right)}{1+\exp\left(\beta_{j,0}^{(1)}+\sum_{k=1}^{K_1}q_{j,k}^{(1)}\beta_{j,k}^{(1)}a_{i,k}^{(1)}\right)}\times\prod_{k=1}^{K_1}\frac{\exp\left(a_{i,k}^{(1)}\left(\beta_{k,0}^{(2)}+\sum_{m=1}^{K_2}q_{k,m}^{(2)}\beta_{k,m}^{(2)}a_{i,m}^{(2)}\right)\right)}{1+\exp\left(\beta_{k,0}^{(2)}+\sum_{m=1}^{K_2}q_{k,m}^{(2)}\beta_{k,m}^{(2)}a_{i,m}^{(2)}\right)}\right\}$$

$$\times\prod_{i=1}^{N}\prod_{\ell=1}^{2^{K_2}}\pi_\ell^{\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell)}\times\prod_{\ell=1}^{2^{K_2}}\pi_\ell^{\delta-1}\times\prod_{j=1}^{J}\left\{N(\beta_{j,0}^{(1)}\mid 0,\sigma_\beta^2)\prod_{k=0}^{K_1}N(\beta_{j,k}^{(1)}\mid 0,\sigma_\beta^2)\mathbb{1}(\beta_{j,k}^{(1)}>0 \text{ if } q_{j,k}^{(1)}=1)\right\}$$

$$\times\prod_{k=1}^{K_1}\left\{N(\beta_{k,0}^{(2)}\mid 0,\sigma_\beta^2)\prod_{m=0}^{K_2}N(\beta_{k,m}^{(2)}\mid 0,\sigma_\beta^2)\mathbb{1}(\beta_{k,m}^{(2)}>0 \text{ if } q_{k,m}^{(2)}=1)\right\}$$

$$\times\prod_{i=1}^{N}\prod_{j=1}^{J}\mathrm{PG}(w_{i,j}^{(1)}\mid 1,0)\cdot\prod_{i=1}^{N}\prod_{k=1}^{K_1}\mathrm{PG}(w_{i,k}^{(2)}\mid 1,0).$$

Our Gibbs sampler iteratively cycles through the following steps.

(1) Recall the notation $\mathcal{K}_j=\{k\in[K_1]:\ q_{j,k}^{(1)}=1\}$. Define

$$\boldsymbol{\beta}_{j,\mathcal{K}_j}^{(1)}=(\beta_{j,0}^{(1)},\ \beta_{j,k}^{(1)};\ k\in\mathcal{K}_j),$$

which is a vector of length $|\mathcal{K}_j|+1$. We introduce a notation $\mathbf{X}_j^{(1)}$, which is a $N\times|\mathcal{K}_j|$ matrix; the entries in this matrix are indexed by $a_{i,k}^{(1)}q_{j,k}^{(1)}$ where $i\in[N]$ and $k\in\{0\}\cup\mathcal{K}_j$. Sample $\boldsymbol{\beta}_{j,\mathcal{K}_j}^{(1)}$ from the (truncated) Multivariate Normal (MVN) distribution:

$$\boldsymbol{\beta}_{j,\mathcal{K}_j}^{(1)}\sim\mathrm{MVN}(\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j}),\quad\text{where}$$
$$\boldsymbol{\Sigma}_{1j}=\left(\mathbf{X}_j^{(1)\top}\mathrm{diag}\left(\mathbf{W}_{:,j}^{(1)}\right)\mathbf{X}_j^{(1)}\right)^{-1},\quad\boldsymbol{\mu}_{1j}=\boldsymbol{\Sigma}_{1j}\mathbf{X}_j^{(1)\top}\left(\mathbf{R}_{:,j}-1/2\right).$$

(2) Define a new notation

$$\mathcal{K}_{2,k}=\{m\in[K_2]:\ q_{k,m}^{(2)}=1\}.$$

Define

$$\boldsymbol{\beta}_{k,\mathcal{K}_{2,k}}^{(2)}=(\beta_{k,0}^{(2)},\ \beta_{k,m}^{(2)};\ m\in\mathcal{K}_{2,k}),$$

which is a vector of length $|\mathcal{K}_{2,k}|+1$. We introduce a notation $\mathbf{X}_k^{(2)}$, which is a $N\times|\mathcal{K}_{2,k}|$ matrix; the entries in this matrix are indexed by $a_{i,m}^{(2)}q_{k,m}^{(2)}$ where $i\in[N]$ and $m\in\{0\}\cup\mathcal{K}_{2,k}$. Sample $\boldsymbol{\beta}_{k,\mathcal{K}_{2,k}}^{(2)}$ from the (truncated) Multivariate Normal (MVN)

distribution:

$$\boldsymbol{\beta}^{(2)}_{k,\mathcal{K}_{2,k}} \sim \text{MVN}(\boldsymbol{\mu}_{2k}, \boldsymbol{\Sigma}_{2k}), \quad \text{where}$$
$$\boldsymbol{\Sigma}_{2k} = \left(\mathbf{X}^{(2)\top}_k \text{diag}\left(\mathbf{W}^{(2)}_{:,k}\right)\mathbf{X}^{(2)}_k\right)^{-1}, \quad \boldsymbol{\mu}_{2k} = \boldsymbol{\Sigma}_{2k}\mathbf{X}^{(2)\top}_k\left(\mathbf{A}^{(1)}_{:,k} - 1/2\right).$$

(3) Sample each $w^{(1)}_{i,j}$, $j \in [J]$ from the Polya-Gamma distribution:

$$w^{(1)}_{i,j} \sim \text{PG}\Big(1, \beta^{(1)}_{j,0} + \sum_{k\in\mathcal{K}_j} \beta^{(1)}_{j,k} a^{(1)}_{i,k}\Big).$$

(4) Sample each $w^{(2)}_{i,k}$, $k \in [K_1]$ from the Polya-Gamma distribution:

$$w^{(2)}_{i,k} \sim \text{PG}\Big(1, \beta^{(2)}_{k,0} + \sum_{m\in\mathcal{K}_{2,k}} \beta^{(2)}_{k,m} a^{(2)}_{i,m}\Big).$$

(5) Sample $\boldsymbol{\pi}^{\text{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\text{deep}} \sim \text{Dirichlet}\left(\delta_1 + \textstyle\sum_{i=1}^N \mathbb{1}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_1), \ldots, \delta_{2^{K_2}} + \sum_{i=1}^N \mathbb{1}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_{2^{K_2}})\right).$$

(6) Sample each entry $a^{(1)}_{i,k}$ from the Bernoulli distribution with the following probability:

$$\mathbb{P}(a^{(1)}_{i,k} = 1 \mid -) = \frac{\mathbb{P}(a^{(1)}_{i,k} = 1 \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = 1, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})}{\sum_{x=0,1}\mathbb{P}(a^{(1)}_{i,k} = x \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = x, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})},$$

where the conditional distributions $\mathbb{P}(a^{(1)}_{i,k} = x \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = x, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})$ both follow the likelihood under the LLM.

(7) Sample each pattern $\boldsymbol{a}^{(2)}_i$ from the categorical distribution with $|\{0,1\}^{K_2}| = 2^{K_2}$ components with the following probabilities:

$$\mathbb{P}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_\ell \mid -) = \mathbb{P}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_\ell \mid \boldsymbol{a}^{(1)}_i, \boldsymbol{\theta}^{(2)}, \boldsymbol{\pi}^{\text{deep}});$$
$$= \frac{\mathbb{P}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_\ell \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}^{(1)}_i \mid \boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_\ell, \boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}}\mathbb{P}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}^{(1)}_i \mid \boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})},$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})$ also directly follow the definition of LLM, with the former being a Dirichlet distribution and the latter following a LLM model conditional distribution.