*Submitted to the Annals of Applied Statistics*

# LATENT CONJUNCTIVE BAYESIAN NETWORK: UNIFY ATTRIBUTE HIERARCHY AND BAYESIAN NETWORK FOR COGNITIVE DIAGNOSIS

BY SEUNGHYUN LEE[a] AND YUQI GU[b]

*Department of Statistics, Columbia University,* [a]*sl4963@columbia.edu;* [b]*yuqi.gu@columbia.edu*

Cognitive diagnostic assessment aims to measure specific knowledge structures in students. To model data arising from such assessments, cognitive diagnostic models with discrete latent variables have gained popularity in educational and behavioral sciences. In a learning context, the latent variables often denote sequentially acquired skill attributes, which is often modeled by the so-called attribute hierarchy method. One drawback of the traditional attribute hierarchy method is that its parameter complexity varies substantially with the hierarchy's graph structure, lacking statistical parsimony. Additionally, arrows among the attributes do not carry an interpretation of statistical dependence. Motivated by these, we propose a new family of *latent conjunctive Bayesian networks* (LCBNs), which rigorously unify the attribute hierarchy method for sequential skill mastery and the Bayesian network model in statistical machine learning. In an LCBN, the latent graph not only retains the hard constraints on skill prerequisites as an attribute hierarchy, but also encodes nice conditional independence interpretation as a Bayesian network. LCBNs are identifiable, interpretable, and parsimonious statistical tools to diagnose students' cognitive abilities from assessment data. We propose an efficient two-step EM algorithm for structure learning and parameter estimation in LCBNs, and establish the consistency of this procedure. Application of our method to an international educational assessment dataset gives interpretable findings of cognitive diagnosis.

**1. Introduction.** Cognitive diagnostic assessment aims to measure specific knowledge structures and processing skills in students (Leighton and Gierl, 2007). To model data arising from such assessments, *cognitive diagnostic models* (CDMs) with discrete latent variables (also called diagnostic classification models; see Rupp et al., 2010; von Davier and Lee, 2019) have recently gained great popularity in educational, psychological, and behavioral applications. CDMs adopt a set of discrete latent *attributes* with substantive meaning to explain a subject's multivariate responses to a set of items. "Attribute" here is a generic term that can represent unobserved psychological constructs including skills, knowledge states, conceptual understandings, cognitive processes, and rules (Wang, 2021). In educational settings, each attribute often represents the mastery/deficiency of a specific latent skill. Adopting CDMs in educational assessment can generate fine-grained diagnoses about students' multiple latent skills, and hence provide detailed feedback about their weaknesses and strengths.

A typical CDM consists of a *structural model* for the latent attributes and a *measurement model* describing the dependence of the observed variables (i.e., item responses in educational assessments) on the latent attributes. The measurement model is accompanied by a so-called $Q$-matrix (Tatsuoka, 1983), summarizing which subset of the attributes each observed variable measures or requires. The $Q$-matrix is often pre-specified by domain experts. Various measurement models have been proposed for different diagnostic purposes. For example, the popular and fundamental Deterministic Input Noisy Output "AND" gate (DINA;

*Keywords and phrases:* Attribute hierarchy, Bayesian network, Cognitive diagnostic model, Directed graphical model, EM algorithm, Identifiability.

Junker and Sijtsma, 2001) model adopts the conjunctive assumption by specifying that a student needs to master all attributes required by an item to be capable of it. The generalized DINA (GDINA; de la Torre, 2011) model generalizes this by incorporating main effects and interaction effects of required attributes into the measurement model. Other popular CDMs include the Deterministic Input Noisy Output "OR" gate (DINO; Templin and Henson, 2006) model, the log-linear CDM (LCDM; Henson et al., 2009), the additive CDM (ACDM; de la Torre, 2011), and general diagnostic models (GDM; von Davier, 2008).

As for the structural model for the latent attributes in a CDM, the *attribute hierarchy method* that models sequential skill mastery has recently attracted increasing attention (Leighton et al., 2004; Gierl et al., 2007; Wang and Gierl, 2011; Templin and Bradshaw, 2014; Gu and Xu, 2019; Wang and Lu, 2021). Students' learning is not instantaneous and often proceeds in a sequential and dependent manner. In a learning context, possessing lower level skills are often believed to be the prerequisite for possessing higher level skills (Simon and Tzur, 2012; Briggs and Alonzo, 2012). Leighton et al. (2004) first proposed the attribute hierarchy method, and Templin and Bradshaw (2014) integrated the attribute hierarchy with a flexible measurement model in a statistical framework to define the family of hierarchical cognitive diagnostic models.

Most existing studies on attribute hierarchy followed Templin and Bradshaw (2014) to adopt unstructured proportion parameters, which characterizes how much proportion of the student population possess this skill pattern. One limitation of this popular approach is that its parameter complexity varies substantially with the graph structure of the hierarchy, lacking statistical parsimony. For instance, with $K$ binary attributes, a chain graph hierarchy requires $K$ free parameters for the latent distribution, whereas a graph with one attribute serving as a common parent to all the other attributes requires $2^{K-1}$ parameters. This lack of parsimony especially creates computational and statistical challenges when there are a large number of attributes and a limited sample size. In addition, the unstructured model for attribute hierarchy does not endow the hierarchy graph with any probabilistic interpretation. Specifically, the hierarchy among the latent attributes is merely treated as a machinery for inducing hard constraints on which latent attribute patterns are permissible (those respecting the hierarchy) and which are forbidden (those violating the hierarchy). As a result, the arrows in such an attribute hierarchy graph do not carry clear interpretation of direct statistical dependence, nor does the lack of arrows indicate conditional independence.

Motivated by the above issues, we propose a new family of *latent conjunctive Bayesian networks* (LCBNs) for cognitive diagnosis. LCBNs are a parsimonious and interpretable class of probabilistic graphical models that rigorously unify attribute hierarchy and Bayesian network. A Bayesian network (Pearl, 1988) is a directed graphical model of random variables, in which directed arrows indicate statistical dependence and the lack of arrows indicate conditional independence. In our LCBN, the directed acyclic graph among the latent attributes *not only* respects the hard constraints on which attribute patterns are permissible/forbidden as under a usual attribute hierarchy, *but also* encodes the nice conditional independence interpretation as in a usual Bayesian network. Therefore, LCBNs enjoy the best of both worlds. Moveover, LCBNs are parsimonious statistical models with a fixed parameter complexity in the latent part – it always only requires $K$ parameters for specifying the joint distribution of $K$ binary latent attributes, regardless of the graph structure of the hierarchy.

In terms of model identifiability, we prove that the attribute hierarchy graph and all the continuous parameters in an LCBN are fully identifiable from the observed data distribution. Our identifiability conditions are transparent requirements on the discrete structure in the model. Identifiability lays the foundation for valid statistical estimation and inference. In terms of estimation, we propose an efficient two-step EM algorithm to perform structure learning and parameter estimation in LCBNs. In the first step, we leverage a penalized EM

algorithm for selecting significant latent patterns (Gu and Xu, 2019) to estimate the discrete structure – the attribute hierarchy graph. In the second step, we fix the attribute hierarchy and propose another EM algorithm to estimate the continuous parameters in the LCBN. Simulation studies demonstrate the estimation accuracy of this procedure. We apply our method to analyze a dataset extracted from an international educational assessment, the Trends in Mathematics and Science Study (TIMSS). The real data analysis gives interpretable finds of cognitive diagnosis and demonstrates the wide applicability of our method.

The remainder of this paper is organized as follows. Section 2 introduces the background of cognitive diagostic modeling, proposes the general framework of LCBNs, and discusses some related work. Section 3 provides identifiability conditions of LCBNs. Section 4 proposes a two-step EM algorithm to estimate the attribute hierarchy graph and model parameters in LCBNs. Section 5 presents simulation studies to empirically assess the proposed method. Section 6 applies the new method to analyze an international educational assessment dataset. Finally, Section 7 provides concluding remarks and discusses future directions. We provide the proofs of the theorems, additional identifiability results, and additional simulation studies in the Supplementary Material.

## 2. Latent Conjunctive Bayesian Network.

2.1. *Cognitive Diagnostic Modeling with an attribute hierarchy.* We first introduce the basic setup of a CDM. Consider a CDM for modeling a cognitive diagnostic assessment. A student's observed variables are his or her correct/wrong responses to a set of $J$ items in the assessment, denoted by $\mathbf{R} = (R_1, \ldots, R_J) \in \{0,1\}^J$, in which $R_j = 1$ indicates the student's response to the $j$th item is correct and $R_j = 0$ otherwise. A student's latent variables are his or her profile of presence/absence of a set of $K$ skill attributes, denoted by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \{0,1\}^K$, in which $\alpha_k = 1$ indicates the student masters the $k$th skill and $\alpha_k = 0$ otherwise. Typically, a CDM consists of two parts: a *structural model* for the latent attributes, and a *measurement model* to describe the distribution of the observed responses given the latent. In a learning context, the skill attributes are often sequentially acquired and form a hierarchy with prerequisite relations among attributes. In a CDM with attribute hierarchy, the key elements of the structural and measurement modeling parts are captured by two discrete graph structures: a directed acyclic graph among the latent attributes, and a bipartite directed graph pointing from the latent attributes to the observed responses. These two graphical structures are illustrated in Figure 1. For clarity of presentation, we next describe the measurement part and the structural part of a CDM separately in subsequent paragraphs.

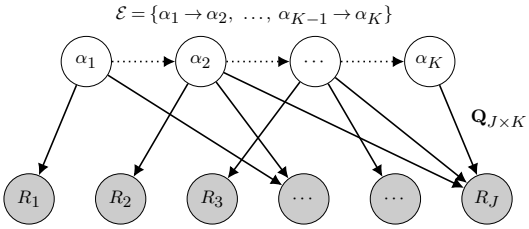


Fig 1: Graphical model representation of a cognitive diagnostic model with a linear attribute hierarchy. White nodes are latent attributes, and grey nodes are observed responses. Dotted arrows denote the prerequisite relationship among the latent attributes, and solid arrows denote the conditional dependence structure of the observed responses given the latent attributes.

For the *measurement* part of a CDM, educational experts who designed the assessment usually provide information about which subset of the $K$ skills each test item measures. All

such information are summarized in a so-called $Q$-matrix (Tatsuoka, 1983). The $Q$-matrix $\mathbf{Q} = (q_{j,k}) \in \{0,1\}^{J\times K}$ is a $J \times K$ matrix with binary entries, with rows indexed by observed items and columns by latent attributes. Each entry $q_{j,k} = 1$ or 0 indicates whether or not the $j$th test item requires/measures the $k$th latent skill. Consequently, the $j$th row vector of $\mathbf{Q}$, denoted by $\boldsymbol{q}_j = (q_{j,1}, \ldots, q_{j,K})$, is the attribute requirement profile of item $j$. For example, in Figure 1 we have $\boldsymbol{q}_1 = (1,0,0,0)$ since the first item only requires the first attribute.

Statistically, a student's responses to the $J$ items are assumed to be conditionally independent given his or her latent attribute profile $\boldsymbol{\alpha}$. Such a local independence assumption is widely adopted in various models for item response data. We collect all the conditional correct response probabilities in a $J \times 2^K$ item parameter matrix $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$, with rows indexed by the $J$ test items and columns by the $2^K$ binary pattern configurations in $\{0,1\}^K$. For any $j \in [J]$ and $\boldsymbol{\alpha} \in \{0,1\}^K$, the entry $\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha})$ defines the conditional probability of giving a correct response to item $j$ given that one has a latent skill profile $\boldsymbol{\alpha}$. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$ of the same length, we write $\boldsymbol{a} \succeq \boldsymbol{b}$ if $a_k \geq b_k$ for all $k \in [K]$ and write $\boldsymbol{a} \not\succeq \boldsymbol{b}$ otherwise. Importantly, since $\boldsymbol{q}_j$ describes which subset of attributes item $j$ measures, the correct response probability $\theta_{j,\boldsymbol{\alpha}}$ only depends on those attributes $\alpha_k$ that are measured by item $j$ (that is, those $\alpha_k$ with $q_{j,k} = 1$). Therefore,

$$\theta_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha}, \boldsymbol{\alpha}' \succeq \boldsymbol{q}_j. \tag{1}$$

Another common feature shared by many measurement models is that item parameters exhibit monotonicity (Xu and Shang, 2018; Gu and Xu, 2019; Balamuta and Culpepper, 2022):

$$\theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha} \succeq \boldsymbol{q}_j \text{ and } \boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j. \tag{2}$$

The above inequality can be interpreted as: if a student possesses all the attributes required by item $j$ (that is, $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$), then this student has a higher probability to give a correct response to this item compared to other subjects who lack some required attribute.

We next review some popular and widely used CDM measurement models.

EXAMPLE 1 (DINA model). *The Deterministic Input Noisy output "And" gate (DINA; Junker and Sijtsma, 2001) model is a very popular and fundamental CDM, with the advantages of parsimony and interpretability. For each item $j$, DINA uses exactly two distinct parameters to describe the correct response probability $\theta_{j,\boldsymbol{\alpha}}$. Specifically, if a student has a latent profile $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, then he/she is considered capable of this item but still has a small probability $s_j$ to make a careless mistake; on the other hand, if the student lacks some of the required attributes with $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j$, then he/she is incapable of this item but still has a small probability $g_j$ to have a lucky guess. The correct response probability can be written as*

$$\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}) = \begin{cases} 1 - s_j, & \text{if } \boldsymbol{\alpha} \succeq \boldsymbol{q}_j; \\ g_j, & \text{if } \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j. \end{cases} \tag{3}$$

*$s_j$ and $g_j$ are called slipping parameter and guessing parameter, respectively. The monotonicity inequality in (2) now boils down to $1 - s_j > g_j$ for all $j$.*

EXAMPLE 2 (All-effect CDMs). *All-effect CDMs generalize the DINA model by considering both the main effects and all the interaction effects of the required attributes. The item parameter $\theta_{j,\boldsymbol{\alpha}}$ can be written as*

$$\theta_{j,\boldsymbol{\alpha}} = f(\delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k} q_{j,k} \alpha_k + \sum_{1\leq k<k'\leq K} \delta_{j,kk'} (q_{j,k}\alpha_k)(q_{j,k'}\alpha_{k'}) + \cdots + \delta_{j,1\cdots K} \prod_{k=1}^{K} (q_{j,k}\alpha_k)), \tag{4}$$

*where $\delta_{j,k}$ is the main effect of the required attribute $\alpha_k$, and $\delta_{j,kk'}$ is the interaction effect between two required attributes $\alpha_k$ and $\alpha_{k'}$, etc. When the link function $f$ is the identity,* (4) *gives the Generalized DINA model (GDINA; de la Torre, 2011); when $f$ is the inverse logit,* (4) *gives the Log-linear CDM (LCDM; Henson et al., 2009); also see the General Diagnostic Model (GDM) framework in von Davier (2008).*

We now describe the *structural* part of a CDM with an attribute hierarchy. The hierarchy is a collection of prerequisite relations between the $K$ latent attributes, in which possessing more basic skills are assumed to be the prerequisite for possessing more advanced ones. For any $1 \le k \ne \ell \le K$, we say that attribute $k$ is a *prerequisite* for attribute $\ell$ (and denote this by $\alpha_k \to \alpha_\ell$ or simply $k \to \ell$) if any latent skill pattern $\boldsymbol{\alpha} \in \{0,1\}^K$ with $\alpha_k = 0$ and $\alpha_\ell = 1$ does not exist in the student population and is not "permissible". In other words, for any student that masters the higher level advanced skill $\alpha_\ell$, he/she must have mastered the lower level basic skill $\alpha_k$. Denote the collection of all the prerequisite relationships by

$$\mathcal{E} = \{k \to \ell : \text{ the } k\text{th skill attribute is a prerequisite for the } \ell\text{th skill attribute}\}.$$

Any attribute hierarchy $\mathcal{E}$ can be visualized as a directed acyclic graph among $K$ nodes, each node representing a latent attribute. For example, Figure 1 illustrates a linear hierarchy among the skills with $\mathcal{E} = \{\alpha_1 \to \alpha_2,\ \alpha_2 \to \alpha_3,\ \ldots,\ \alpha_{K-1} \to \alpha_K\}$.

For a CDM with an attribute hierarchy, most existing studies followed Templin and Bradshaw (2014) to adopt an unstructured statistical model for the hierarchy. Specifically, such a model is based on the observation that any nonempty $\mathcal{E}$ induces a sparsity structure on the $2^K$–dimensional proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$. For example, if $k \to \ell$, then as aforementioned, any pattern $\boldsymbol{\alpha}$ with $\alpha_k = 0$ but $\alpha_\ell = 1$ does not exist in the population and hence its population proportion $p_{\boldsymbol{\alpha}} = 0$. In this way, we can define the set of permissible latent skill patterns under a hierarchy $\mathcal{E}$ as follows:

$$\mathcal{A}(\mathcal{E}) = \{\boldsymbol{\alpha} \in \{0,1\}^K : \boldsymbol{\alpha} \text{ is permissible under } \mathcal{E}\} = \{\boldsymbol{\alpha} \in \{0,1\}^K : p_{\boldsymbol{\alpha}} > 0 \text{ under } \mathcal{E}\}. \tag{5}$$

Note that $\mathcal{A}(\mathcal{E})$ is fully determined by the attribute hierarchy $\mathcal{E}$.

Since the hierarchy $\mathcal{E}$ is a directed acyclic graph among $K$ attributes, it can also be equivalently represented by a $K \times K$ *reachability matrix* $\mathbf{G}(\mathcal{E})$ (also denoted by $\mathbf{G}$ for short) in the graph theory terminology. The $(k,\ell)$th entry of $\mathbf{G}$ is a binary indicator of whether the $k$th skill is the prerequisite for the $\ell$th skill, that is, $G_{k,\ell} = \mathbb{1}(k \to \ell)$. Here we assume the diagonal entries of $\mathbf{G}$ are all zero.

EXAMPLE 3. *Consider an example with $K = 4$ skill attributes and a hierarchy $\mathcal{E} = \{1 \to 3,\ 1 \to 4,\ 2 \to 3,\ 2 \to 4\}$. This hierarchy means that the first two skills are the basic ones that serve as the prerequisites for the last two advanced skills. This $\mathcal{E}$ is visualized in the left panel of Figure 2. There are seven permissible attribute patterns under $\mathcal{E}$:*

$$\mathcal{A}(\mathcal{E}) = \{0000,\ 1000,\ 0100,\ 1100,\ 1110,\ 1101,\ 1111\}. \tag{6}$$

*The patterns in $\mathcal{A}(\mathcal{E})$ can also be viewed as forming a distributive lattice, a concept in combinatorics (Gratzer, 2009), as shown in the middle panel of Figure 2. The corresponding reachability matrix $\mathbf{G}$ under $\mathcal{E}$ is shown in the rightmost panel of Figure 2.*

It is worth emphasizing the distinction between a directed acyclic graph (DAG) in the usual attribute hierarchy method and that in a conventional Bayesian network model (Pearl, 1988, or equivalently, a probabilistic directed graphical model). Specifically, the arrows in the DAG among the latent attributes (as shown in Figure 2) generally *cannot* be interpreted as encoding direct statistical dependence, nor does the lack of arrows indicate conditional independence. Rather, such a DAG merely encodes certain hard constraints on what attribute patterns are

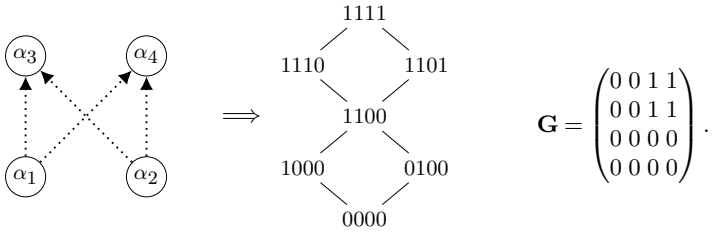


Fig 2: An example with $K = 4$ skill attributes. Left: attribute hierarchy graph $\mathcal{E}$. Middle: all the allowable attribute patterns in $\mathcal{A}(\mathcal{E})$. Right: $K \times K$ reachability matrix $\mathbf{G}$.

permissible (those $\boldsymbol{\alpha} \in \mathcal{A}(\mathcal{E})$) and which are forbidden (those $\boldsymbol{\alpha} \in \{0,1\}^K \setminus \mathcal{A}(\mathcal{E})$). In contrast, the DAG in a Bayesian network has arrows capturing the statistical dependence between the random variables, and generally does not forbid any configurations.

A natural and interesting question is – can we introduce a new family of models that rigorously unify the above two models and inherit the advantages of both? This question is particularly relevant considering the drawbacks of the existing attribute hierarchy method, including not only the lack of interpretability, but also the lack of statistical parsimony. To see this, consider an attribute hierarchy $\mathcal{E} = \{1 \to 2,\ 1 \to 3,\ \ldots,\ 1 \to K\}$ where the first attribute serves as a common prerequisite for all the $K-1$ remaining attributes. This $\mathcal{E}$ implies $\mathcal{A}(\mathcal{E}) = \{\mathbf{0}_{1\times K},\ (1, \boldsymbol{\alpha}') \text{ for } \boldsymbol{\alpha}' \in \{0,1\}^{K-1}\}$ with $2^{K-1}+1$ permissible patterns. To model this $\mathcal{E}$, a conventional attribute hierarchy method would require $2^{K-1}$ free parameters in the latent distribution. Such a lack of parsimony especially creates statistical and computational challenges when there are a large number of attributes but a limited sample size, as would be the case in fine-grained cognitive diagnosis of many skills in small classroom settings.

2.2. *Latent Conjunctive Bayesian Networks.* This subsection introduces a new family of models for attribute hierarchy in cognitive diagnosis: the Latent Conjunctive Bayesian Networks (LCBNs). LCBNs rigorously unify the attribute hierarchy method in educational measurement and the Bayesian network model in statistical machine learning, and inherit the advantages of both. Our proposal of LCBNs is inspired by another seemingly remote research area – graphical modeling of genetic mutations in bioinformatics. Specifically, the conjunctive Bayesian network (CBN) proposed by Beerenwinkel et al. (2005) and analyzed by Beerenwinkel et al. (2007), models a set of *observed* binary genetic mutations by a partial order, and assign zero probabilities to genotypes (analogue of our skill attribute patterns) that are not compatible with this partial order (analogue of our attribute hierarchy). An important difference is that, genetic mutations are often assumed to be entirely observed without any latent variables (Beerenwinkel et al., 2005, 2006, 2007). In contrast, in our cognitive diagnostic modeling of educational assessment data, the skill attributes are latent constructs that are not directly observable, but rather indirectly measured by item responses. We will further discuss the differences between the proposed LCBN and the CBN in Section 2.3, after elaborating on their common conjunctive modeling framework for multiple binary random variables.

We formally define the latent conjunctive Bayesian network for the attribute hierarchy. Introduce $K$ Bernoulli parameters $\boldsymbol{t} = (t_1, ..., t_K)^\top \in (0,1)^K$. For any $k \in [K]$, denote the set of "parent" attributes of $\alpha_k$ in the attribute hierarchy graph by $\text{pa}(k)$. The parent attribute of $\alpha_k$ here has the identical definition as the prerequisite attribute of $\alpha_k$. For example, for the attribute hierarchy shown in Figure 2, $\text{pa}(1) = \text{pa}(2) = \varnothing$ and $\text{pa}(3) = \text{pa}(4) = \{\alpha_1,\ \alpha_2\}$. Now define the probability mass function of the attribute pattern as follows:

$$\forall \boldsymbol{\alpha} \in \{0,1\}^K,\ \ p_{\boldsymbol{\alpha}} = \mathbb{P}(\boldsymbol{\alpha} \mid \boldsymbol{t}) = \prod_{k=1}^{K} \mathbb{P}(\alpha_k \mid \text{pa}(k)),\ \text{where} \tag{7}$$

$$\mathbb{P}(\alpha_k \mid \text{pa}(k)) = t_k^{\alpha_k \prod_{\ell=1}^K \alpha_\ell^{G_{l,k}}} (1-t_k)^{(1-\alpha_k)\prod_{\ell=1}^K \alpha_\ell^{G_{l,k}}}$$

$$= \begin{cases} t_k, & \text{if } \alpha_k = 1 \text{ and } \prod_{\ell \to k} \alpha_\ell = 1; \\ 1-t_k, & \text{if } \alpha_k = 0 \text{ and } \prod_{\ell \to k} \alpha_\ell = 1; \\ 0, & \text{if } \alpha_k = 1 \text{ and } \prod_{\ell \to k} \alpha_\ell = 0; \\ 1, & \text{if } \alpha_k = 0 \text{ and } \prod_{\ell \to k} \alpha_\ell = 0. \end{cases} \tag{8}$$

Eq. (7) follows the conventional definition of a Bayesian network (i.e., a probabilistic directed graphical model), where the joint distribution of random variables factorizes into the product of conditional distributions of each variable given its parents (Bishop, 2006). The conjunctive Bayesian network defined above has an intuitive and natural interpretation. This model states that a student can only master attribute $\alpha_k$ if he/she has already mastered every prerequisite attribute for $\alpha_k$; in this case, the mastery of $\alpha_k$ happens with probability

$$t_k = \mathbb{P}(\alpha_k = 1 \mid \alpha_\ell = 1 \text{ for all } \ell \in [K] \text{ such that } \ell \to k);$$

and $1 - t_k$ represents the probability of failing to master $\alpha_k$ given the student has already mastered all of its prerequisite attributes. The last line in (8) states that, if a student lacks some of $\alpha_k$'s prerequisite skills, then the probability of mastering $\alpha_k$ is zero. Therefore, this model respects the usual constraints on permissible/forbidden patterns as a conventional attribute hierarchy method. One can readily show that the model in (7)-(8) defines a valid joint distribution of attributes: $\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \mathcal{A}(\mathcal{E})} p_{\boldsymbol{\alpha}} = 1$ for any $\boldsymbol{t}$. The next example illustrates how the proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}})$ are parameterized by CBN parameters $\boldsymbol{t}$.

EXAMPLE 4 (Example 3 continued). *We revisit the attribute hierarchy in Example 3 and give it an LCBN parametrization. By (7), the proportion parameters for the permissible attribute patterns in $\mathcal{A}(\mathcal{E})$ in (6) can be written as*

$$p_{0000} = (1-t_1)(1-t_2), \quad p_{1000} = t_1(1-t_2), \quad p_{0100} = (1-t_1)t_2,$$
$$p_{1100} = t_1 t_2 (1-t_3)(1-t_4), \quad p_{1110} = t_1 t_2 t_3 (1-t_4),$$
$$p_{1101} = t_1 t_2 (1-t_3) t_4, \quad p_{1111} = t_1 t_2 t_3 t_4.$$

*For any $\boldsymbol{\alpha} \notin \mathcal{A}(\mathcal{E})$, $p_{\boldsymbol{\alpha}} = 0$ is naturally guaranteed by following the CBN definition. Note that if without the CBN assumption, the proportion parameters $p_{\boldsymbol{\alpha}}$ would be only subject to the sparsity constraint $p_{\boldsymbol{\alpha}} = 0$ for $\boldsymbol{\alpha} \notin \mathcal{A}(\mathcal{E})$; in this case, six free parameters would be needed to specify the latent distribution. In contrast, under the CBN, $\boldsymbol{\alpha}$ can be modeled using four Bernoulli parameters: $t_1, t_2, t_3, t_4$. In addition to such statistical parsimony, the LCBN model provides intuitive conditional independence statements about the skill attributes. In the current toy example, LCBN asserts that given a student's latent states of the first two basic skills, their mastery of the third and fourth skills are conditionally independent.*

Under our LCBN-based cognitive diagnostic model, the marginal distribution of the observed item response vector of the $i$th student takes the form:

$$\mathbb{P}(\mathbf{R}_i = \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \underbrace{t_k^{\alpha_k \prod_{\ell=1}^K \alpha_\ell^{G_{\ell,k}}} (1-t_k)^{(1-\alpha_k)\prod_{\ell=1}^K \alpha_\ell^{G_{\ell,k}}}}_{p_{\boldsymbol{\alpha}}} \prod_{j=1}^J \theta_{j,\boldsymbol{\alpha}}^{r_j} (1-\theta_{j,\boldsymbol{\alpha}})^{1-r_j}, \tag{9}$$

for any $\boldsymbol{r} \in \{0,1\}^J$. The hierarchy $\mathcal{E}$ implicitly appears above through the reachability matrix entries $G_{\ell,k}$. The item parameters $\theta_{j,\boldsymbol{\alpha}}$ in (9) are subject to the constraints imposed by the $Q$-matrix and can follow various measurement models described in Examples 1–2.

2.3. *Comparison of LCBNs and existing models.* We now discuss the difference between our LCBN-based cognitive diagnostic model and the CBN model for genetic mutations proposed by Beerenwinkel et al. (2005). In a CBN, each binary variable $\alpha_k = 1$ or $0$ represents a genetic event of whether an amino acid in the genome has mutated or not. There is a partial order (i.e., $\mathcal{E}$ in our notation) defined on the genetic events such that certain mutations are the prerequisite for others. Any patient's genetic mutation profile is fully observed as a binary vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})$, and the probability mass function of $\mathbf{X}_i$ is

$$\mathbb{P}(\mathbf{X}_i = \boldsymbol{\alpha} \mid \boldsymbol{t}, \mathcal{E}) = t_k{}^{\alpha_k} \textstyle\prod_{\ell=1}^{K} \alpha_\ell{}^{G_{\ell,k}} (1 - t_k)^{(1-\alpha_k)} \prod_{\ell=1}^{K} \alpha_\ell{}^{G_{\ell,k}}, \quad \forall \boldsymbol{\alpha} \in \{0,1\}^K.$$

In this fully observed CBN model, the hierarchy graph $\mathcal{E}$ can be directly read off from the set of all the observed binary patterns (*genotypes*) of the patients. In addition, Beerenwinkel et al. (2007) showed that the maximum likelihood estimator of parameters $\boldsymbol{t}$ in a CBN actually has a closed-form solution. In contrast, in our LCBN, students' $J$-dimensional item response vectors $\mathbf{R}_i$ in (9) do not readily reveal the attribute hierarchy graph $\mathcal{E}$ among the $K$ latent attributes; furthermore, the LCBN parameters $\boldsymbol{t}$ only enter the likelihood through those mixture proportion parameters $p_{\boldsymbol{\alpha}}$ in (9) rather than directly. Therefore, the identifiability issue of LCBNs is nontrivial, and the estimation of the attribute hierarchy and model parameters in LCBNs is not straightforward.

In terms of modeling the binary latent variables, LCBNs have the advantages of interpretability and statistical parsimony over conventional attribute hierarchy methods and conventional Bayesian networks. Comparing these two conventional models, the usual attribute hierarchy method has fewer parameters when the hierarchy graph $\mathcal{E}$ is dense with many arrows, whereas a Bayesian network without the conjunctive assumption (employed by Hu and Templin (2020) for cognitive diagnosis) has fewer parameters when the graph $\mathcal{E}$ is sparse. As concrete examples, consider the three different hierarchies in Figure 3 among $K = 7$ binary attributes. The numbers of free parameters needed to specify the distribution for the latent $\boldsymbol{\alpha}$ are shown in Table 1, from which it is clear that neither a conventional attribute hierarchy method nor a conventional Bayesian network is universally parsimonious. On the other hand, the number of parameters in LCBNs is $K$ for all hierarchies and is universally parsimonious.

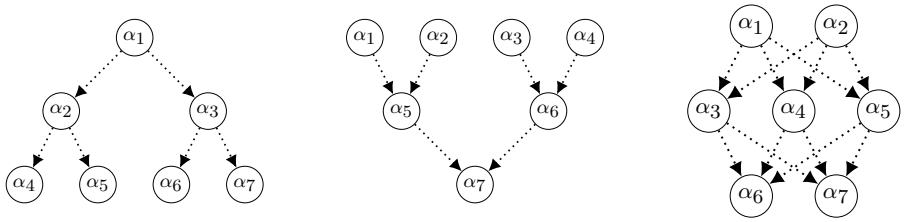


Fig 3: Different attribute hierarchies with $K = 7$ attributes. Divergent (left), convergent (middle), three-layer fully connected (right).

TABLE 1

*Number of free parameters needed for modeling the latent attributes in the conventional attribute hierarchy method (AHM), Bayesian network (BN), and LCBN with $K = 7$ attributes.*

| Hierarchy \ Model | AHM | BN | LCBN |
|---|---|---|---|
| Linear ($\mathcal{E} = \{1 \to 2 \to \cdots \to K\}$) | 7 | 13 | 7 |
| Divergent (left in Fig. 3) | 25 | 13 | 7 |
| Convergent (middle in Fig. 3) | 25 | 16 | 7 |
| 3-layer fully connected (right in Fig. 3) | 13 | 30 | 7 |
| No hierarchy | 127 | 7 | 7 |

A related model in the applied psychological measurement literature is the sequential higher order latent structural model for hierarchical attributes in Zhan et al. (2020). Specifically, motivated by the higher-order latent trait modeling in de la Torre and Douglas (2004) and the attribute hierarchy method, Zhan et al. (2020) proposed a conjunctive model with a higher-order continuous latent variable to model the attributes. It was assumed that every attribute $\alpha_k$ depends on the higher-order variable through an item response theory model. Our current work differs from this existing work in several fundamental ways. First, we do not assume the existence of any higher-order latent variables, which helps achieve the greatest amount of statistical parsimony. Only in this most parsimonious possible LCBN, the lack of arrows between the skills would encode nice conditional independence interpretation; in Zhan et al. (2020)'s higher-order model, all the skills are always conditionally dependent due to the higher-order latent trait. Second, we establish identifiability for the family of LCBN-based cognitive diagnostic models (see Section 3) and propose a general two-step method to perform both structure learning of $\mathcal{E}$ and parameter estimation of $(\boldsymbol{\Theta}, \boldsymbol{t})$ (see Section 4). In previous studies such as Zhan et al. (2020), identifiability issues were not examined and estimation was performed by assuming the hierarchy $\mathcal{E}$ is known.

**3. Identifiability of LCBNs for Cognitive Diagnosis.** Identifiability is a fundamental property of statistical models and a prerequisite for valid parameter estimation and hypothesis testing. If a model is not identifiable, then there exist multiple and possibly an infinite number of parameter sets that lead to the same observed distribution, and it is impossible to distinguish them. In the applied context of using LCBNs for cognitive diagnosis, it is crucial to guarantee that the model is identifiable, so that any practical interpretation made about the cognitive structure and student diagnosis is statistically valid. In this section, we provide transparent conditions for LCBNs to be identifiable.

Because the DINA model in Example 1 is the most popular and fundamental cognitive diagnostic model due to its interpretability and parsimony, we next focus on the LCBN-based DINA model and provide tight and explicit identifiability conditions for it. We remark that LCBN-based CDMs with other measurement models (such as main-effect and all-effect CDMs) are also identifiable under slightly stronger conditions. In light of the space constraint and for notational simplicity, we defer those identifiability results to Section S.1. in the Supplementary Material.

As mentioned earlier, the identifiability of LCBN-based cognitive diagnostic models is a nontrivial and challenging issue, unlike the fully observed CBNs. Fortunately, thanks to our model formulation, the LCBN parameters $\boldsymbol{t}$ enter the observed distribution in (9) only through the mixture proportion parameters $p_{\boldsymbol{\alpha}}$. Therefore, we are able to leverage existing techniques for conventional CDMs with an unstructured attribute hierarchy model in Gu and Xu (2023a) to establish identifiability for LCBNs. Specifically, we next provide conditions that ensure the identifiability of not only the continuous parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{t})$, but also the discrete hierarchy graph structure $\mathcal{E}$ in an LCBN. We first define the concept of strict identifiability of the LCBN-based DINA model.

DEFINITION 1 (Strict identifiability for LCBN-based DINA). *The parameters $(\mathcal{E}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{t})$ of an LCBN-based DINA model are identifiable if for any $(\mathcal{E}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{t})$ and $(\bar{\mathcal{E}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{t}})$ where $\bar{\mathcal{E}}$ induces at most $|\mathcal{A}(\mathcal{E})|$ permisible skill patterns, the following holds if and only if $(\bar{\mathcal{E}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{t}}) = (\mathcal{E}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{t})$ holds.*

$$\mathbb{P}(\mathbf{R} = r \mid \bar{\mathcal{E}}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{t}}) = \mathbb{P}(\mathbf{R} = r \mid \mathcal{E}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{t}) \text{ for all } r \in \{0,1\}^J. \tag{10}$$

We introduce some new notation before presenting the identifiability results. Following the definition in Gu and Xu (2023a), we categorize the latent attributes into four different types:

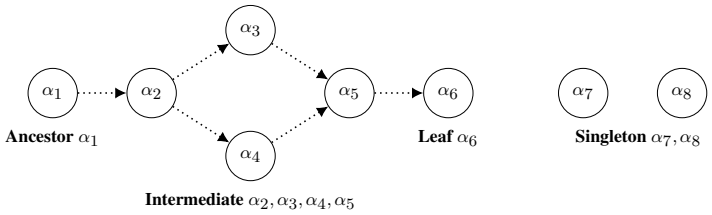


Fig 4: Illustrating all the four types of attributes in an attribute hierarchy graph.

ancestor, intermediate, leaf, and singleton. An attribute is an "ancestor attribute" when it has a child but no parent attribute; an "intermediate attribute" when it has both a child and a parent; a "leaf attribute" when it has a parent but no child attribute; a "singleton attribute" when it has no child nor parent attribute. These definitions are illustrated in Figure 4. Interestingly, the identifiability conditions of LCBN-based DINA model can be stated in terms of different types of the attributes in the hierarchy graph. Still following the definition in Gu and Xu (2023a), we define a "sparsified" $Q$-matrix given $\mathcal{E}$ by setting $q_{j,k} = 0$ for any $j, k$ such that $q_{j,h} = 1$ for some child attribute $\alpha_h$ of $\alpha_k$.

THEOREM 1. *The LCBN-based DINA model is strictly identifiable when the* $\mathbf{Q}$ *and* $\mathcal{E}$ *satisfy the following conditions.*

*A.* $\mathbf{Q}$ *contains a* $K \times K$ *submatrix* $\mathbf{Q}_0$ *whose sparsified version under* $\mathcal{E}$ *is* $I_K$*. Without the loss of generality, write* $\mathbf{Q} = [\mathbf{Q}_0^\top, \mathbf{Q}^{*\top}]^\top$.

*B. In the sparsified version of* $\mathbf{Q}$*, any intermediate attribute is measured at least once, any ancestor or leaf attribute is measured at least twice, and any singleton attribute is measured at least three times.*

*C. For any singleton attributes* $\alpha_k$ *and* $\alpha_\ell$*, the* $k$*th and* $l$*th columns of* $\mathbf{Q}^*$ *are different.*

Theorem 1 is adapted from Theorem 2 in Gu and Xu (2023a) to our LCBN-based model setting. In general, it is difficult to derive the necessary and sufficient conditions for identifiability of complicated models such as LCBN-based CDMs. Nevertheless, we next show our sufficient identifiability conditions in Theorem 1 may not be far from being necessary by considering a special hierarchy. The next proposition states that our conditions $A, B, C$ in Theorem 1 become the minimal requirement for identifiability under the linear hierarchy.

PROPOSITION 1. *Suppose* $\mathcal{E}$ *is a linear hiearchy, i.e.* $\alpha_1 \to \alpha_2 \to \cdots \to \alpha_K$*. Then, the conditions in Theorem 1 are necessary and sufficient for strict identifiability of an LCBN-based DINA model. In particular, conditions* $B$ *and* $C$ *reduce exactly to be:*

*B*$^\star$*. In the sparsified version of* $\mathbf{Q}$*, the leaf attribute and the ancestor attribute are each measured at least twice.*

The proofs of Theorem 1 and Proposition 1, and additional identifiability results (sufficient conditions for strict and generic identifiability for LCBNs with other measurement models) are included in Sections S.1 and S.2 in the Supplementary Material.

**4. Two-step Estimation Method for LCBN-based Cognitive Diagnostic Models.** This section proposes a two-step estimation method to recover both the attribute hierarchy graph $\mathcal{E}$ and the model parameters $(\boldsymbol{\Theta}, \boldsymbol{t})$. Our first step (Algorithm 1) uses a penalized EM algorithm under a saturated attribute model to estimate the graph $\mathcal{E}$, and our second step (Algorithm 2) develops another EM algorithm to estimate the continuous LCBN parameters.

We first write out the likelihood given the responses from a sample of $N$ students. Denote the response vectors for the $N$ students by $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,J})^\top$, for $i = 1, \ldots, N$. The marginal likelihood under an LCBN-based cognitive diagnostic model is

(11)
$$L(\boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E}) = \prod_{i=1}^{N} \Big[ \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} t_k{}^{\alpha_k \prod_{\ell=1}^K \alpha_\ell^{G_{\ell,k}}} (1 - t_k)^{(1-\alpha_k)\prod_{\ell=1}^K \alpha_\ell^{G_{\ell,k}}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big]$$
$$= \prod_{i=1}^{N} \Big[ \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big] =: L(\boldsymbol{\Theta}, \boldsymbol{p}),$$

where the last line uses the equivalent parameterization of the mixture proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}}; \boldsymbol{\alpha} \in \{0,1\}^K)$ instead of the LCBN parameters $\boldsymbol{t}$. Write the marginal log-likelihood as $\ell(\boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E}) = \log L(\boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E})$ and $\ell(\boldsymbol{\Theta}, \boldsymbol{p}) = \log L(\boldsymbol{\Theta}, \boldsymbol{p})$. We next describe the two steps of the proposed estimation procedure in Sections 4.1 and 4.2, respectively.

4.1. *First step: structure learning of $\mathcal{E}$ via a penalized EM algorithm.* In the first step, we focus on estimating the discrete graph structure in an LCBN: the attribute hierarchy $\mathcal{E}$. Estimating $\mathcal{E}$ amounts to performing structure learning of a directed graphical model, and this graphical model is among the $K$ latent skills.

The key idea in learning $\mathcal{E}$ in an LCBN is to realize that, as an attribute hierarchy $\mathcal{E}$ naturally defines a set of permissible binary skill patterns $\mathcal{A} = \mathcal{A}(\mathcal{E})$, a set of permissible patterns $\mathcal{A}$ also allows for reconstructing an attribute hierarchy graph $\mathcal{E} = \mathcal{E}(\mathcal{A})$. Specifically, one can inversely infer $\mathcal{E}$ by examining the sparsity structure of $\boldsymbol{p}$. For a set of permissible patterns $\mathcal{A} \subseteq \{0,1\}^K$, we can read that $\alpha_k$ is a prerequisite for $\alpha_\ell$ if for any permissible pattern $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathcal{A}$, we have $\alpha_\ell = 1$ holds only if $\alpha_k = 1$ holds. In this way, we can define an attribute hierarchy graph $\mathcal{E}$ by collecting these prerequisite relationships:

(12) $$\mathcal{E} = \mathcal{E}(\mathcal{A}) = \{k \to \ell : \text{ if for any } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathcal{A}, \ \alpha_\ell = 1 \text{ only if } \alpha_k = 1\}.$$

EXAMPLE 5 (Example 3 continued). *We revisit the attribute hierarchy $\mathcal{E} = \{1 \to 3,\ 1 \to 4,\ 2 \to 3,\ 2 \to 4\}$ in Example 3 and show it can be recovered from the set of permissible patterns $\mathcal{A}$. First, we rewrite $\mathcal{A}$ into a $|\mathcal{A}| \times K$ matrix denoted by $\mathbf{C}$. Each row of $\mathbf{C}$ corresponds to one pattern $\boldsymbol{\alpha} \in \mathcal{A}$ and each column corresponds to a skill. Then we compare the column vectors of $\mathbf{C}$ to obtain a partial order among the skills. For example, if $\mathbf{C}_{:,1} \succeq \mathbf{C}_{:,3}$ (the first column of $\mathbf{C}$ is elementwisely greater than or equal to the third column of $\mathbf{C}$), attribute $\alpha_3$ is present only if attribute $\alpha_1$ is present; this indicates $1 \to 3$. In the current toy example, such a procedure gives the following reconstruction of the hierarchy $\mathcal{E}$.*

$$\mathbf{C} = \begin{pmatrix} 0\,0\,0\,0 \\ 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 1\,1\,0\,0 \\ 1\,1\,1\,0 \\ 1\,1\,0\,1 \\ 1\,1\,1\,1 \end{pmatrix} \xrightarrow{\text{get a partial order between columns}} \begin{matrix} \mathbf{C}_{:,1} \succeq \mathbf{C}_{:,3} \\ \mathbf{C}_{:,1} \succeq \mathbf{C}_{:,4} \\ \mathbf{C}_{:,2} \succeq \mathbf{C}_{:,3} \\ \mathbf{C}_{:,2} \succeq \mathbf{C}_{:,4} \end{matrix} \implies \mathcal{E} = \begin{Bmatrix} 1 \to 3, \\ 1 \to 4, \\ 2 \to 3, \\ 2 \to 4. \end{Bmatrix}$$

To estimate $\mathcal{E}$, now the problem boils down to estimating $\mathcal{A}$. To this end, we leverage the log penalty and penalized EM algorithm proposed in Gu and Xu (2019) for selecting significant latent patterns. Consider the truncated log function

$$\log_{\rho_N}(p_{\boldsymbol{\alpha}}) = \log(p_{\boldsymbol{\alpha}}) \cdot \mathbb{1}(p_{\boldsymbol{\alpha}} > \rho_N) + \log(\rho_N) \cdot \mathbb{1}(p_{\boldsymbol{\alpha}} \le \rho_N),$$

where $\rho_N$ is a small threshold that avoids the singularity issue of the log function at zero. The penalized marginal log likelihood $\ell^\lambda(\boldsymbol{\Theta}, \boldsymbol{p})$ is defined as

$$\ell^\lambda(\boldsymbol{\Theta}, \boldsymbol{p}) = \ell(\boldsymbol{\Theta}, \boldsymbol{p}) + \lambda \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \log_{\rho_N}(p_{\boldsymbol{\alpha}}), \tag{13}$$

where $\lambda < 0$ is a tuning parameter controlling the sparsity of $\boldsymbol{p}$. We maximize $\ell^\lambda(\boldsymbol{\Theta}, \boldsymbol{p})$ instead of the original marginal log likelihood $\ell(\boldsymbol{\Theta}, \boldsymbol{p})$ using the Penalzed EM (PEM) algorithm in Gu and Xu (2019). We restate this algorithm in Algorithm 1. A smaller tuning parameter $\lambda$ (i.e. larger $-\lambda = |\lambda| > 0$) leads to a stronger penalty and encourages a sparser $\boldsymbol{p}$.

REMARK 1. *One main reason for choosing the log penalty on the proportion parameters $\boldsymbol{p}$ over other popular sparsity-inducing penalties is computational convenience. Among sparsity-inducing penalties, the $L_0$ penalty is the most direct one that penalizes the number of nonzero entries. Although the $L_0$ penalty encourages sparsity and theoretically leads to consistent selection, it is computationally inefficient due to its discontinuous and nonconvex nature. There exist various attempts to replace the $L_0$ penalty with a similar but more tractable objective. One such example is the popular $L_1$ (Lasso, Tibshirani, 1996) penalty. But actually, Lasso turns out to not induce any sparsity on our proportion parameters $\boldsymbol{p}$ because $\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} |p_{\boldsymbol{\alpha}}| = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$ for any $\boldsymbol{p}$.*

*Compared to the aforementioned penalties, the log penalty proposed by Gu and Xu (2019) is preferable as it not only induces nice sparsity on $\boldsymbol{p}$, but also allows for efficient and explicit M-step updates for $\boldsymbol{p}$ in an EM algorithm. This follows from the fact that the log penalty can be alternatively viewed as a Dirichlet prior for $\boldsymbol{p}$, which is a conjugate prior for the complete data log likelihood. For more discussions on the connection between the log penalty and the Dirichlet prior in a Bayesian context, please see Remark 12 in Gu and Xu (2019).*

We denote the estimator of the item parameters by $\boldsymbol{\Theta}^\lambda$ and that of the mixture proportion parameters by $\boldsymbol{p}^\lambda = (p^\lambda_{\boldsymbol{\alpha}};\ \boldsymbol{\alpha} \in \{0,1\}^K)$. Further, we define the following estimated set of existing skill patterns: $\mathcal{A}^\lambda = \{\boldsymbol{\alpha} \in \{0,1\}^K :\ p^\lambda_{\boldsymbol{\alpha}} > \rho_N\}$; that is, $\mathcal{A}^\lambda$ collects those skill patterns with estimated proportions greater than the threshold $\rho_N$. This $\mathcal{A}^\lambda$ is the key quantity that would give an estimate of the attribute hierarchy $\mathcal{E}^\lambda$.

We consider a sequence of values for $\lambda$ and select an optimal tuning parameter $\widehat{\lambda}$ based on the Extended Bayesian Information Criterion (EBIC, Chen and Chen, 2008), that is $\widehat{\lambda} = \arg\min_\lambda \text{EBIC}_\lambda$. The exact definition of EBIC is given in Supplementary Material S.3.1. Compared to BIC, EBIC has an additional penalty term for the number of selected parameters and hence favors a more parsimonious model. EBIC has been used in related existing works (Gu and Xu, 2019; Ma et al., 2023), and it also turns out to be especially useful for estimating $\mathcal{E}$ in LCBNs. In fact, our simulations suggest that the stronger penalty in EBIC is desirable because overselecting non-existing patterns often leads to error in estimating the graph $\mathcal{E}$, whereas underselecting truly existing patterns can sometimes still suffice for correctly estimating $\mathcal{E}$ (see Section 5). In addition, other popular criteria for model selection such as cross-validation is not suitable for selecting $\lambda$ here, because it does not take the model sparsity into account. In fact, in simulation studies in Supplementary Material S.4.2, we show that cross-validation tends to select a larger $\lambda < 0$ with a smaller magnitude than needed, hence resulting in selecting a non-sparse model. Finally, given the estimated set $\mathcal{A}^{\widehat{\lambda}}$, define our estimate of the attribute hierarchy graph $\mathcal{E}$ following (12):

$$\widehat{\mathcal{E}} = \{k \to \ell :\ \text{if for any } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathcal{A}^{\widehat{\lambda}},\ \ \alpha_\ell = 1 \text{ only if } \alpha_k = 1\}.$$

Next, we show that our estimator $\widehat{\mathcal{E}}$ is statistically consistent under suitable conditions. We consider the conventional asymptotic setting where the sample size $N$ goes to infinity, but

the number of skills $K$ and the number of items $J$ are fixed. Following the assumption in Gu and Xu (2019), we also assume that the convergence rate of the MLE satisfies

$$(\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) - \ell(\widehat{\boldsymbol{\Theta}}^{\mathcal{E}}, \widehat{\boldsymbol{p}}^{\mathcal{E}}))/N = O_p(N^{-\delta}) \tag{14}$$

for some $\delta \in (0, 1]$. Here, $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ is the MLE obtained from maximizing $L(\boldsymbol{\Theta}, \boldsymbol{p})$ in (11) and $(\widehat{\boldsymbol{\Theta}}^{\mathcal{E}}, \widehat{\boldsymbol{p}}^{\mathcal{E}})$ is the oracle MLE assuming that the true hierarchy $\mathcal{E}$ is known. Similar to Gu and Xu (2019), we impose this assumption because the convergence rate of the MLE with an unknown hierarchy (or equivalently, an unknown number of mixture components $|\mathcal{A}|$) can be slower than the usual parametric rate with $\delta = 1$ (Ho and Nguyen, 2016). The following theorem shows the consistency conclusion.

THEOREM 2. *Consider an identifiable LCBN-based CDM with parameters* $(\boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E})$. *Suppose the item parameter* $\boldsymbol{\Theta}$ *satisfies*

$$\theta_{j,\mathbf{1}} - \max_{\boldsymbol{\alpha} \not\supseteq \boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} \geq c > 0, \quad \forall j,$$

$\rho_N = O(N^{-\delta})$, *and* (14) *holds. Then, for any sequence* $\{\lambda_N\}$ *satisfying* $\frac{N^{1-\delta}}{|\log \rho_N|} \lesssim -\lambda_N \lesssim \frac{N}{|\log \rho_N|}$, $\mathbb{P}(\widehat{\mathcal{E}}^{\lambda_N} = \mathcal{E}) \to 1$ *as* $N \to \infty$. *Here,* $\widehat{\mathcal{E}}^{\lambda_N}$ *is the estimated hierarchy based on* $\mathcal{A}^{\lambda_N}$.

Theorem 2 also provides theoretical guidelines on choosing the tuning parameters. In particular, we choose $\rho_N = \frac{1}{2N}$ so that it satisfies the condition in Theorem 2 for any $\delta$.

In addition to the above estimation consistency result for the attribute hierarchy, one could further quantify uncertainty via formal hypothesis testing. Specifically, we can consider testing the null hypothesis $H_0 : \mathcal{E} = \widehat{\mathcal{E}}$ using additional response data, where $\widehat{\mathcal{E}}$ is the estimated attribute hierarchy. To this end, one may conduct standard goodness of fit tests such as the likelihood ratio test with a $\chi^2$ asymptotic reference distribution. We leave the detailed development of such hypothesis testing procedures for future research.

4.2. *Second step: parameter estimation of* $(\boldsymbol{\Theta}, \boldsymbol{p})$ *via another EM algorithm.* We next propose another EM algorithm to estimate the continuous LCBN parameters $\boldsymbol{t}$ and $\boldsymbol{\Theta}$. The previous Algorithm 1 does not take into account the LCBN structure, but merely focuses on estimating which skill patterns have nonzero proportions in the student population. Importantly, note that although the hierarchy graph $\widehat{\mathcal{E}}$ can be read off from the sparsity structure of $\widehat{\boldsymbol{p}}^{\lambda}$, the LCBN parameters $\boldsymbol{t}$ *cannot* be read off from the estimated proportion parameters $\boldsymbol{p}$. This is because the latter is an overparametrization of the former, and it is not guaranteed that a freely estimated $\boldsymbol{p}$ will correspond to a $K$-dimensional LCBN parameters $\boldsymbol{t} = (t_1, \ldots, t_K)$.

We next propose another EM algorithm to re-estimate the continuous parameters in LCBN-based cognitive diagnostic models given $\widehat{\mathcal{E}}$. For each individual $i = 1, ..., N$, denote their latent skill profile by $\mathbf{A}_i = (A_{i,1}, ..., A_{i,K})$. We maximize the likelihood in (11) with respect to $(\boldsymbol{t}, \boldsymbol{\Theta})$ when holding $\mathcal{E} = \widehat{\mathcal{E}}$ fixed. The log likelihood for the complete data $(\mathbf{A}_i, \mathbf{R}_i)$, $i = 1, \ldots, N$ takes the following form:

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}, \boldsymbol{t} \mid \mathcal{E}) = & \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{N} \mathbb{1}(\mathbf{A}_i = \boldsymbol{\alpha}) \prod_{k=1}^{K} \left[ {t_k}^{\alpha_k \prod_{\ell=1}^{K} \alpha_\ell^{G_{\ell,k}}} (1 - t_k)^{(1-\alpha_k) \prod_{\ell=1}^{K} \alpha_\ell^{G_{\ell,k}}} \right] \\ & + \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{N} \mathbb{1}(\mathbf{A}_i = \boldsymbol{\alpha}) \sum_{j=1}^{J} \Big[ R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}}) \Big]. \end{aligned} \tag{15}$$

**Algorithm 1:** Penalized EM to learn the attribute hierarchy graph $\mathcal{E}$ (Algorithm 1 in Gu and Xu (2019))

**Data:** $Q$-matrix $\mathbf{Q} = (q_{j,k})$, response vectors $(\mathbf{R}_1^\top, \ldots, \mathbf{R}_N^\top)^\top$.

Initialize $\boldsymbol{\Delta} = (\Delta_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$ from the $(2^K - 1)$-dimensional probability simplex.

**while** *not converged* **do**

In the $(t+1)$th iteration,

**for** $(i, \boldsymbol{\alpha}) \in [N] \times \{0,1\}^K$ **do**

$$\varphi_{i,\boldsymbol{\alpha}}^{(t+1)} = \frac{\delta_{\boldsymbol{\alpha}}^{(t)} \cdot \exp\left\{\sum_{j=1}^J \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}}^{(t)}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}}^{(t)})\right]\right\}}{\sum_{\boldsymbol{\alpha}' \in \{0,1\}^K} \delta_{\boldsymbol{\alpha}'}^{(t)} \cdot \exp\left\{\sum_{j=1}^J \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}'}^{(t)}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}'}^{(t)})\right]\right\}};$$

**for** $\boldsymbol{\alpha} \in \{0,1\}^K$ **do**

$\delta_{\boldsymbol{\alpha}}^{(t+1)} = \max\{c,\ \lambda + \sum_{i=1}^N \varphi_{i,\boldsymbol{\alpha}}^{(t+1)}\}$; ($c > 0$ is a pre-specified small constant, set to $c = 0.01$ throughout the experiments following the suggestion of Gu and Xu (2019));

$\boldsymbol{p}^{(t+1)} \leftarrow \boldsymbol{\delta}^{(t+1)} / \left(\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \delta_{\boldsymbol{\alpha}}^{(t+1)}\right)$;

**for** $j \in [J]$ **do**

$$\boldsymbol{\Theta}_j^{(t+1)} = \arg\max_{\boldsymbol{\Theta}_j} \left\{\sum_{\boldsymbol{\alpha}} \sum_i \varphi_{i,\boldsymbol{\alpha}}^{(t+1)} \sum_j \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}}^{(t)}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}}^{(t)})\right]\right\};$$

After convergence, use $\mathcal{A}^\lambda, \boldsymbol{\Theta}^\lambda, \boldsymbol{p}^\lambda$ to calculate the EBIC for a sequence of $\lambda < 0$.

Select $\widehat{\lambda}$ with the minimum EBIC and recover the hierarchy structure $\mathcal{E}^{\widehat{\lambda}}$ from $\mathcal{A}^{\widehat{\lambda}}$.

**Output:** Attribute hierarchy $\mathcal{E}$.

Recall that the prerequisite relationships in $\mathcal{E}$ completely define the reachability matrix entries $G_{\ell,k} = \mathbb{1}(\ell \to k)$ in the above expression. So the only things that vary in (15) are $(\boldsymbol{\Theta}, \boldsymbol{t})$.

In the E-step, we evaluate the conditional expectation of (15) given the current parameter values $\boldsymbol{\Theta}^{(t)}$ and $\boldsymbol{t}^{(t)}$ from the previous iteration. It suffices to evaluate the conditional probability of $\mathbb{1}(\mathbf{A}_i = \boldsymbol{\alpha})$, denoted by $\varphi_{i,\boldsymbol{\alpha}} = \mathbb{P}(\mathbf{A}_i = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{t}^{(t)})$. See the detailed formula for $\varphi_{i,\boldsymbol{\alpha}}$ in Algorithm 2. Then we obtain the following function of $(\boldsymbol{\Theta}, \boldsymbol{t})$:

$$Q(\boldsymbol{\Theta}, \boldsymbol{t} \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{t}^{(t)}) = \mathbb{E}\Big[\ell_{\mathrm{c}}(\boldsymbol{\Theta}, \boldsymbol{t} \mid \mathcal{E}) \;\Big|\; \boldsymbol{\Theta}^{(t)}, \boldsymbol{t}^{(t)}\Big].$$

Next, in the M-step, we seek the maximizers of the above function and obtain new estimates of the model parameters:

$$(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{t}^{(t+1)}) = \arg\max_{\boldsymbol{\Theta}, \boldsymbol{t}} Q(\boldsymbol{\Theta}, \boldsymbol{t} \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{t}^{(t)}). \tag{16}$$

Every parameter in $(\boldsymbol{\Theta}, \boldsymbol{t})$ is continuous, so we set the partial derivative with respect to each of them to zero to seek $(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{t}^{(t+1)})$. We present detailed derivations of Algorithm 2 and closed form updates for item parameters in Section S.3.2 of the Supplementary Material.

Next, we show that our two-stage estimation method based on Algorithms 1 and 2 can consistently estimate both the attribute hierarchy graph and the continuous parameters.

THEOREM 3. *Consider an identifiable LCBN-based CDM with parameters $(\boldsymbol{\Theta}, \boldsymbol{t}, \mathcal{E})$, and suppose that the conditions in Theorem 2 hold. Let $\widehat{\mathcal{E}}$ be the hierarchy estimated from Algorithm 1, and let $(\widehat{\boldsymbol{\Theta}}_N, \widehat{\boldsymbol{t}}_N)$ be the maximum likelihood estimator of $(\boldsymbol{\Theta}, \boldsymbol{t})$ given $\widehat{\mathcal{E}}$. Then, entries in $(\widehat{\boldsymbol{\Theta}}_N, \widehat{\boldsymbol{t}}_N)$ converge to corresponding entries in $(\boldsymbol{\Theta}, \boldsymbol{t})$ in probability as $N \to \infty$.*

4.3. *Estimation under unknown Q.* In the previous subsections, we have focused on estimating the LCBN parameters assuming a known and fixed $Q$-matrix. This is a common

**Algorithm 2:** EM to estimate LCBN parameters.

**Data:** $Q$-matrix $\mathbf{Q}$, response patterns $\{\mathbf{R}_i : i = 1, \ldots, N\}$, attribute hierarchy $\mathcal{E}$.
Initialize $\boldsymbol{t} = (t_1, ..., t_K)$, $\boldsymbol{\Theta}$ (subject to the constraints of the $Q$-matrix), and $\mathbf{G}$.
**while** *not converged* **do**
In the $(t+1)$th iteration:
**for** $\boldsymbol{\alpha} \in \mathcal{A}$ **do**

$$p_{\boldsymbol{\alpha}}^{(t+1)} = \prod_{k=1}^{K} \left(t_k^{(t)}\right)^{\alpha_k \prod_\ell \alpha_\ell^{G_{\ell,k}}} \left(1 - t_k^{(t)}\right)^{(1-\alpha_k)\prod_\ell \alpha_l^{G_{\ell,k}}};$$

**for** $(i, \boldsymbol{\alpha}) \in [N] \times \mathcal{A}(\mathcal{E})$ **do**

$$\varphi_{i,\boldsymbol{\alpha}}^{(t+1)} = \frac{p_{\boldsymbol{\alpha}}^{(t)} \cdot \exp\left\{\sum_{j=1}^{J}\left[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}}^{(t)}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}}^{(t)})\right]\right\}}{\sum_{\boldsymbol{\alpha}' \in \mathcal{A}(\mathcal{E})} p_{\boldsymbol{\alpha}'}^{(t)} \cdot \exp\left\{\sum_{j=1}^{J}\left[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}'}^{(t)}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}'}^{(t)})\right]\right\}};$$

**for** $k \in [K]$ **do**

$$t_k^{(t+1)} = \frac{\sum_{i,\boldsymbol{\alpha}} \alpha_k \prod_{\ell=1}^{K} \alpha_\ell^{G_{\ell,k}} \varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_{i,\boldsymbol{\alpha}} \prod_{\ell=1}^{K} \alpha_\ell^{G_{\ell,k}} \varphi_{i,\boldsymbol{\alpha}}^{(t+1)}};$$

**for** $j \in [J]$ **do**

$$\boldsymbol{\Theta}_j^{(t+1)} = \arg\max_{\boldsymbol{\Theta}_j} \left\{\sum_{\boldsymbol{\alpha}} \sum_i \varphi_{i,\boldsymbol{\alpha}}^{(t+1)} \sum_j \left[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}}^{(t)}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}}^{(t)})\right]\right\};$$

After the total $T$ iterations,
**Output:** Estimated parameters $\boldsymbol{t}, \boldsymbol{\Theta}$.

assumption in cognitive diagnostic assessments, because domain experts and test designers often have specified how the test items depend on the latent attributes. But sometimes it is of interest to estimate the $\mathbf{Q}$-matrix directly from data together with other model parameters. Our two-step estimation procedure for LCBNs can be readily extended to such unknown $Q$-matrix settings by leveraging existing $\mathbf{Q}$-matrix estimation methods for traditional CDMs. We next briefly describe how the exploratory estimation method in Ma et al. (2023) can be incorporated into our LCBN estimation procedure with an unknown $K$, $\mathbf{Q}$, and $\mathcal{E}$.

We briefly sketch the method proposed by Ma et al. (2023) in Algorithm 3. This algorithm includes an additional truncated Lasso penalty (TLP; Shen et al., 2012) term on the item parameter matrix $\boldsymbol{\Theta}$ to encourage row-wise sparsity, and consequently recover the $Q$-matrix. The attribute hierarchy $\mathcal{E}$ is estimated by comparing the partial orders of the columns of $\boldsymbol{\Theta}$, and assigning binary representations to these columns as attribute patterns. This Algorithm 3 can serve as our new first step in the two-step estimation procedure. Given the estimated $\mathbf{Q}$-matrix and $\mathcal{E}$, we can then apply our proposed Algorithm 2 to estimate the continuous LCBN parameters: $\boldsymbol{t}$ and $\boldsymbol{\Theta}$. We present simulation study results in Supplementary Material S.4.3 that demonstrate the good performance of the above estimation method.

**5. Simulation Studies.** In this section, we conduct simulation studies under different models and parameter settings to assess the performance of our proposed method. We consider the LCBN-based DINA and GDINA models (see Examples 1 and 2 for their definition)

**Algorithm 3:** Estimate $K$ and discrete structures $\mathbf{Q}$ and $\mathcal{E}$
(Brief sketch of Algorithms 1 and 2 in Ma et al. (2023))

**Data:** Responses $(\mathbf{R}_1^\top, \ldots, \mathbf{R}_N^\top)^\top$.
Set an upper bound for $|\mathcal{A}|$, the number of latent configurations
**Step 1:** Use penalized EM assuming sparsity of $\boldsymbol{p}$ and $\boldsymbol{\Theta}$ to estimate $\boldsymbol{\Theta}$ and $|\mathcal{A}|$
**Step 2:** Construct the $J \times |\mathcal{A}|$ indicator matrix $\Gamma = \mathbb{1}(\theta_{j,m} = \max_{l \in [|\mathcal{A}|]} \theta_{j,l})$
**Step 3:** Plot a DAG based on the partial orders of the columns of $\Gamma$
**Step 4:** Assign binary representations bassed on this DAG and recover $K$ and $\mathcal{E}$
**Step 5:** Reconstruct each rows of $\mathbf{Q}$ based on the corresponding row of the $\Gamma$ matrix
**Output:** Number of latent attributes $K$, Hierarchy structure $\mathcal{E}$, $Q$-matrix $\mathbf{Q}$

with $K = 8$ latent attributes and $J = 24$ items. The $Q$-matrix takes the following form:

$$\text{(17)} \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \\ \mathbf{I}_K \end{pmatrix}, \quad \text{where } \mathbf{Q}_1 = \begin{pmatrix} 1 & 1 & & & 0 \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & 1 & \\ 0 & & 1 & 1 & \end{pmatrix}_{K \times K} \text{ and } \mathbf{Q}_2 = \begin{pmatrix} 1 & 1 & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \\ 0 & & & 1 & \end{pmatrix}_{K \times K}.$$

Note that $K = 8$ is already a relatively large number of attributes in the educational cognitive diagnosis applications. In all of our simulations, we specify the true hierarchy $\mathcal{E}$ to be the diamond hierarchy defined in Figure 5. This is a complex multi-layer hierarchy which encodes $|\mathcal{A}(\mathcal{E})| = 15$ permissible patterns. We set the true LCBN parameters as $\boldsymbol{t} = (0.9, 0.8, 0.8, 0.7, 0.7, 0.7, 0.6, 0.6)^\top$. Following (7), we can obtain the mixture proportion parameters for the permissible skill patterns. All the permissible patterns indexed by $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{15}$ and their corresponding true proportion parameters are presented in Table 2.

We vary the following three aspects in the simulation studies: (1) *measurement model*: DINA and GDINA; (2) *sample size* $N = 500, 1000, 2000$; and (3) *noise level of item parameters*. In the LCBN-based DINA model, we use a noise level $r$ to define the slipping and guessing parameters $\boldsymbol{s}, \boldsymbol{g}$ by $s_j = g_j = r$ for all $j = 1, \ldots, J$. The larger the noise level $r$ is, the more challenging it is to estimate the model parameters. Specifically, under DINA, if $r = 0$ then there is no uncertainty in one's responses given their latent skills, whereas if $r = 0.5$ the responses are purely random noise. For the LCBN-based GDINA model, we set $\theta_{j,\mathbf{0}_K} = r$ and $\theta_{j,\mathbf{1}_K} = 1 - r$, and define the remaining item parameters by setting all main effects and interaction effects of the required attributes in (4) to be equal.

In each simulation setting, we run 100 independent simulation replications. We apply our two-step estimation method described in Section 4. The tuning parameter $\lambda$ in Algorithm 1 (PEM algorithm) is selected from a grid of ten values $\lambda \in \{-0.4, -0.8, \ldots, -3.6, -4.0\}$. We evaluate the root mean squared errors (RMSE) of the continuous parameters and also the estimation accuracy of the permissible patterns in $\mathcal{A}$ (this is same as the estimation accuracy of the hierarchy $\mathcal{E}$). The RMSE of the proportion parameters $\widehat{\boldsymbol{p}}$ is computed using the $2^K$–dimensional sparse vector in the probability simplex, i.e. for $C = 100$ simulations,

$$\text{RMSE}(\widehat{\boldsymbol{p}}) = \sqrt{\frac{1}{2^K C} \sum_{c=1}^{C} \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} (\widehat{p}_{\boldsymbol{\alpha}}^{(c)} - p_{\boldsymbol{\alpha}})^2},$$

where $\widehat{\boldsymbol{p}}^{(c)} = (\widehat{p}_{\boldsymbol{\alpha}}^{(c)})$ denotes the estimator from the $c$th simulation replicate. We sum over all $\boldsymbol{\alpha} \in \{0,1\}^K$ instead of $\boldsymbol{\alpha} \in \mathcal{A}$ in order to compute an accurate RMSE even when the estimated $\widehat{\mathcal{A}}$ is incorrect. The RMSEs for the item parameters and the LCBN parameters $\boldsymbol{t}$ are similarly defined. The estimation accuracy of $\mathcal{E}$ is defined as $\text{Acc}(\widehat{\mathcal{E}}) = \frac{1}{C} \sum_{c=1}^{C} \mathbb{1}(\widehat{\mathcal{E}}^{(c)} = \mathcal{E})$,

where $\widehat{\mathcal{E}}^{(c)} = \mathcal{E}$ indicates that the entire hierarchy is exactly recovered. We also compare our final estimated model (denoted by LCBN in the table) to the first-stage estimate (denoted by PEM in the table) by comparing their EBIC values. The simulation results for the LCBN-based DINA and GDINA are summarized in Tables 3 and 4, respectively. The "argmin EBIC" column in Table 3 (or 4) records the percentage of each method (PEM or our two-step procedure) that achieves the minimum EBIC value among the 100 simulation replicates. We report additional simulation details (convergence criteria and the choice of tuning parameters) and computation time in Supplementary Material S.4.1.

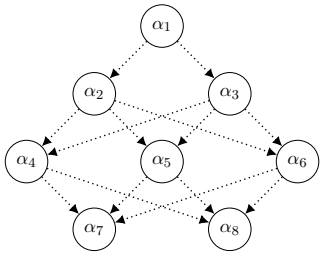


Fig 5: Diamond hierarchy.

| $\mathcal{A}(\mathcal{E})$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ | $p_{\boldsymbol{\alpha}}$ |
|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{\alpha}_1$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.100 |
| $\boldsymbol{\alpha}_2$ | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.036 |
| $\boldsymbol{\alpha}_3$ | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.144 |
| $\boldsymbol{\alpha}_4$ | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.144 |
| $\boldsymbol{\alpha}_5$ | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0.016 |
| $\boldsymbol{\alpha}_6$ | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0.036 |
| $\boldsymbol{\alpha}_7$ | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0.036 |
| $\boldsymbol{\alpha}_8$ | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0.085 |
| $\boldsymbol{\alpha}_9$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0.036 |
| $\boldsymbol{\alpha}_{10}$ | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0.085 |
| $\boldsymbol{\alpha}_{11}$ | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0.085 |
| $\boldsymbol{\alpha}_{12}$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0.032 |
| $\boldsymbol{\alpha}_{13}$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0.047 |
| $\boldsymbol{\alpha}_{14}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0.047 |
| $\boldsymbol{\alpha}_{15}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.071 |

Table 2: Permissible patterns under the diamond hierarchy

Tables 3 and 4 show that our method is effective in recovering the attribute hierarchy $\mathcal{E}$. In particular, the recovery accuracy improves as the sample size $N$ increases and noise level $r$ decreases. In particular, when $N$ is large ($N = 2000$), $\text{Acc}(\widehat{\mathcal{E}})$ is above 0.97 in all of our simulation settings. This observation empirically verifies the identifiability and estimation consistency of $\mathcal{E}$. The accuracy of recovering the hierarchy $\mathcal{E}$ in Tables 3 and 4 is close to 90% or higher in all scenarios except for the slightly lower values of 74% and 52% when $N = 500$ and $r = 0.2$. These two lower accuracy values correspond to the smallest signal-to-noise settings under DINA and GDINA models. Additionally, the estimation accuracy under GDINA is lower than that under DINA, as it has more item parameters that need to be estimated (in our settings, GDINA has 108 parameters whereas DINA has 48 parameters).

Tables 3 and 4 also show that the proposed method can accurately estimate the continuous parameters $\boldsymbol{p}$, $\boldsymbol{\Theta}$, and $\boldsymbol{t}$. Similar to the estimation of $\widehat{\mathcal{E}}$, the estimation error of the continuous parameters is smaller under a smaller noise level $r$, and it decreases as sample size $N$ increases. This observation again corroborates our identifiability and consistency results of the LCBN model parameters. One can also see that the RMSE of $\boldsymbol{p}$ and $\boldsymbol{\Theta}$ after our second-step algorithm is smaller than the RMSE after just the first-step. This indicates that our second-step estimation procedure improves the overall estimation accuracy, by properly taking into account the LCBN structure. In addition, even when the hierarchy is incorrectly estimated in the first-step, the error for estimating the continuous parameters in the second step is still not large. For example, for the $N = 500, r = 0.1$ row in Table 3, the RMSEs of $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{t}}$ when the hierarchy is incorrect are 0.031 and 0.103, respectively. These numbers are comparable to the overall average RMSEs of 0.029 and 0.042 in the corresponding row of the table.

Finally, the model selected after the second-step tends to have a lower EBIC value compared to the first-step selected model. This demonstrates that our parsimonious LCBN is preferable to the unstructured attribute hierarchy model fitted by PEM.

TABLE 3

*Estimation accuracy of attribute hierarchy and RMSE for the estimated parameters for the DINA-based LCBN. The "argmin EBIC" column shows the percentage of each method (PEM or proposed) having a smaller EBIC.*

| Model | $N$ | $r$ | Method | Acc($\widehat{\mathcal{E}}$) | argmin EBIC | RMSE($\widehat{\boldsymbol{\Theta}}$) | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{t}}$) |
|---|---|---|---|---|---|---|---|---|
| DINA | 500 | 0.1 | PEM | – | 7% | 0.042 | 0.005 | – |
| | | | Proposed | 0.92 | 93% | 0.029 | 0.004 | 0.042 |
| | | 0.2 | PEM | – | 6% | 0.053 | 0.008 | – |
| | | | Proposed | 0.74 | 94% | 0.046 | 0.006 | 0.053 |
| | 1000 | 0.1 | PEM | – | 2% | 0.033 | 0.004 | – |
| | | | Proposed | 0.98 | 98% | 0.021 | 0.003 | 0.027 |
| | | 0.2 | PEM | – | 6% | 0.040 | 0.006 | – |
| | | | Proposed | 0.94 | 94% | 0.033 | 0.004 | 0.038 |
| | 2000 | 0.1 | PEM | – | 2% | 0.021 | 0.002 | – |
| | | | Proposed | 0.98 | 98% | 0.015 | 0.001 | 0.021 |
| | | 0.2 | PEM | – | 0% | 0.029 | 0.004 | – |
| | | | Proposed | 1.00 | 100% | 0.021 | 0.002 | 0.022 |

TABLE 4

*Estimation accuracy of attribute hierarchy and RMSE for the estimated parameters for the GDINA-based LCBN. The "argmin EBIC" column shows the percentage of each method (PEM or proposed) having a smaller EBIC.*

| Model | $N$ | $r$ | Method | Acc($\widehat{\mathcal{E}}$) | argmin EBIC | RMSE($\widehat{\boldsymbol{\Theta}}$) | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{t}}$) |
|---|---|---|---|---|---|---|---|---|
| GDINA | 500 | 0.1 | PEM | – | 0% | – | 0.005 | – |
| | | | Proposed | 0.99 | 100% | 0.109 | 0.003 | 0.039 |
| | | 0.2 | PEM | – | 5% | – | 0.009 | – |
| | | | Proposed | 0.52 | 95% | 0.176 | 0.004 | 0.086 |
| | 1000 | 0.1 | PEM | – | 1% | – | 0.003 | – |
| | | | Proposed | 0.99 | 99% | 0.072 | 0.002 | 0.030 |
| | | 0.2 | PEM | – | 0% | – | 0.007 | – |
| | | | Proposed | 0.89 | 100% | 0.121 | 0.003 | 0.055 |
| | 2000 | 0.1 | PEM | – | 3% | – | 0.002 | – |
| | | | Proposed | 0.97 | 97% | 0.052 | 0.001 | 0.025 |
| | | 0.2 | PEM | – | 0% | – | 0.005 | – |
| | | | Proposed | 0.99 | 100% | 0.080 | 0.002 | 0.035 |

We report additional simulation results in the Supplementary Material to further support our proposed method. In the Supplementary Material, Section S.4.4 includes simulations when the proportion parameters $\boldsymbol{p}$ respect the hierarchy graph but attributes do not exhibit the induced conditional independence asserted by LCBNs; Section S.4.5 includes sensitivity analysis for choosing the tuning parameter $\lambda$ in the log penalty; Section S.4.6 provides detailed analysis on the uncertainty of estimating the attribute hierarchy $\mathcal{E}$.

**6. Application to Data from the Trends in Mathematics and Science Study.** In this section, we apply the proposed method to analyze an educational assessment dataset from the Trends in Mathematics and Science Study (TIMSS). TIMSS is a series of international assessments of fourth and eighth graders' mathematics and science knowledge, involving students in over 60 countries (Mullis et al., 2012). We analyze the TIMSS 2011 Austrian fourth-

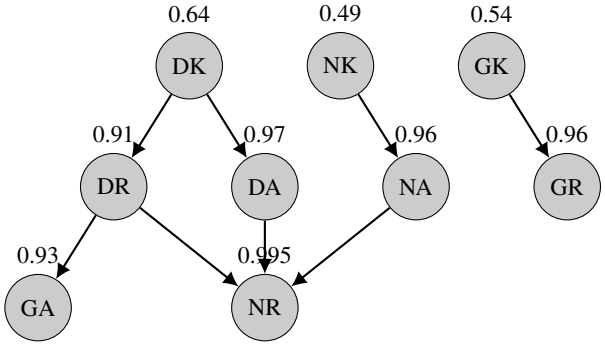


Fig 6: Initial hierarchy of the TIMSS 2011 dataset. The LCBN parameters $t_k$ are displayed above each skill.

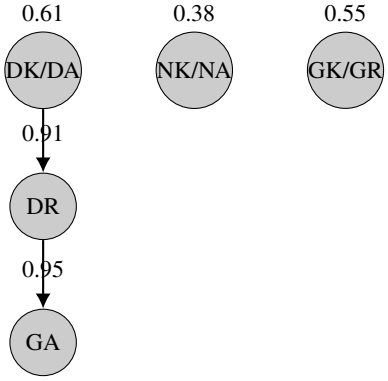


Fig 7: Re-estimated hierarchy

grade mathematics test data, which is publicly available in the R package CDM (George et al., 2016). The data contains the responses of $N = 4668$ Austrian students to $J = 174$ test items. Educational experts have specified the $K = 9$ fine-grained skill attributes to be: (DA) Data and Applying, (DK) Data and Knowing, (DR) Data and Reasoning, (GA) Geometry and Applying, (GK) Geometry and Knowing, (GR) Geometry and Reasoning, (NA) Numbers and Applying, (NK) Numbers and Knowing, (NR) Numbers and Reasoning (George and Robitzsch, 2015). These nine skill attributes were defined by considering the combinations of three content skills (Data, Geometry, and Number) and three cognitive skills (Applying, Knowing, and Reasoning). This attribute definition follows George and Robitzsch (2015), where a corresponding $Q$-matrix was also specified. This $Q$-matrix assumes that each item measures exactly one attribute, and it satisfies our identifiability conditions in Theorem 1.

One structure specific to large scale assessments such as TIMSS is that only a subset of all items in the entire study is presented to each of the students (George and Robitzsch, 2015). This results in many missing entries in the $N \times J$ data matrix. Nevertheless, these entries are missing at random because the missingness patterns do not depend on the students' latent skills or model parameters. Our estimation algorithms can be easily adapted to this setting. Specifically, in the complete data log likelihood used in our EM algorithms, we can just replace the summation range from $\sum_{i=1}^{N}\sum_{j=1}^{J}$ to $\sum_{(i,j)\in\Omega}$, where $\Omega$ is the collection of indices $(i, j)$ that correspond to the observed entries in the data matrix.

As a first analysis, we apply the two-step method in Section 4 to estimate the latent hierarchy graph and the continuous parameters. Since each row in the $Q$-matrix has exactly one nonzero entry, the DINA and GDINA models under this $Q$-matrix are equivalent. So we just adopt the DINA model in the analysis. Algorithm 1 selected 15 attribute patterns, and Figure 6 shows the estimated attribute hierarchy with the latent CBN parameter $t_1, \ldots, t_9$. Figure 6 reveals that there are three ancestor attributes, DK, NK, GK that serve as the prerequisite attributes for each type of content skills in Data, Number, and Geometry. This implies that among the three cognitive skills Knowing, Applying, and Reasoning, the skill Knowing is the most basic. If a student "Knows" a certain content skill, then they possesses the prerequisite to "Apply" or "Reason" the same content skill, sometimes with the aid of other content skills. For instance, NR (Number and Reasoning) requires DR (Data and Reasoning) and DA (Data and Applying) in addition to NA (Number and Applying) as prerequisites.

Recall that each $t_k$ gives the conditional probability of mastering attribute $\alpha_k$ provided that one has already mastered all of $\alpha_k$'s prerequisites. One consequence of this definition is that $t_k$ does not capture the individual effect of mastering any specific parent on the mastery of $\alpha_k$. As pointed out by a reviewer, sometimes it may also be interesting to consider such individual effects, e.g., the skill NR in Figure 6 has three parents and one may wish to distinguish their individual influences. One possible way to indirectly think about this could be to compare the

values of marginal mastery probability $\mathbb{P}(\alpha_l = 1)$ for each parent skill $l \in \text{pa}(k)$. The parent skill $\alpha_l$ with the smallest marginal mastery probability $\mathbb{P}(\alpha_l = 1)$ could be viewed as having the largest influence on the mastery of the child skill $\alpha_k = 1$. Going back to the current data example with $k =$NR, the skill NA has the largest influence on NR among the three parent skills since $\mathbb{P}(\text{NA} = 1)$ is the smallest among those three. We also include more discussions on potential alternative models to distinguish individual effects in Section 7.

Additionally, Figure 6 shows that many $t_k$ parameters in the second or third layer are larger than 0.9, whereas the ancestor attributes have much smaller $t_k$ values. Specifically, consider $t_{DA} = 0.97$. Then, $\mathbb{P}(\alpha_{DA} = 0 \mid \alpha_{DK} = 0) = 1$ and $\mathbb{P}(\alpha_{DA} = 0 \mid \alpha_{DK} = 1) = 0.03$, whereas $\mathbb{P}(\alpha_{DA} = 1 \mid \alpha_{DK} = 1) = 0.97$. This implies that DA may not be a meaningful attribute, as it does not offer additional discrimination of students compared to DK. Therefore, we conduct a second analysis and merge those attributes whose $t_k > 0.95$. For instance, we combine the attributes "DA" and "DK" into one "meta" attribute. This simplification reduces the number of attributes $K$ from nine to five and the number of permissible attribute patterns $|\mathcal{A}|$ from 69 to 16. Then we fit an LCBN with this new attribute hierarchy in Figure 7, where the new $Q$-matrix can be obtained by summing the corresponding columns in the original $Q$-matrix. The fitted LCBN parameters are shown in Figure 7. The final result has the log likelihood equal to $-5.88 \times 10^4$ and EBIC equal to $1.205 \times 10^5$, which is a great improvement compared to the values in our first analysis (previous log likelihood equal to $-6.43 \times 10^4$ and EBIC equal to $1.327 \times 10^5$). This implies that merging the attributes and fitting an even more parsimonious LCBN model provides better fit to data. In summary, our LCBN model is a parsimonious and interpretable alternative to existing cognitive diagnostic models, and is especially useful to make sense of data arising from modern large-scale educational assessments such as TIMSS.

**7. Discussion.** We have proposed a new family of latent variable models, the latent conjunctive Bayesian networks, for modeling cognitive diagnostic assessment data in education. The LCBN family rigorously unifies the attribute hierarchy method in educational cognitive diagnosis and Bayesian networks in statistical machine learning. Compared to existing modeling approaches, our model is identifiable, parsimonious, and provides nice interpretation of conditional independence. We propose a two-step method that efficiently estimates the discrete attribute hierarchy graph and the continuous model parameters, and establish the consistency of this procedure. We have also shown that our method can be easily extended to more challenging settings with an unknown $\mathbf{Q}$-matrix. Simulation studies and real data analysis demonstrate that our method has good empirical performance.

Our estimation procedure is scalable and can be easily applied to analyze modern large-scale assessment data, such as TIMSS and Program for International Student Assessment data. Most existing studies of attribute hierarchy focused on the cases when $K = 3$ or 4 due to the computational cost of estimating potentially exponentially many proportion parameters under an unstructured attribute hierarchy model (e.g., Templin and Bradshaw, 2014; Wang and Lu, 2021). On the contrary, our LCBN only requires a linear number of $K$ parameters to specify the latent attribute distribution and is much more parsimonious.

This work proposes the most parsimonious Bayesian network model, LCBN, for attribute hierarchy. In the future, it would also be interesting to explore other Bayesian network models in the cognitive diagnostic applications. For example, sometimes the conjunctive assumption in LCBN may be too strong or there may exist multiple paths to master a skill. To this end, one could consider a latent *disjunctive* Bayesian network:

$$\mathbb{P}(\alpha_k = 1 \mid \alpha_{\text{pa}(k)}) = \begin{cases} 0, & \text{if } \prod_{l \in \text{pa}(k)} (1 - \alpha_l) = 1, \\ t_k, & \text{otherwise.} \end{cases}$$

The above model assumes that as long as a student masters one of the parent attributes of $\alpha_k$, they will have a probability $t_k$ to master $\alpha_k$. Alternatively, we could define the following latent *additive* Bayesian network model that defines the conditional mastery probability as a linear combination of those parent attributes: $\mathbb{P}(\alpha_k = 1 \mid \alpha_{\text{pa}(k)}) = \sum_{l \in \text{pa}(k)} t_{k,l}\alpha_l$, where $t_{l,k} \geq 0$ and $\sum_{l \in \text{pa}(k)} t_{l,k} \leq 1$. In this model, mastering each parent attribute $\alpha_l$ increases the mastery probability of the child attribute $\alpha_k$ by $t_{k,l}$. This model is less parsimonious than LCBNs, but could model different paths to mastering $\alpha_k$ with different probabilities.

An interesting future direction is to study the double-asymptotic regime where $N$ and $J$ both go to infinity and aim to also consistently estimate the individual-level latent profiles $\mathbf{A}_i$'s. In this work, we study identifiability in the fixed $J$ regime and focus on identifying and estimating the *population* quantities $(\mathcal{E}, \boldsymbol{\Theta}, \boldsymbol{t})$. On the other hand, when $J$ goes to infinity with increasing information provided by each student, it may be possible to consistently estimate the individual students' latent skills $\mathbf{A}_i$ in the *sample* (e.g., Gu and Xu, 2023b). Such sample estimates would provide reliable personalized diagnosis. Furthermore, if individual students' skills are consistently estimated, then the LCBN parameters can be estimated via a closed form MLE (Beerenwinkel et al., 2006, 2007). This can be an alternative estimation method suitable for the double-asymptotic regime without using regularization. Another interesting future direction is to employ LCBNs in adaptive learning or reinforcement learning settings (Chen et al., 2018; Tang et al., 2019) to help design recommendation strategies and enhance learning. Thanks to LCBNs' parsimony, interpretability, and identifiability, it is attractive to incorporate LCBNs in these computationally intensive applications to help achieve more reliable decision making and recommendations. We leave these directions for future research.

**Acknowledgments.** The authors thank the Editors and referees for their constructive and helpful comments. This work is partially supported by NSF Grant DMS-2210796.

## SUPPLEMENTARY MATERIAL

The Supplementary Material includes proofs of the theorems, additional identifiability results, various additional simulation studies and details.

## REFERENCES

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

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2006). Evolution on distributive lattices. *Journal of Theoretical Biology*, 242(2):409–420.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2007). Conjunctive Bayesian networks. *Bernoulli*, pages 893–909.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., and Lengauer, T. (2005). Learning multiple evolutionary pathways from cross-sectional data. *Journal of Computational Biology*, 12(6):584–598.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Briggs, D. C. and Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In *Learning progressions in science*, pages 293–316. Brill.

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

Chen, Y., Li, X., Liu, J., and Ying, Z. (2018). Recommendation system for adaptive learning. *Applied Psychological Measurement*, 42(1):24–41.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2):179–199.

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

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

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., and Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74:1–24.

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.

Gratzer, G. (2009). *Lattice theory: First concepts and distributive lattices*. Courier Corporation.

Gu, Y. and Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(115):1–58.

Gu, Y. and Xu, G. (2023a). Identifiability of hierarchical latent attribute models. *Statistica Sinica*, 33:2561–2591.

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

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(2):191–210.

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755.

Hu, B. and Templin, J. (2020). Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivariate Behavioral Research*, 55(2):300–311.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.

Leighton, J. and Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41(3):205–237.

Ma, C., Ouyang, J., and Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, 88(1):175–207.

Mullis, I. V., Martin, M. O., Minnich, C. A., Stanco, G. M., Arora, A., Centurino, V. A., and Castle, C. E. (2012). *TIMSS 2011 Encyclopedia: Education Policy and Curriculum in Mathematics and Science. Volume 1: AK*. ERIC.

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

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

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232.

Simon, M. A. and Tzur, R. (2012). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. In *Hypothetical Learning Trajectories*, pages 91–104. Routledge.

Tang, X., Chen, Y., Li, X., Liu, J., and Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *British Journal of Mathematical and Statistical Psychology*, 72(1):108–135.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, pages 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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2):287–307.

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

Wang, C. (2021). Using penalized EM algorithm to infer learning trajectories in latent transition CDM. *Psychometrika*, 86(1):167–189.

Wang, C. and Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48(2):165–187.

Wang, C. and Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics*, 46(1):58–84.

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

Zhan, P., Ma, W., Jiao, H., and Ding, S. (2020). A sequential higher order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Applied Psychological Measurement*, 44(1):65–83.