

# Hypothesis Testing for Hierarchical Structures in Cognitive Diagnosis Models

CHENCHEN MA<sup>1</sup> AND GONGJUN XU<sup>1,\*</sup>

<sup>1</sup>*Department of Statistics, University of Michigan, West Hall, 1085 S University Ave, Ann Arbor, MI 48109, USA*

## Abstract

Cognitive Diagnosis Models (CDMs) are a special family of discrete latent variable models widely used in educational, psychological and social sciences. In many applications of CDMs, certain hierarchical structures among the latent attributes are assumed by researchers to characterize their dependence structure. Specifically, a directed acyclic graph is used to specify hierarchical constraints on the allowable configurations of the discrete latent attributes. In this paper, we consider the important yet unaddressed problem of testing the existence of latent hierarchical structures in CDMs. We first introduce the concept of testability of hierarchical structures in CDMs and present sufficient conditions. Then we study the asymptotic behaviors of the likelihood ratio test (LRT) statistic, which is widely used for testing nested models. Due to the irregularity of the problem, the asymptotic distribution of LRT becomes nonstandard and tends to provide unsatisfactory finite sample performance under practical conditions. We provide statistical insights on such failures, and propose to use parametric bootstrap to perform the testing. We also demonstrate the effectiveness and superiority of parametric bootstrap for testing the latent hierarchies over non-parametric bootstrap and the naïve Chi-squared test through comprehensive simulations and an educational assessment dataset.

**Keywords** *bootstrapping; latent hierarchical structure; likelihood ratio test*

## 1 Introduction

Cognitive Diagnosis Models (CDMs) are a popular family of discrete latent variable models that have been widely used in social and biological sciences. In CDMs, a set of latent attributes are assumed to exist that can explain or govern the observed variables. Under the framework of CDMs, one can get fine-grained inference on subjects' latent attribute profiles based on their multivariate observed responses, which further can be used to infer the subgroups of the population. In practice, the discrete latent attributes often have special scientific meanings. For example, in educational assessments, the latent attributes are assumed to be mastery or deficiency of target skills (Junker and Sijtsma, 2001; de la Torre, 2011); in psychiatric diagnosis, they are modeled as presence or absence of some underlying mental disorders (Templin and Henson, 2006; de la Torre et al., 2018); and in epidemiological and medical measurement studies, the latent attributes are interpreted as existence or nonexistence of some disease pathogens (Wu et al., 2016b,a; O'Brien et al., 2019).

Various CDMs have been developed in the literature by modeling the interactions between the observed variables and the latent attributes differently under different cognitive diagnostic

---

\*Corresponding author. Email: [gongjun@umich.edu](mailto:gongjun@umich.edu).

assumptions. The Deterministic Input Noisy Output “AND” gate (DINA; Haertel, 1989) model and the Deterministic Input Noisy Output “OR” gate (DINO; Templin and Henson, 2006) model are among the most basic ones. More general CDMs include the Generalized DINA (GDINA) model (de la Torre, 2011), the General Diagnostic Model (GDM; von Davier, 2005), the reduced Reparameterized Unified Model (reduced-RUM; DiBello et al., 1995), and the Log-linear Cognitive Diagnosis Models (LCDM; Henson et al., 2009).

Recently the relationship among the latent attributes, especially hierarchical structures among them, has gained increasing research interests. For example, in a learning context, prerequisite relations are often postulated among the latent skill/concept attributes, which can be used to characterize the learning trajectory of the students (Dahlgren et al., 2006; Jimoyiannis and Komis, 2001; Simon and Tzur, 2004; Wang and Gierl, 2011). Based on the prerequisite relationships among the latent attributes, learning materials or test items can be designed, and recommendations or remedy strategies can be generated accordingly. The Attribute Hierarchy Model (Leighton et al., 2004) is a variation of Tatsuoka’s rule-space approach (Tatsuoka, 1983) and uses an adjacency matrix to explicitly defined the hierarchical attribute structures. Templin and Bradshaw (2014) proposed the Hierarchical Cognitive Diagnosis Models (HCDMs), which enforce hard constraints on hierarchical configurations of the latent attributes. In particular, a Directed Acyclic Graph (DAG) was used to characterize the hierarchical dependence structure among the latent attributes, and impose hard constraints on possible latent attribute profiles under hierarchies.

Despite the popularity of CDMs, statistical inference of relationships among the latent attributes especially hierarchical structures has been a challenging yet unaddressed problem in the literature. For example, hypothesis testing plays an important role in validating the presence of suspected attribute hierarchies, which can provide guidance to practitioners for experiment design or data modeling (Templin and Bradshaw, 2014). However, to our best knowledge, there are no systematical testing procedures or statistical theories on hypothesis testing of latent hierarchical structures. Two natural questions about such testings are (1) when the hierarchical structures are testable and (2) how to conduct the hypothesis testing. On the one hand, under the framework of CDMs, if the hierarchical structure under the null hypothesis cannot be distinguished from those under the alternative, we cannot test such a hierarchical structure and therefore it is untestable. In fact, the testability of hierarchical structures is closely related to the identifiability of the models. On the other hand, under the hierarchical constraints, the problem of testing latent hierarchical structures is equivalent to testing the sparsity structure of the set of latent attribute profiles in the population, that is, the sparsity structure of the population proportion parameter vector. However, due to the identifiability and the irregularity issue that the true proportion parameters are on the boundary of the parameter space under hierarchical structures, the conventional asymptotic Chi-squared distribution may not hold for the likelihood ratio test.

Non-regularity issues of the likelihood ratio test are known to exist in many latent variable models such as finite mixture models, factor analysis, structural equation models, and random effects models (Chen, 2017; Chen et al., 2020). In particular, testing the sparsity structure of the proportion parameter vector in CDMs is closely related to the problem of testing the number of components in finite mixture models and latent class models (Nylund et al., 2007; Chen, 2017). However, testing the hierarchical structures in CDMs is even more challenging, since it tests whether a specific set of the proportion parameters specified by the hierarchical structure under the null hypothesis is zero; and such a problem is further complicated due to the restrictions imposed by the structural  $Q$ -matrix and the discrete nature of the latent variables in CDMs.

In this paper, we focus on the problem of hypothesis testing for latent hierarchical structures. We first discuss the testability of latent hierarchical structures and present sufficient conditions under which hierarchical structures are testable in CDMs. Then under such conditions, we examine the asymptotic behaviors of the popularly used likelihood ratio test. Since the true proportion parameter is on the boundary of the parameter space, the asymptotic distribution of LRT becomes nonstandard due to the lack of regularity (Self and Liang, 1987). Moreover, the nonstandard limiting distribution of LRT is observed to not provide satisfactory finite-sample results under practical settings, and we provide statistical insights on such failures. Specifically, we find that when the number of items is large or the item parameters are close to the boundary, the convergence of the nonstandard limiting distribution can be very slow and the test tends to fail. Therefore we do not recommend using the nonstandard limiting distribution to conduct the hypothesis testing in practice. Instead, based on these findings, we propose to use resampling-based methods to test hierarchical structures. We conduct comprehensive simulations and comparisons between parametric bootstrap and nonparametric bootstrap and recommend using parametric bootstrap for testing latent hierarchies in CDMs.

The rest of the paper is organized as follows: the model setup of hierarchical CDMs and the motivations of the problem are introduced in Section 2. Sufficient conditions for testability of hierarchical structures and several illustrative examples are provided in Section 3. Studies on the likelihood ratio test and numerical results are presented in Section 4. Specifically, Section 4.1 studies the asymptotic behaviors of LRT and provides insights on its failures in some situations. Section 4.2 presents simulation studies that compare parametric bootstrap and nonparametric bootstrap for testing hierarchical structures. In Section 5, we perform hypothesis testing for a linear attribute hierarchy in an educational assessment dataset and compare different testing procedures. Finally Section 6 concludes with some discussions.

## 2 Model Setup and Motivations

In this section, we introduce the model setup of CDMs and the problem of interest. We start with some notations. For an integer  $K$ , we use  $[K]$  to denote the set  $\{1, 2, \dots, K\}$ . For two vectors  $\mathbf{a} = (a_1, \dots, a_K)$  and  $\mathbf{b} = (b_1, \dots, b_K)$  of the same length, we define a partial order “ $\succeq$ ” that we say  $\mathbf{a} \succeq \mathbf{b}$  if  $a_k \geq b_k$  for  $\forall k \in [K]$  and  $\mathbf{a} \not\preceq \mathbf{b}$  otherwise. “ $\preceq$ ” and “ $\not\preceq$ ” are defined similarly. We use  $|\cdot|$  to denote the cardinality of a set.

### 2.1 Cognitive Diagnosis Models

We first introduce the model setup of CDMs and provide some illustrative examples. In CDMs,  $J$  items are assumed to depend on  $K$  latent attributes of interest. In this work, both the responses and the latent attributes are assumed to be binary. Based on a subject’s observed responses to the items  $\mathbf{R} = (R_1, \dots, R_J) \in \{0, 1\}^J$ , the latent attribute profile of the subject  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K) \in \{0, 1\}^K$  denoting the profile of possession of the latent attributes needs to be inferred. With  $K$  latent binary attributes, there are  $2^K$  possible latent attribute profiles, and we use  $\mathbf{p} = (p_\alpha : \boldsymbol{\alpha} \in \{0, 1\}^K)$  to denote the proportion parameter vector for the latent attribute profiles which satisfies  $p_\alpha \in [0, 1]$  and  $\sum_{\boldsymbol{\alpha} \in \{0, 1\}^K} p_\alpha = 1$ . We use  $\Delta^n$  to denote the standard  $n$ -simplex, that is,  $\Delta^n = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} : \sum_{i=0}^n t_i = 1, t_i \geq 0 \text{ for all } i\}$ , and then we have  $\mathbf{p} \in \Delta^{2^K-1}$ . In CDMs, the latent attribute profiles are assumed to follow a categorical distribution with proportion parameter vector  $\mathbf{p} \in \Delta^{2^K-1}$ .

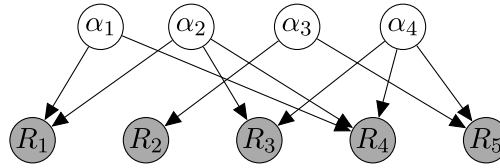


Figure 1: Illustration of Q-matrix.

A key ingredient of CDMs is a structural binary matrix, the so-called  $\mathbf{Q}$ -matrix (Tatsuoka, 1990),  $\mathbf{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$ , which imposes constraints on items to reflect the dependence structure between the items and the latent attributes. Specifically,  $q_{j,k} = 1$  indicates that item  $j$  requires (or depends on) attribute  $k$ . Then the  $j$ th row vector of  $\mathbf{Q}$  denoted by  $\mathbf{q}_j$  describes the full dependence of item  $j$  on  $K$  latent attributes. Usually in applications, the  $\mathbf{Q}$ -matrix is pre-specified by domain experts to reflect some scientific assumptions (George and Robitzsch, 2015; Junker and Sijtsma, 2001; von Davier, 2005). See Figure 1 for illustration of the Q-matrix.

As in classical latent class analysis, given a subject's latent attribute profile  $\boldsymbol{\alpha}$ , the responses to  $J$  items are assumed to be independent and follow Bernoulli distributions with parameters  $\theta_{1,\boldsymbol{\alpha}}, \dots, \theta_{J,\boldsymbol{\alpha}}$ , which are called item parameters in this work. Specifically,  $\theta_{j,\boldsymbol{\alpha}} := \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha})$ . In the literature of psychometrics, the item parameters often refer to the discrimination and difficulty parameters in item response theory (e.g., Reckase, 2009). Here for convenience of discussions, we use item parameters to represent the probabilities of getting a positive response, which are different from the discrimination and difficulty parameters in item response theory. We use  $\boldsymbol{\theta}_j = (\theta_{j,\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0, 1\}^K)$  to denote the item parameter vector for the  $j$ th item, and  $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}} : j \in [J], \boldsymbol{\alpha} \in \{0, 1\}^K)$  to denote the item parameter matrix. Under the local independence assumption, the probability mass function of a subject's response vector  $\mathbf{R} = (R_1, \dots, R_J) \in \{0, 1\}^J$  then can be written as  $\mathbb{P}(\mathbf{R} \mid \boldsymbol{\Theta}, \mathbf{p}) = \sum_{\boldsymbol{\alpha} \in \{0, 1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^J \theta_{j,\boldsymbol{\alpha}}^{R_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_j}$ .

Under the CDM framework, for subject  $i$ , we have latent attribute profile  $\boldsymbol{\alpha}_i$  to indicate subject  $i$ 's possession of  $K$  attributes, and for item  $j$ , we have a structural vector  $\mathbf{q}_j$  to reflect the item's dependence on  $K$  latent attributes. Under different model assumptions, the structural matrix  $\mathbf{Q}$  puts constraints on item parameters in different ways. One important common assumption in CDMs is that the item parameter  $\theta_{j,\boldsymbol{\alpha}}$  only depends on whether the latent attribute profile  $\boldsymbol{\alpha}$  contains the required attributes by item  $j$ , i.e., the attributes in the set  $\mathcal{K}_j = \{k \in [K] : q_{j,k} = 1\}$ , which is called the *set of required attributes of item  $j$*  in Gu and Xu (2019). Therefore, for item  $j$ , the latent attribute profiles which are only different in the attributes outside of  $\mathcal{K}_j$  would have the same item parameters. In such a way, the structural matrix  $\mathbf{Q}$  forces some entries in the item parameter matrix  $\boldsymbol{\theta}$  to be the same. The dependences of item parameters on the required attributed are modeled differently in different CDMs. See Example 1 and Example 2 for two popular families of such models.

**Example 1** (DINA and DINO Models). *The DINA model (Junker and Sijtsma, 2001) and the DINO model (Templin and Henson, 2006) are two basic models in cognitive diagnosis. The DINA model assumes a conjunctive "AND" relationship among the latent binary attributes while the DINO model assumes a conjunctive "OR" relationship. We introduce a binary matrix  $\boldsymbol{\Gamma} = (\Gamma_{j,\boldsymbol{\alpha}} : j \in [J], \boldsymbol{\alpha} \in \{0, 1\}^K) \in \{0, 1\}^{J \times 2^K}$ , which is called the ideal response matrix. In the DINA model, an ideal response is defined as  $\Gamma_{j,\boldsymbol{\alpha}}^{DINA} = \mathbb{I}(\boldsymbol{\alpha} \succeq \mathbf{q}_j) = \prod_{k=1}^K \alpha_k^{q_{j,k}}$ , indicating whether the latent profile contains all the required attributes. In the DINO model, an ideal response is defined as  $\Gamma_{j,\boldsymbol{\alpha}}^{DINO} = \mathbb{I}(\exists k \in [K], q_{j,k} = \alpha_k = 1)$ , indicating whether the latent profile contains any of the*



required attributes. The item level uncertainties are characterized by two parameters, the slipping parameter and the guessing parameter. For item  $j$ , the slipping parameter  $s_j := \mathbb{P}(R_j = 0 \mid \Gamma_{j,\alpha} = 1)$  denotes the probability of a subject giving a negative response despite possessing all the necessary skills; while the guessing parameter  $g_j := \mathbb{P}(R_j = 1 \mid \Gamma_{j,\alpha} = 0)$  denotes the probability of giving a positive response despite deficiency of some necessary skills. Then the item parameter for item  $j$  and latent profile  $\alpha$  can be written as  $\theta_{j,\alpha} := \mathbb{P}(R_j = 1 \mid \alpha) = (1 - s_j)^{\Gamma_{j,\alpha}} g_j^{1 - \Gamma_{j,\alpha}}$ .

**Example 2** (GDINA model). The Generalized DINA model (GDINA, de la Torre, 2011) is a more general model where all interactions among the required latent attributes are considered. The item parameters for the GDINA model are written as

$$\theta_{j,\alpha}^{GDINA} = \beta_{j,0} + \sum_{k=1}^K \beta_{j,k} \alpha_k q_{j,k} + \sum_{k=1}^K \sum_{k'=k+1}^K \beta_{j,k,k'} \alpha_k \alpha_{k'} q_{j,k} q_{j,k'} + \cdots + \beta_{j,1,2,\dots,K} \prod_{k=1}^K \alpha_k q_{j,k}. \quad (1)$$

The coefficients in the GDINA model can be interpreted as following:  $\beta_{j,0}$  is the probability of a positive response for the most incapable subjects with no required attributes present;  $\beta_{j,k}$  is the change in the probability of a positive response due to the main effect of  $\alpha_k$ ;  $\beta_{j,k,k'}$  is the change in the probability of a positive response due to the interaction of  $\alpha_k$  and  $\alpha_{k'}$ ;  $\beta_{j,1,2,\dots,K}$  is the change in the positive probability due to the interaction of all the latent attributes. In the GDINA model, the intercept and main effects are typically assumed to be nonnegative, while the interactions can take negative values. By incorporating all the interactions among the required attributes, the GDINA model is one of the most general CDMs.

## 2.2 Problem and Motivations

Researchers in many applications assume certain hierarchical structures among the latent attributes to characterize their dependence. For example, in cognitive diagnosis modeling, the possession of lower-level skills is often regarded as the prerequisite for gaining higher-level skills (Leighton et al., 2004; Templin and Bradshaw, 2014). If some hierarchical structure exists, any latent profile  $\alpha$  that does not respect the hierarchy is deemed not to exist with the corresponding population proportion  $p_\alpha = 0$ . For  $1 \leq k \neq l \leq K$ , we use  $\alpha_k \rightarrow \alpha_l$  (or  $k \rightarrow l$ ) to denote the hierarchy that attribute  $\alpha_k$  is a prerequisite for attribute  $\alpha_l$ . Under the hierarchy  $\alpha_k \rightarrow \alpha_l$ , the latent profiles with  $\alpha_l = 1$  but  $\alpha_k = 0$  will not exist in the population and therefore we have  $p_\alpha = 0$  if  $\alpha_l = 1$  but  $\alpha_k = 0$ . The set of prerequisite relationships is denoted by  $\mathcal{E} = \{k \rightarrow l : \text{attribute } k \text{ is a prerequisite for } l, 1 \leq k \neq l \leq K\}$ , and the induced set of existing latent attribute profiles is denoted by  $\mathcal{A} = \{\alpha \in \{0, 1\}^K : p_\alpha \neq 0 \text{ under } \mathcal{E}\}$ . It is noted that an attribute hierarchy results in the sparsity of the proportion parameter vector, which will reduce the number of model parameters especially when  $K$  is large. Example hierarchical structures and the corresponding induced latent profile sets are shown in Figure 2.

In many applications, certain hierarchical structures are postulated by domain experts. In this paper, we consider the problem of hypothesis testing for the existence of such pre-specified latent hierarchical structures. As we illustrated above, the hierarchical structure of the latent attributes results in the sparsity structure of the proportion parameter vector for the latent attribute profiles, since the latent profiles that do not follow the hierarchical structure will not exist in the population. Therefore the problem of testing latent hierarchy is equivalent to testing the sparsity structure of the proportion parameter vector. More formally, we aim to test the following hypothesis:

$$H_0 : p_\alpha = 0, \forall \alpha \notin \mathcal{A}_0 \text{ under hierarchy } \mathcal{E}_0,$$

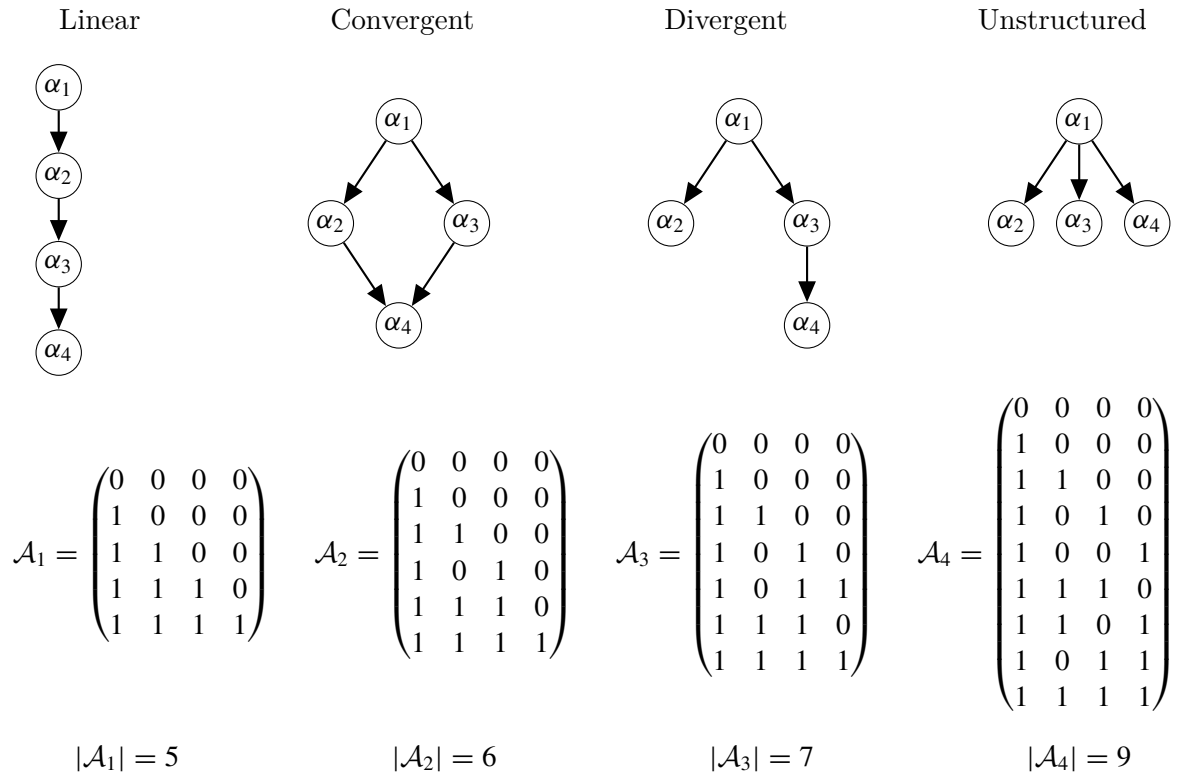


Figure 2: Examples of hierarchical structures of latent attributes. For  $i = 1, \dots, 4$ , each  $\mathcal{A}_i$  represents the induced set of existent attribute profiles under the hierarchical structure above it, where each row in  $\mathcal{A}_i$  represents an attribute profile  $\alpha$  with  $p_\alpha \neq 0$ .

where  $\mathcal{E}_0$  is the hierarchical structure under the null hypothesis and  $\mathcal{A}_0$  is the induced latent attribute profile set under  $\mathcal{E}_0$ .

Even though CDMs can be viewed as a special family of finite mixture models, there is a key difference between testing hierarchical structures in CDMs and testing the number of components in finite mixture models. When testing the number of components in finite mixture models, there are no restrictions on the components' distributions. See Chen (2017) for a review of testing the number of components in finite mixture models. However, when testing latent hierarchical structures, we are in fact testing whether the proportion parameters of the nonexistent latent attribute profiles corresponding to the hierarchy are zeros. Moreover, the constraints imposed by the structural  $Q$ -matrix makes it more restrictive and complicated.

### 3 Testability Requirements and Conditions

Before we introduce concrete testing procedures, we first need to understand when the hierarchical structures are testable. For instance, consider the case when the item parameters are the same for two latent attribute profiles and we want to test the nonexistence of one of them. In this situation, we cannot distinguish these two profiles and thus cannot identify their proportion parameters, not mention to test whether the corresponding proportion is zero. Example 3 provides an illustrative example. Therefore, the testability issue is of fundamental importance before performing concrete testing procedures. Moreover, the testability conditions would also

provide guidance for practitioners and scientific researchers to design experiments.

**Example 3.** Assume that there are two latent attributes of interest and there is a linear attribute hierarchy  $\mathcal{E}_0 = \{1 \rightarrow 2\}$ , which results in the induced attribute profile set  $\mathcal{A}_0 = \{(0, 0), (1, 0), (1, 1)\}$ . If the  $\mathbf{Q}$ -matrix is specified as,

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix},$$

then under the DINA model assumption, the item parameter vector for the latent profile  $(0, 0)$ ,  $\boldsymbol{\theta}_{(0,0)}$ , would be the same as  $\boldsymbol{\theta}_{(0,1)}$ . In this case, we cannot distinguish the profiles  $(0, 0)$  and  $(0, 1)$ , and therefore the proportion parameters  $p_{(0,0)}$  and  $p_{(0,1)}$  cannot be identified. Furthermore, the induced profile set  $\mathcal{A}_0$  is not identifiable, which makes the latent hierarchical structure untestable.

To ensure the testability of hierarchical structures, some conditions need to be met. Before we dive into these conditions, let's first define the concept of the testability of latent hierarchies.

**Definition 1** (Strict testability of  $\mathcal{E}_0$ ). Given the  $\mathbf{Q}$ -matrix and certain cognitive diagnosis model assumptions, consider the following hypothesis testing:

$$\begin{aligned} H_0: & \text{the latent attributes respect the hierarchy } \mathcal{E}_0, \\ H_1: & \text{the latent attributes do not respect the hierarchy } \mathcal{E}_0. \end{aligned}$$

Then the latent hierarchy  $\mathcal{E}_0$  is said to be testable, if there is no parameter under the alternative hypothesis gives the same distribution as the parameters under the null hypothesis.

In fact, the testability is closely related to the identifiability of CDMs (e.g., Xu and Zhang, 2016; Xu, 2017; Xu and Shang, 2018). The identifiability refers to that if two parameters give the same distribution, then the two parameters must be the same. Nevertheless, the testability of hierarchical structure is actually less restrictive compared with the identifiability. In testing latent hierarchies, we only need to distinguish the latent attribute profiles under the null hierarchical structure with the others under the alternative, while in terms of the identifiability, we need to identify all the model parameters and all the latent attribute profiles. Therefore the concept of testability is weaker than the definitions of identifiability. In particular, identifiability is a sufficient but not necessary condition for testability.

We first consider the DINA model. For the DINA model, since the item parameters only depend on the highest interactions among the required latent attributes, we have equivalent  $\mathbf{Q}$ -matrices under hierarchical structures. Here we say two  $\mathbf{Q}$ -matrices are equivalent under hierarchical structure  $\mathcal{E}$ , denoted by  $\mathbf{Q}_1 \stackrel{\mathcal{E}}{\sim} \mathbf{Q}_2$ , if they give the same item parameter matrices, that is,  $\boldsymbol{\Theta}(\mathbf{Q}_1, \mathcal{A}_{\mathcal{E}}) = \boldsymbol{\Theta}(\mathbf{Q}_2, \mathcal{A}_{\mathcal{E}})$ , where  $\mathcal{A}_{\mathcal{E}}$  is the induced latent attribute profile set under hierarchy  $\mathcal{E}$ . For example, consider three latent attributes with a linear hierarchy, that is,  $\mathcal{E} = \{1 \rightarrow 2 \rightarrow 3\}$ . We have

$$\mathbf{Q}^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \stackrel{\mathcal{E}}{\sim} \mathbf{Q}^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \stackrel{\mathcal{E}}{\sim} \mathbf{Q}^{(*)} = \begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & * & 1 \end{pmatrix}, \tag{2}$$

where “\*” can be either 0 or 1. Based on this observation, following Gu and Xu (2021), we introduce two useful operations on the  $\mathbf{Q}$ -matrix.

**Definition 2.** Given an attribute hierarchy  $\mathcal{E}$  and a  $Q$ -matrix  $\mathbf{Q}$ . For any  $q_{j,l} = 1$  and  $k \rightarrow l$ , set  $q_{j,k}$  to 0 and obtain a modified matrix  $\mathcal{S}^{\mathcal{E}}(\mathbf{Q})$ , which is called the “sparsified” version of  $\mathbf{Q}$ .

**Definition 3.** Given an attribute hierarchy  $\mathcal{E}$  and a  $Q$ -matrix  $\mathbf{Q}$ . For any  $q_{j,l} = 1$  and  $k \rightarrow l$ , set  $q_{j,k}$  to 1 and obtain a modified matrix  $\mathcal{D}^{\mathcal{E}}(\mathbf{Q})$ , which is called the “densified” version of  $\mathbf{Q}$ .

As we discussed previously, the identifiability conditions are sufficient conditions for the testability. We present some identifiability results in Gu and Xu (2021) for the DINA model.

**Proposition 1** (Strict testability for the DINA model). Consider a DINA model with a given  $\mathbf{Q}$ . A hierarchy  $\mathcal{E}_0$  is testable if  $\mathbf{Q}$  satisfies the following conditions:

- (1)  $\mathbf{Q}$  contains a  $K \times K$  identity submatrix  $I_K$ . (Without loss of generality, assume the first  $K$  rows of  $\mathbf{Q}$  form  $I_K$  and denote the remaining submatrix of  $\mathbf{Q}$  by  $\mathbf{Q}^*$ .)
- (2)  $\mathcal{S}^{\mathcal{E}_0}(\mathbf{Q})$ , the sparsified version of  $\mathbf{Q}$ , has at least three entries of “1” in each column.
- (3)  $\mathcal{D}^{\mathcal{E}_0}(\mathbf{Q}^*)$ , the densified version of  $\mathbf{Q}^*$ , contains  $K$  distinct column vectors.

Among the above conditions, condition (1) is in fact necessary to ensure the testability of any hierarchy  $\mathcal{E}$ , which is however not satisfied in Example 3. Since the  $Q$ -matrix in Example 3 does not contain (0, 1), we can not distinguish the latent attribute profiles (0, 1) and (0, 0), making the hierarchy not testable. Conditions (2) and (3), on the other hand, may be further weakened. For example, Gu and Xu (2021) provided necessary conditions for the identifiability of DINA-based hierarchical cognitive diagnosis models and discussed different necessary conditions for different hierarchical structures. As testability is a less restrictive concept than identifiability, we would expect the identifiability conditions (2) and (3) can be further weakened for testability. We leave such an interesting yet challenging question to future study. We next revisit Example 3 with a different  $Q$ -matrix and demonstrate its testability.

**Example 4** (Example 3 revisited). Consider the same setting as in Example 3, but with a different  $Q$ -matrix specified as below:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{I}_2 \\ \mathbf{Q}^* \end{pmatrix}, \text{ where } \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } \mathbf{Q}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Then the modified  $Q$  matrices are:

$$\mathcal{S}^{\mathcal{E}_0}(\mathbf{Q}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ \mathbf{0} & 1 \end{pmatrix}, \quad \mathcal{D}^{\mathcal{E}_0}(\mathbf{Q}^*) = \begin{pmatrix} 1 & 0 \\ \mathbf{1} & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

Since the  $Q$ -matrix contains an identity matrix  $\mathbf{I}_2$ , the sparsified version  $\mathcal{S}^{\mathcal{E}_0}(\mathbf{Q})$  has three “1” entries in each column, and the densified version  $\mathcal{D}^{\mathcal{E}_0}(\mathbf{Q}^*)$  contains two distinct columns, all three conditions in Theorem 1 are satisfied. Therefore, for a DINA model with this  $Q$ -matrix, the linear hierarchy  $\mathcal{E}_0$  is strictly testable.

Given an attribute hierarchy  $\mathcal{E}_0$ , if we are interested in testing a subset of prerequisite relations conditioned on that the other prerequisite relations are assumed, we can further relax Condition (1) in Proposition 1. For example, assume that there are three latent attributes and the full hierarchical structure is  $\mathcal{E}_0 = \{1 \rightarrow 2 \rightarrow 3\}$ . If we are interested in testing  $\mathcal{E} = \{1 \rightarrow 2\}$  given  $\mathcal{E}_0 \setminus \mathcal{E} = \{2 \rightarrow 3\}$ , that is  $H_0 : \mathcal{E}_0$  vs.  $H_1 : \mathcal{E}_0 \setminus \mathcal{E}$ , we can relax Condition (1) in Proposition 1 to Condition (1\*) in Corollary 1.

**Corollary 1.** *Consider a DINA model with a given  $\mathbf{Q}$  and a given attribute hierarchy  $\mathcal{E}_0$ . Suppose we are interested in testing a subset  $\mathcal{E} \subset \mathcal{E}_0$  given that  $\mathcal{E}_0 \setminus \mathcal{E}$  has already been assumed. It is testable if Condition (2) and (3) in Proposition 1 and the following condition are satisfied: (1\*)  $\mathcal{S}^{\mathcal{E}_0}(\mathbf{Q})$ , the sparsified version of  $\mathbf{Q}$ , contains an identity submatrix  $I_K$  and for any attribute  $\alpha_k$  involved in  $\mathcal{E}$ , there is an item which is only targeted on this attribute.*

The proof of Corollary 1 directly follows Theorem 1 in Gu and Xu (2021). As we mentioned previously, the testability is a weaker requirement than the identifiability, in that we only need to differentiate the latent attribute profiles between the null and alternative hypothesis for testability. We next provide examples in which the hierarchical structures are testable but the models are not identifiable.

**Example 5** (Testability vs. Identifiability). *Consider a DINA model with three latent attributes. Further assume the slipping and guessing parameters are known. We want to test the linear hierarchy which is specified as  $\mathcal{E}_0 = \{1 \rightarrow 2 \rightarrow 3\}$ . Then the induced latent attribute profile set is  $\mathcal{A}_0 = \{(0, 0, 0), (1, 0, 0), (1, 1, 0), (1, 1, 1)\}$ . We denote the set of latent attribute profiles that do not exist under the hierarchy as  $\mathcal{A}_0^c = \{(0, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1)\}$ . Consider the following  $\mathbf{Q}$ -matrix and get the corresponding ideal response matrices for  $\mathcal{A}_0$  and  $\mathcal{A}_0^c$ :*

$$\mathbf{Q} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}; \quad \Gamma^{\mathcal{A}_0} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma^{\mathcal{A}_0^c} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then the ideal response vectors for classes in  $\mathcal{A}_0$  and  $\mathcal{A}_0^c$  are different, making the hierarchical structure testable. However, the ideal response vectors for  $(0, 0, 0)$  and  $(1, 0, 0)$  are the same, and those for  $(0, 0, 1)$  and  $(1, 0, 1)$  are also the same, making the model not identifiable.

**Example 6** (Condition on Alternative). *Consider the same model and hierarchical structure in Example 5 but with unknown slipping and guessing parameters. Here we are interested in testing  $\{1 \rightarrow 2\}$  given  $\{2 \rightarrow 3\}$ . Consider the  $\mathbf{Q}$ -matrix:*

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since  $\boldsymbol{\theta}_{(0,0,0)} = \boldsymbol{\theta}_{(0,0,1)}$ , we cannot distinguish the latent profiles  $(0, 0, 0)$  and  $(0, 0, 1)$ , and thus not all the latent profiles are identifiable. However, the conditions in Proposition 1 are satisfied, so the hierarchical structure  $\{1 \rightarrow 2\}$  given  $\{2 \rightarrow 3\}$  is testable.

For more general CDMs, we adapt the identifiability results from Gu and Xu (2019) to establish sufficient conditions for the testability of latent attribute hierarchies. Following the same notations in Gu and Xu (2019), we first introduce the so-called constraint matrix  $\Gamma$ . The constraint matrix for a set of latent attribute profiles  $\mathcal{A}$  is defined as  $\Gamma^{\mathcal{A}} = (\mathbb{I}(\alpha \succeq q_j) : \alpha \in \mathcal{A}, j \in [J]) \in \{0, 1\}^{J \times |\mathcal{A}|}$ , which is a binary matrix indicating whether an attribute profile  $\alpha \in \mathcal{A}$  possesses all the required attributes of item  $j$ . Note that for the DINA model, the constraint matrix is also its ideal response matrix, while they are not the same for the DINO model. The defined constraint matrix is used as a tool to study the testability conditions for general CDMs. Based on the constraint matrix, we define a partial order among the latent attribute profiles “ $\succeq_S$ ” for any subset of items  $S \subset [J]$ . For  $\alpha, \alpha' \in \mathcal{A}$ , we say  $\alpha \succeq_S \alpha'$  under  $\Gamma^{\mathcal{A}}$  if  $\Gamma_{j,\alpha}^{\mathcal{A}} \geq \Gamma_{j,\alpha'}^{\mathcal{A}}$  for  $j \in S$ . And for two item sets  $S_1$  and  $S_2$ , we say “ $\succeq_{S_1} = \succeq_{S_2}$ ” if for any  $\alpha, \alpha' \in \mathcal{A}$ , we have  $\alpha \succeq_{S_1} \alpha'$  if and only if  $\alpha \succeq_{S_2} \alpha'$ . Example 7 provides illustrations for the partial orders.

**Example 7.** Consider two latent attributes with a linear hierarchical structure. The  $Q$ -matrix considered and the corresponding constraint matrix for the latent attribute profile set  $\mathcal{A} = \{(0, 0), (1, 0), (1, 1)\}$  are specified as:

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \Gamma^{\mathcal{A}} = \begin{matrix} & \begin{matrix} (0, 0) & (1, 0) & (1, 1) \end{matrix} \\ \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \end{matrix}.$$

For the item set  $S = \{1, 2\}$ , we can see that  $\Gamma_{j,(1,0)} \geq \Gamma_{j,(0,0)}$  and  $\Gamma_{j,(1,1)} \geq \Gamma_{j,(1,0)}$  for  $j \in S$ . Therefore  $(1, 0) \succeq_S (0, 0)$  and  $(1, 1) \succeq_S (1, 0)$ . Moreover, if we take  $S_1 = \{1, 2\}$  and  $S_2 = \{3, 4\}$ , then for any  $\alpha, \alpha' \in \mathcal{A}$ , we have  $\alpha \succeq_{S_1} \alpha'$  if and only if  $\alpha \succeq_{S_2} \alpha'$ . Therefore,  $\succeq_{S_1} = \succeq_{S_2}$ .

In the following testability results for general CDMs, we focus on equal size cases or underfitted cases when  $|\mathcal{A}| \leq |\mathcal{A}_0|$ , where  $\mathcal{A}_0$  is the set of the latent attribute profiles under the null hypothesis and  $\mathcal{A}$  is the set of the latent attribute profiles under the alternative hypothesis. Note that for overfitted cases with  $|\mathcal{A}| > |\mathcal{A}_0|$ , if  $\mathcal{A}$  and  $\mathcal{A}_0$  lead to the same distribution, the model complexity of  $\mathcal{A}$  is larger than that of  $\mathcal{A}_0$ , and therefore practically we can still distinguish them using information-based criteria or penalized likelihood methods.

**Proposition 2** (Strict testability for general CDMs). Consider a general CDM with a given  $Q$  and an arbitrary hierarchy  $\mathcal{E}_0$ . The hierarchy is testable when the alternative is restricted to the latent profile sets of the same or a smaller size than that under the null hypothesis, if the following conditions of the constraint matrix  $\Gamma^{\mathcal{A}_0}$  corresponding to the induced latent profile set  $\mathcal{A}_0$  under the hierarchy  $\mathcal{E}_0$  are satisfied:

- (1) There exist two disjoint item sets  $S_1$  and  $S_2$ , such that  $\Gamma^{(S_i, \mathcal{A}_0)}$  has distinct column vectors for  $i = 1, 2$  and “ $\succeq_{S_1} = \succeq_{S_2}$ ” under  $\Gamma^{\mathcal{A}_0}$ .
- (2) For any  $\alpha, \alpha' \in \mathcal{A}_0$  where  $\alpha' \succeq_{S_i} \alpha$  under  $\Gamma^{\mathcal{A}_0}$  for  $i = 1$  or  $2$ , there exists some  $j \in (S_1 \cup S_2)^c$  such that  $\Gamma_{j,\alpha}^{\mathcal{A}_0} \neq \Gamma_{j,\alpha'}^{\mathcal{A}_0}$ .
- (3) Any column vector of  $\Gamma^{\mathcal{A}_0}$  is different from any column vector of  $\Gamma^{\mathcal{A}_0^c}$ , where  $\mathcal{A}_0^c = \{0, 1\}^K \setminus \mathcal{A}_0$ .

Based on the conditions in Proposition 2, one can see that having three identity submatrices in the  $Q$ -matrix is sufficient for testability. However, having several identity submatrices is in fact a strong requirement in practice. Under a general CDM, these conditions can be further relaxed



if we consider  $\mathcal{E}_0$  to be testable with the true model parameter ranging almost everywhere in the restricted parameter space except a set of Lebesgue measure zero. Specifically, we have the following definition of generic testability.

**Definition 4** (Generic testability of  $\mathcal{E}_0$ ). *Denote the parameter space under  $\mathcal{E}_0$  by  $\mathbf{\Omega}_0$ . The latent hierarchy  $\mathcal{E}_0$  is said to be generically testable, if there exists a subset  $\mathcal{V}$  of  $\mathbf{\Omega}_0$  that has Lebesgue measure zero, such that there is no parameter under the alternative hypothesis gives the same distribution as the parameters in  $\mathbf{\Omega}_0 \setminus \mathcal{V}$ .*

For generic testability, following the generic identifiability results in Gu and Xu (2019) and Gu and Xu (2020), a nice corollary can be derived where the requirements are directly characterized by the structure of the  $Q$ -matrix.

**Corollary 2.** *If the  $Q$ -matrix satisfies the following conditions, then for any hierarchy  $\mathcal{E}_0$  such that the induced latent attribute profile set  $\mathcal{A}_0$  satisfies Condition (3) in Proposition 2, the hierarchy  $\mathcal{E}_0$  is generically testable:*

(1) *The  $Q$ -matrix contains two  $K \times K$  sub-matrices  $\mathbf{Q}_1$  and  $\mathbf{Q}_2$ , such that for  $i = 1, 2$ ,*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \\ \mathbf{Q}' \end{pmatrix}_{J \times K}; \quad \mathbf{Q}_i = \begin{pmatrix} 1 & * & \cdots & * \\ * & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 1 \end{pmatrix}_{K \times K}, \quad i = 1, 2,$$

where each “\*” can be either zero or one.

(2) *With  $\mathbf{Q}$  in the form as above,  $\sum_{j=2K+1}^J q_{j,k} \geq 1$  for each  $k \in [K]$ .*

By relaxing strict testability to generic testability, less stringent conditions in Corollary 2 have been established. Moreover, the requirements in Corollary 2 can be checked directly from the  $Q$ -matrix, making it easier to use in practice. Next, we present an illustrative example about strict testability and generic testability of general CDMs.

**Example 8.** *Consider a general CDM setting with two latent attributes and a linear attribute hierarchy  $\mathcal{E}_0 = \{1 \rightarrow 2\}$ . Consider the  $Q$ -matrix:*

$$\mathbf{Q} = \begin{pmatrix} \mathbf{I}_2 \\ \mathbf{I}_2 \\ \mathbf{Q}' \end{pmatrix}; \quad \mathbf{Q}' = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

*By directly looking at the  $Q$ -matrix, we know the conditions in Corollary 2 are satisfied and therefore the hierarchical structure is generically testable. Moreover, the constraint matrix under attribute hierarchy  $\mathcal{E}_0 = \{1 \rightarrow 2\}$  is*

$$\mathbf{\Gamma}^{\mathcal{A}_0} = \begin{pmatrix} (0, 0) & (1, 0) & (1, 1) \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}; \quad \mathbf{\Gamma}^{\mathcal{A}_0^c} = \begin{pmatrix} (0, 1) \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

If we set  $S_1 = \{1, 2\}$ ,  $S_2 = \{3, 4\}$ , then  $\Gamma^{(S_i, \mathcal{A}_0)}$  has distinct columns for  $i = 1, 2$ . Moreover, “ $\succeq_{S_1} = \succeq_{S_2}$ ” under  $\Gamma^{\mathcal{A}_0}$ . For  $(1, 0) \succeq_{S_i} (0, 0)$  for  $i = 1$  or  $2$ , we have  $\Gamma_{5,(1,0)}^{\mathcal{A}_0} \neq \Gamma_{5,(0,0)}^{\mathcal{A}_0}$ . For  $(1, 1) \succeq_{S_i} (1, 0)$  for  $i = 1$  or  $2$ , we have  $\Gamma_{6,(1,1)}^{\mathcal{A}_0} \neq \Gamma_{6,(1,0)}^{\mathcal{A}_0}$ . For  $(1, 1) \succeq_{S_i} (0, 0)$  for  $i = 1$  or  $2$ , we have  $\Gamma_{6,(1,1)}^{\mathcal{A}_0} \neq \Gamma_{6,(0,0)}^{\mathcal{A}_0}$ . Finally, the columns of  $\Gamma^{\mathcal{A}_0}$  are different from that of  $\Gamma^{\mathcal{A}_0^c}$ . Therefore, based on the constraint matrix, we can see that the conditions in Proposition 2 are met, and thus the linear attribute hierarchy is also strictly testable.

**Example 9.** Consider a general CDM setting with three latent attributes and a linear hierarchical structure  $\mathcal{E}_0 = \{1 \rightarrow 2 \rightarrow 3\}$ . The induced latent attribute profile set under  $\mathcal{E}_0$  is  $\mathcal{A}_0 = \{(0, 0, 0), (1, 0, 0), (1, 1, 0), (1, 1, 1)\}$ . We denote the complement set of latent attribute profiles as  $\mathcal{A}_0^c = \{(0, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1)\}$ . Consider the  $\mathcal{Q}$ -matrix and the corresponding constraint matrices for  $\mathcal{A}_0$  and  $\mathcal{A}_0^c$ :

$$\mathcal{Q} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad \Gamma^{\mathcal{A}_0} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma^{\mathcal{A}_0^c} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Based on the specified  $\mathcal{Q}$ -matrix and the corresponding constraint matrices, one can easily see that the conditions in Corollary 2 are satisfied and therefore the hierarchical structure is generically testable. However, Condition (1) in Proposition 2 is not satisfied and the model is not strictly identifiable since  $\Gamma^{(0,0,0)}$  and  $\Gamma^{(1,0,0)}$  are the same.

## 4 Likelihood Ratio Test

With the sufficient conditions for the testability of the hierarchical structures specified in Section 3, the next question becomes how to conduct the hypothesis testing. As we illustrated in Section 2.2, when some hierarchical structure exists, the number of truly existing latent attribute profiles will be less than  $2^K$ , and the corresponding model will be a nested model of the full model with all possible latent attribute profiles. Testing the latent hierarchical structure is then equivalent to testing the sparsity structure of the proportion parameter vector. A popular choice of testing a nested model is the likelihood ratio test with an asymptotic Chi-squared distribution under some regularity conditions. One commonly assumed regularity condition is that the true parameter vector is in the interior of the parameter space. However, in our testing problem, the true proportion parameter vector  $\mathbf{p}$  lies on the boundary of the simplex under the null hypothesis, making the conventional Chi-squared limiting distribution no longer hold. In this section, we review the nonstandard asymptotic behaviors of the LRT statistic and provide statistical insights on the failures of such limiting distributions under practical conditions. Then we propose to use resampling-based methods to test hierarchical structures and conduct a comprehensive simulation study to compare different testing procedures.

#### 4.1 Failure of Limiting Distribution of LRT

When the parameter of the null model lies on the boundary of the parameter space, the LRT statistic has been shown to often follow a mixture of  $\chi^2$  distributions asymptotically (Self and Liang, 1987). We first present some general asymptotic theories on the LRT statistic under such nonstandard conditions and discuss the application to our testing problem for latent hierarchies.

Let  $f(\mathbf{x}; \boldsymbol{\theta})$  be the probability density function of a random variable  $\mathbf{X}$ , where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$  takes values in the parameter space  $\boldsymbol{\Omega}$ , a subset of  $\mathbb{R}^p$ . When the model is identifiable, distinct values of  $\boldsymbol{\theta}$  correspond to distinct probability distributions. Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be  $N$  independent observations of  $\mathbf{X}$  and denote the log-likelihood function,  $\sum_{i=1}^N \log[f(\mathbf{x}_i; \boldsymbol{\theta})]$ , by  $l_N(\boldsymbol{\theta})$ . Consider the hypothesis testing

$$H_0 : \boldsymbol{\theta}_0 \in \boldsymbol{\Omega}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta}_0 \in \boldsymbol{\Omega} \setminus \boldsymbol{\Omega}_0,$$

where  $\boldsymbol{\theta}_0$  is the true parameter and  $\boldsymbol{\Omega}_0$  is a subset of  $\boldsymbol{\Omega}$ . When  $\boldsymbol{\Omega}_0$  is an  $r$ -dimensional subset of  $\boldsymbol{\Omega}$ ,  $\boldsymbol{\theta}_0$  is a boundary point of both  $\boldsymbol{\Omega}_0$  and  $\boldsymbol{\Omega} \setminus \boldsymbol{\Omega}_0$  but an interior point of  $\boldsymbol{\Omega}$ , under some regularity conditions, by the Wilk's theorem, the asymptotic distribution of the LRT statistic,  $\lambda_N := -2(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Omega}_0} l_N(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \boldsymbol{\Omega}} l_N(\boldsymbol{\theta}))$ , will be  $\chi^2(p-r)$ . However, when  $\boldsymbol{\theta}_0$  is a boundary point of  $\boldsymbol{\Omega}$ , the regularity condition is not satisfied and the conventional Chi-squared limiting distribution does not hold either.

In Self and Liang (1987), the authors studied the nonstandard tests where the parameter of the null model is on the boundary of the parameter space. It is shown that when some of the true parameter values are on the boundary of the parameter space, under certain regularity conditions, the limiting distribution of the LRT statistic is the same as the distribution of the projection of the Gaussian random variable onto the region of admissible values for the mean. Specifically, both the whole parameter space  $\boldsymbol{\Omega}$  and the null parameter space  $\boldsymbol{\Omega}_0$  are assumed to be regular enough to be approximated by cones with vertices at the true parameter  $\boldsymbol{\theta}_0$ , which is defined as below.

**Definition 5.** *The set  $\boldsymbol{\Omega} \subset \mathbb{R}^p$  is approximated at  $\boldsymbol{\theta}_0$  by a cone with vertex at  $\boldsymbol{\theta}_0$ ,  $C_{\boldsymbol{\Omega}}$ , if*

$$\begin{aligned} (1) \quad & \inf_{\mathbf{x} \in C_{\boldsymbol{\Omega}}} \|\mathbf{x} - \mathbf{y}\| = o(\|\mathbf{y} - \boldsymbol{\theta}_0\|), \quad \forall \mathbf{y} \in \boldsymbol{\Omega}, \\ (2) \quad & \inf_{\mathbf{y} \in \boldsymbol{\Omega}} \|\mathbf{x} - \mathbf{y}\| = o(\|\mathbf{x} - \boldsymbol{\theta}_0\|), \quad \forall \mathbf{x} \in C_{\boldsymbol{\Omega}}. \end{aligned}$$

When the model is identifiable, with further regularity conditions (see Section 1 in Self and Liang (1987) for details), the following asymptotic distribution of the LRT statistic is derived.

**Theorem 1** (Self and Liang, 1987). *Let  $\mathbf{Z}$  be a random variable with a multivariate Gaussian distribution with mean  $\boldsymbol{\theta}_0$  and covariance matrix  $I^{-1}(\boldsymbol{\theta}_0)$ , where  $I(\boldsymbol{\theta}) = N^{-1}I_N(\boldsymbol{\theta})$  and  $I_N(\boldsymbol{\theta})$  is the second derivative of the log-likelihood function  $l_N(\boldsymbol{\theta})$ . Let  $C_{\boldsymbol{\Omega}_0}$  and  $C_{\boldsymbol{\Omega}}$  be non-empty cones approximating  $\boldsymbol{\Omega}_0$  and  $\boldsymbol{\Omega}$  at  $\boldsymbol{\theta}_0$ , respectively. Then the asymptotic distribution of the likelihood ratio test statistic is the same as the distribution of the likelihood ratio test of  $\boldsymbol{\theta} \in C_{\boldsymbol{\Omega}_0}$  versus the alternative  $\boldsymbol{\theta} \in C_{\boldsymbol{\Omega}}$  based on a single realization  $\mathbf{Z}$  when  $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ .*

Following Self and Liang (1987), the asymptotic representation of the LRT statistic given by Theorem 1 can be written as

$$\sup_{\boldsymbol{\theta} \in C_{\Omega} - \boldsymbol{\theta}_0} \{-(\mathbf{Z} - \boldsymbol{\theta})^\top I(\boldsymbol{\theta}_0)(\mathbf{Z} - \boldsymbol{\theta})\} - \sup_{\boldsymbol{\theta} \in C_{\Omega_0} - \boldsymbol{\theta}_0} \{-(\mathbf{Z} - \boldsymbol{\theta})^\top I(\boldsymbol{\theta}_0)(\mathbf{Z} - \boldsymbol{\theta})\}, \quad (3)$$

where  $\mathbf{Z}$  has a multivariate Gaussian distribution with mean  $\mathbf{0}$  and covariance matrix  $I^{-1}(\boldsymbol{\theta}_0)$ . We can further rewrite it as

$$\inf_{\boldsymbol{\theta} \in \tilde{C}_0} \|\tilde{\mathbf{Z}} - \boldsymbol{\theta}\|^2 - \inf_{\boldsymbol{\theta} \in \tilde{C}} \|\tilde{\mathbf{Z}} - \boldsymbol{\theta}\|^2, \quad (4)$$

where  $\tilde{C} = \{\tilde{\boldsymbol{\theta}} : \tilde{\boldsymbol{\theta}} = \Lambda^{1/2} P^\top \boldsymbol{\theta}, \forall \boldsymbol{\theta} \in C_{\Omega} - \boldsymbol{\theta}_0\}$ ,  $\tilde{C}_0 = \{\tilde{\boldsymbol{\theta}} : \tilde{\boldsymbol{\theta}} = \Lambda^{1/2} P^\top \boldsymbol{\theta}, \forall \boldsymbol{\theta} \in C_{\Omega_0} - \boldsymbol{\theta}_0\}$ ,  $\tilde{\mathbf{Z}}$  follows a multivariate Gaussian distribution with mean  $\mathbf{0}$  and the identity covariance matrix, and  $P \Lambda P^\top$  represents the spectral decomposition of  $I(\boldsymbol{\theta}_0)$ . Therefore, after the orthogonal transformation, the distribution in equation (3) can be computed using the standard Gaussian distribution.

This result provides a promising direction for the hypothesis testing in the hierarchical CDM setting, and we consider a simple example in Example 10.

**Example 10.** Consider a DINA model with two latent attributes. Suppose that we want to test whether the first attribute is a prerequisite for the second attribute, that is, the hierarchical structure  $\mathcal{E}_0 = \{1 \rightarrow 2\}$ . Assume that the identifiability conditions in Proposition 1 are satisfied. The model parameters include the proportion parameters and item parameters  $\{p_{(0,0)}, p_{(0,1)}, p_{(1,0)}, p_{(1,1)}, s_j, g_j, j = 1, \dots, J\}$ , so the total number of parameters is  $3 + 2 \times J$ , noting that the proportion parameter vector  $\mathbf{p} = (p_{(0,0)}, p_{(0,1)}, p_{(1,0)}, p_{(1,1)})$  lies in the 3-simplex. To test the hierarchy  $\mathcal{E}_0$ , it is equivalent to test

$$H_0 : p_{(0,1)} = 0 \quad \text{vs} \quad H_1 : p_{(0,1)} \neq 0.$$

Therefore we have one parameter of interest that has true value on the boundary and  $2 + 2 \times J$  nuisance parameters with true values not on the boundary. After an orthogonal transformation, we have  $\tilde{C} = [0, \infty) \times \mathbb{R}^{2+2 \times J}$  and  $\tilde{C}_0 = \{0\} \times \mathbb{R}^{2+2 \times J}$  and thus the asymptotic distribution of the LRT statistic is reduced to

$$\tilde{Z}_1^2 \cdot I(\tilde{Z}_1 > 0),$$

where  $\tilde{Z}_1$  follows a standard univariate gaussian distribution. Therefore the limiting distribution of the LRT statistic is a mixture of Chi-squared distribution  $\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2$ .

In Example 10, we derive the closed form of the limiting distribution of the LRT statistic in the DINA model with two latent attributes and a linear hierarchy. In this example, we take the advantage of the fact that there is only one boundary parameter and it occurs as the parameter of interest. However, the asymptotic distribution of the LRT statistic in fact becomes considerably more complicated if there are more latent attributes and more complex hierarchical structures. Moreover, even in the simple setting as in Example 10, the convergence may be very slow if the number of items  $J$  is large or the guessing parameter  $g_j$  and slipping parameter  $s_j$  are close to the boundary, as illustrated in Figure 3. Specifically, in Figure 3, we present the p-values under various settings for Example 10. The observed p-values are plotted by blue points, and the p-values for the reference distribution  $\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2$  are plotted as the red lines. The first row in Figure 3 contains three plots of p-values with the same sample size and item parameters but different numbers of items. It is noted that when the number of items was small, the observed p-values were very close to those of the mixture Chi-squared limiting distribution.

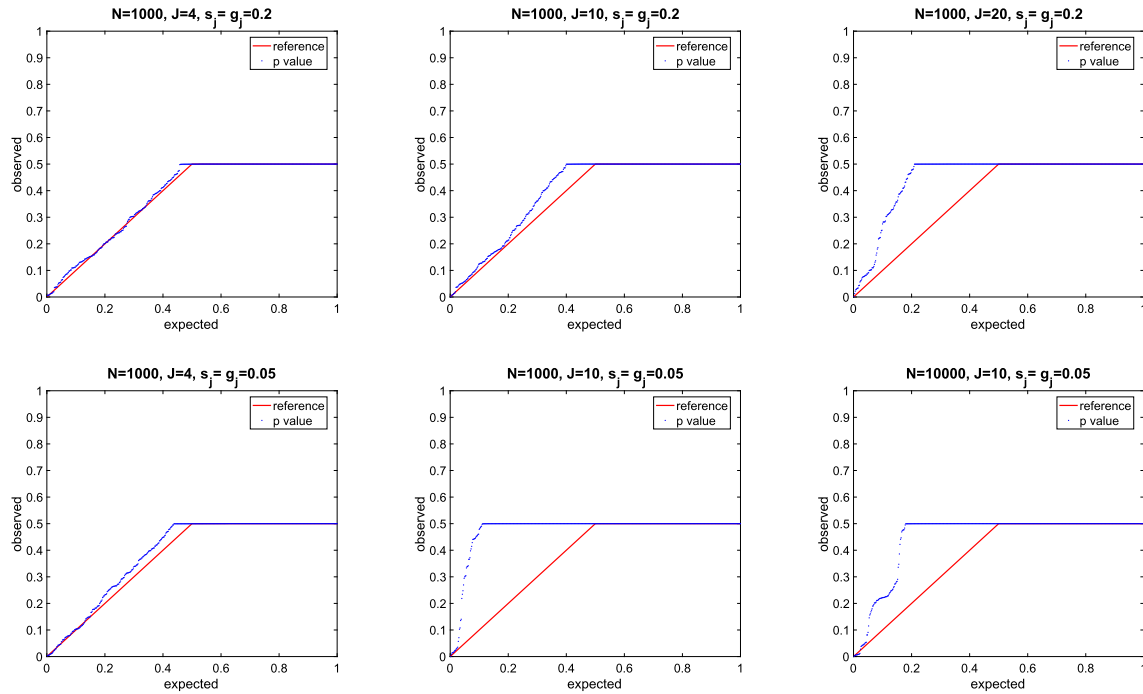


Figure 3: QQ-plots for Example 10 under various settings. The x-axis is the expected percentile of the p-values under the null hypothesis, and the y-axis is the percentile of the observed p-values. The observed p-values are plotted by blue points, and the p-values for the reference limiting distribution  $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$  are plotted as red lines. If the blue points are close to the red lines, it indicates that the empirical distribution of the observed p-values approximates the asymptotic distribution well.

However, as the number of items increased, the gap between the observed p-values and the reference limiting distribution became larger. The second row in Figure 3 contains three plots of p-values with more extreme item parameters. Compared with the plots in the first row, it is shown that when the item parameters were close to the boundary, the convergence of the LRT statistic became much slower, and such testing tended to fail even with a large sample size  $N = 10,000$ .

As pointed out in Self and Liang (1987), even though based on Theorem 1 we can derive the asymptotic distribution of the LRT statistic for any fixed  $\theta_0 \in \Omega_0$ , this distribution is generally different for different  $\theta_0$ . Moreover, these distributions typically vary over  $\Omega_0$  in a discontinuous way when some of the nuisance parameters may also be on the boundary. This discontinuity can affect the quality of the asymptotic approximation much. As in our example, when the slipping parameter  $s_j$  and the guessing parameter  $g_j$  in the DINA model are close to the boundary, the distribution with a given finite sample may be far away from the weighted Chi-squared mixture as described in Example 10.

We provide further insights on why the convergence was slow when the number of items was large or the item parameters were close to the boundaries as shown in Figure 3. The average log-likelihood of CDMs is given by  $l_N/N := \sum_{i=1}^N \log \left( \sum_{\alpha} p_{\alpha} P(\mathbf{R}_i | \alpha) \right) / N$ . Consider Example 10 where we are interested in the hierarchy  $\{1 \rightarrow 2\}$  and want to test whether  $p_{(0,1)} = 0$ . If we write  $p_{(0,0)} = 1 - \sum_{\alpha \neq (0,0)} p_{\alpha}$ , then  $l_N/N = \sum_{i=1}^N \log((1 - \sum_{\alpha \neq (0,0)} p_{\alpha}) P(\mathbf{R}_i | \alpha = (0,0)) +$

$\sum_{\alpha \neq (0,0)} p_\alpha P(\mathbf{R}_i | \alpha) / N$ . The derivative of the log-likelihood w.r.t.  $p_{(0,1)}$  becomes

$$\begin{aligned} \frac{\partial l_N / N}{\partial p_{(0,1)}} &= \frac{1}{N} \sum_{i=1}^N \frac{P(\mathbf{R}_i | \alpha = (0, 1)) - P(\mathbf{R}_i | \alpha = (0, 0))}{\sum_{\alpha} p_\alpha P(\mathbf{R}_i | \alpha)} \\ &= \frac{1}{N} \sum_{i=1}^N \frac{P(\mathbf{R}_i | \alpha = (0, 1)) - P(\mathbf{R}_i | \alpha = (0, 0))}{P(\mathbf{R}_i)} \\ &= \frac{1}{N} \sum_{i=1}^N \sum_{\mathbf{r} \in \{0,1\}^J} I(\mathbf{R}_i = \mathbf{r}) \frac{P(\mathbf{r} | \alpha = (0, 1)) - P(\mathbf{r} | \alpha = (0, 0))}{P(\mathbf{r})}. \end{aligned} \quad (5)$$

When the null hypothesis that  $p_{(0,1)} = 0$  is true, by the strong law of large number, we have

$$\begin{aligned} \left. \frac{\partial l_N / N}{\partial p_{(0,1)}} \right|_{p_{(0,1)}=0} &\xrightarrow{a.s.} \mathbb{E}_0 \left[ \sum_{\mathbf{r} \in \{0,1\}^J} I(\mathbf{R} = \mathbf{r}) \frac{P(\mathbf{r} | \alpha = (0, 1)) - P(\mathbf{r} | \alpha = (0, 0))}{P(\mathbf{r})} \right] \\ &= \sum_{\mathbf{r} \in \{0,1\}^J} \left( P(\mathbf{r} | \alpha = (0, 1)) - P(\mathbf{r} | \alpha = (0, 0)) \right) \\ &= \sum_{\mathbf{r} \in \{0,1\}^J} P(\mathbf{r} | \alpha = (0, 1)) - \sum_{\mathbf{r} \in \{0,1\}^J} P(\mathbf{r} | \alpha = (0, 0)) \\ &= 0. \end{aligned}$$

However, since the number of possible response patterns is  $|\{0, 1\}^J| = 2^J$  which grows exponentially with the number of items  $J$ , it requires an exponentially growing sample size to cover all the possible response patterns, and therefore the convergence can be slow when  $J$  is large.

Next, consider the case when the item parameters are close to the boundary. Note that

$$P(\mathbf{R} | \alpha) = \prod_{j=1}^J P(R_j | \alpha) = \prod_{j=1}^J \left( g_j^{1-\Gamma_{j,\alpha}} (1 - s_j)^{\Gamma_{j,\alpha}} \right)^{R_j} \left( (1 - g_j)^{1-\Gamma_{j,\alpha}} s_j^{\Gamma_{j,\alpha}} \right)^{1-R_j}.$$

When the item parameters are very close to the boundaries, that is,  $s_j$  and  $g_j$  are very close to 0, the model becomes near deterministic. For simplicity, let  $s_j = g_j = \delta$  which is very close to 0 for all  $j \in [J]$ . Then

$$\begin{aligned} P(\mathbf{R} = \mathbf{r} | \alpha) &= \prod_{j=1}^J \left( \delta^{(1-\Gamma_{j,\alpha})} (1 - \delta)^{\Gamma_{j,\alpha}} \right)^{r_j} \left( (1 - \delta)^{(1-\Gamma_{j,\alpha})} \delta^{\Gamma_{j,\alpha}} \right)^{1-r_j} \\ &= \prod_{r_j=1} \delta^{(1-\Gamma_{j,\alpha})} (1 - \delta)^{\Gamma_{j,\alpha}} \prod_{r_j=0} (1 - \delta)^{(1-\Gamma_{j,\alpha})} \delta^{\Gamma_{j,\alpha}} \\ &= \delta^{\sum_{r_j=1} (1-\Gamma_{j,\alpha}) + \sum_{r_j=0} \Gamma_{j,\alpha}} \cdot (1 - \delta)^{\sum_{r_j=1} \Gamma_{j,\alpha} + \sum_{r_j=0} (1-\Gamma_{j,\alpha})}. \end{aligned}$$

For  $\mathbf{r} = \mathbf{\Gamma}_{\cdot, \alpha}$ , we have  $P(\mathbf{R} = \mathbf{\Gamma}_{\cdot, \alpha} | \alpha) = (1 - \delta)^J$ . And for any  $\mathbf{r} \neq \mathbf{\Gamma}_{\cdot, \alpha}$ ,  $\delta^J \leq P(\mathbf{R} = \mathbf{r} | \alpha) \leq \delta$ . Moreover, when  $p_{(0,1)} = 0$ , since  $P(\mathbf{R} = \mathbf{r}) = \sum_{\alpha \neq (0,1)} p_\alpha P(\mathbf{R} = \mathbf{r} | \alpha)$ , we have

$$P(\mathbf{R} = \mathbf{\Gamma}_{\cdot, \alpha}) \geq p_\alpha (1 - \delta)^J, \text{ for } \alpha \in \mathcal{A} = \{(0, 0), (1, 0), (1, 1)\},$$

$$P(\mathbf{R} \neq \mathbf{\Gamma}_{\cdot, \alpha}, \alpha \in \mathcal{A}) \leq 1 - (1 - \delta)^J \rightarrow 0 \text{ as } \delta \rightarrow 0.$$



Therefore from the above discussions, when  $p_{(0,1)} = 0$ , the probability mass is concentrated around three response patterns  $\Gamma_{\cdot,\alpha}$  for  $\alpha \in \mathcal{A} = \{(0, 0), (1, 0), (1, 1)\}$ . For the terms in the RHS of (5), when  $\mathbf{r} = \Gamma_{\cdot,(0,0)}$ , we have

$$\begin{aligned} & [P(\mathbf{R} = \mathbf{r} \mid \alpha = (0, 1)) - P(\mathbf{R} = \mathbf{r} \mid \alpha = (0, 0))](P(\mathbf{R} = \mathbf{r}))^{-1} \\ & \in \left[ \frac{\delta^J - (1 - \delta)^J}{(1 - \delta)^J p_{(0,0)}}, \frac{\delta - (1 - \delta)^J}{(1 - \delta)^J p_{(0,0)} + \delta(1 - p_{(0,0)})} \right] \rightarrow -1/p_{(0,0)} \text{ as } \delta \rightarrow 0. \end{aligned}$$

When  $\mathbf{r} = \Gamma_{\cdot,(1,0)}$  (or  $\Gamma_{\cdot,(1,1)}$ ), we have

$$\begin{aligned} & [P(\mathbf{R} = \mathbf{r} \mid \alpha = (0, 1)) - P(\mathbf{R} = \mathbf{r} \mid \alpha = (0, 0))](P(\mathbf{R} = \mathbf{r}))^{-1} \\ & \in \left[ \frac{\delta^J - \delta}{(1 - \delta)^J p_{(1,0)}}, \frac{\delta - \delta^J}{(1 - \delta)^J p_{(1,0)} + \delta(1 - p_{(1,0)})} \right] \rightarrow 0 \text{ as } \delta \rightarrow 0. \end{aligned}$$

Therefore the terms in the RHS of (5) also concentrate around two points,  $-1/p_{(0,0)}$  and 0, making the convergence slow since more data points are needed to have it converge to 0.

Based on the above discussions about the asymptotic behaviors of the LRT statistic under the nonstandard conditions and in the hierarchical CDM setting, it has been shown that even in the simple setting where we could derive a closed form of the limiting distribution, the convergence can be very slow. Moreover, the asymptotic distribution of the LRT will be much more complicated if we have more latent attributes and more complex hierarchical structures. Therefore, it is not practical to use the theoretical limiting distribution of the LRT statistic to test latent hierarchical structures in CDMs, especially considering that the number of test items is usually relatively large (e.g. more than 20).

## 4.2 Bootstrap and Numerical Studies

From the discussions about the LRT in Section 4.1, we learn that the limiting distribution of the LRT statistic under latent hierarchical structures can be very complicated and the convergence can be slow when the number of items is large or the item parameters are close to the boundary even in simple settings. To overcome these difficulties, we propose to use the bootstrap method as an alternative to the asymptotic limiting distribution method. The bootstrap method (Efron, 1979) has been shown to be successful in many nonstandard situations. The basic idea of bootstrap is treating inference of the true probability distribution, given the original data, as being analogous to the inference of the empirical distribution, given the resampled data. If the empirical distribution is a reasonable approximation to the true distribution, then the bootstrap method will provide good inferences.

In this section, we consider two different bootstrap procedures: nonparametric bootstrap and parametric bootstrap. The idea of nonparametric bootstrap is to simulate data from the empirical distribution by directly resampling from the original data. To be specific, in nonparametric bootstrap, we draw samples of the same size from the original data with replacement. Then the statistic of interest is computed based on the resampled data set and we repeat this routine many times. The steps for nonparametric bootstrap are summarized as below:

- Step 1. Initially estimate the model with the specified hierarchy under the null hypothesis, and the model under the alternative hypothesis (without the null hypothesis hierarchy constraints), and calculate the LRT statistic.
- Step 2. Draw a sample of the same size with replacement from the original data and calculate the LRT statistic.

- Step 3. Repeat Step 2 independently many times and estimate the distribution of the LRT statistic.
- Step 4. Estimate the p-value by comparing the distribution obtained in Step 3 with the LRT statistic obtained in Step 1. Then this p-value is used to determine whether the null model with the specified null hierarchy should be rejected in favor of the model without the hierarchical constraints.

The idea of parametric bootstrap is to simulate data based on good estimates of distribution parameters, often by maximum likelihood. In parametric bootstrap, a parametric model is fitted to the original data, and samples are drawn from this fitted model. The steps for parametric bootstrap are similar to those of nonparametric bootstrap except for Step 2:

Step 2\*. Based on the estimates of the model with specified hierarchy from step 1, generate a bootstrap sample from the fitted model and calculate the LRT statistic.

Next we conduct comprehensive simulation studies to compare parametric bootstrap and nonparametric bootstrap for testing latent hierarchical structures under various settings. We considered four different hierarchical structures shown in Figure 2. For the data generating process, we considered the DINA model and the GDINA model respectively. For both models we included three different sample sizes ( $N = 200, 500, \text{ or } 1000$ ) and the number of items was set to 30 ( $J = 30$ ). In terms of uncertainty, two levels of guessing and slipping parameters in the DINA model were included ( $s_j = g_j = 0.1 \text{ or } 0.2$  for  $j \in [J]$ ). For the GDINA model, we also considered two different uncertainty levels, where the highest item parameter was 0.9 or 0.8, and the lowest item parameter was 0.1 or 0.2. The other item parameters in between were equally spaced. To satisfy testability conditions, the  $Q$ -matrix contained two identity sub-matrices and the remaining items were randomly generated. For each scenario, we performed 500 independent repetitions and in each repetition we generated bootstrap samples for 500 times. To fit the models under the null and alternative hypotheses, we used R package “CDM”.

The type I errors with significance level  $\alpha = 0.05$  under different settings are plotted in the supplementary material. There we also provide corresponding error bars for uncertainty quantification of the Monte Carlo errors. The naïve Chi-squared test is included for comprehensive comparison. From the plots, we can see that the type I errors for parametric bootstrap were around 0.05 in most cases and therefore parametric bootstrap controlled the type I errors generally well. By contrast, nonparametric bootstrap was too conservative and the type I errors for nonparametric bootstrap were very close to 0. In terms of the naïve Chi-squared test, it was also very conservative in most cases even though the type I errors for the GDINA model with larger noises under the “unstructured” hierarchy were closer to the significance level 0.05.

To further examine the behaviors of the testing procedures, the QQ plots for p-values under the null hypothesis are provided in Figure 4 and Figure 6. For presentation brevity, we only present results of four hierarchies with the same noise level and sample size. More comprehensive simulation results are presented in the supplementary material. It is known that under the null hypothesis, the p-values should follow a uniform distribution on  $[0, 1]$ . In the QQ-plots, if the points are lying closer to the identity line, it indicates that it approximates the uniform distribution better. From Figure 4 and 6, one can see that under the null hypothesis, the p-values of parametric bootstrap approximated the uniform distribution on  $[0, 1]$  very well in almost all the settings. By contrast, the p-values of nonparametric bootstrap and the naïve Chi-squared test were far away from the uniform distribution, indicating that these testing procedures are not reliable. We also conducted power analysis where all the latent attribute profiles existed in the data generation process. The true proportion parameters were equally assigned. The QQ plots for p-values under the alternative hypothesis are shown in Figure 5 and

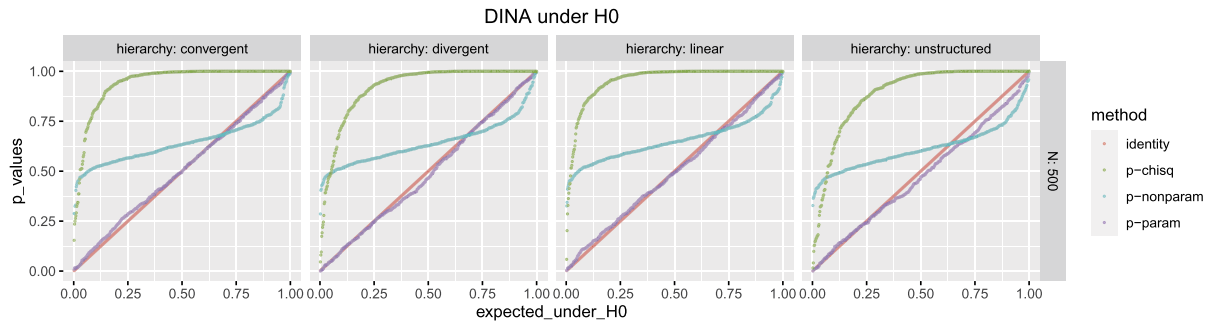


Figure 4: QQ-plots for the p-values of the DINA model under the null hypothesis where  $\theta_j^+ = 0.8$ ,  $\theta_j^- = 0.2$  that corresponds to the case with high noises.

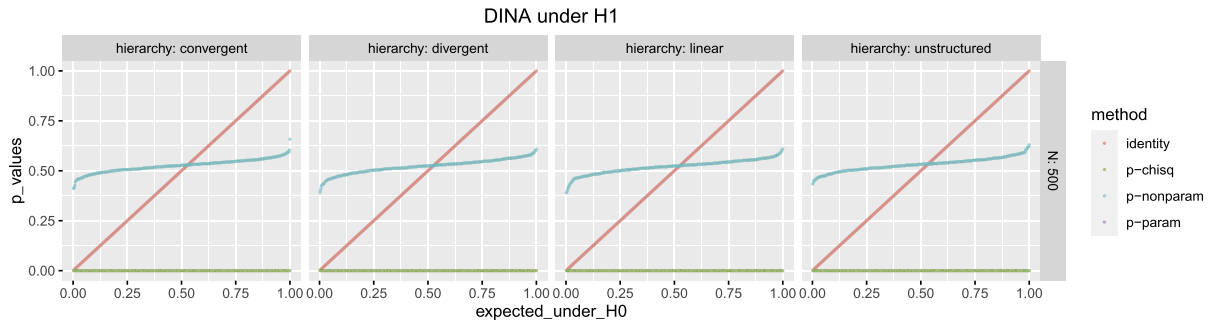


Figure 5: QQ-plots for the p-values of the DINA model under the alternative hypothesis where  $\theta_j^+ = 0.8$ ,  $\theta_j^- = 0.2$  corresponding to high noises. The expected quantiles on the x-axis are the expected quantiles of the p-values under the null hypothesis, i.e., uniform distribution on  $[0,1]$ .

Figure 7 respectively. To have more power, we expect the p-values to be small so that we would reject the null hypothesis. Therefore in the QQ-plots, the closer to 0 the points are, the more powerful the test is. From Figure 5 and 7, one can see that the p-values of parametric bootstrap and the naïve Chi-squared test were almost 0 and therefore the power was close to 1. However, the p-values of nonparametric bootstrap were close or above 0.5, which means we would not reject the null hypothesis, making the power almost 0. Taking both the type I error and power into consideration, the parametric bootstrap outperformed the other two testing procedures.

In order for nonparametric bootstrap to work, the empirical distribution of the sample data should be close to the true distribution, which may not hold for the cases here especially when the number of items is relatively large. As we discussed in Section 4.1, the total number of possible response patterns  $2^J$  grows exponentially with the number of items  $J$ . Therefore, in nonparametric bootstrap, we need a very large sample size to cover all the possible response patterns, which may explain the failures of nonparametric bootstrap. On the contrary, in parametric bootstrap, we first fit a model from the original data and resample data from the fitted model, which incorporates the variance in data generation better and thus makes parametric bootstrap perform better. Note that in Templin and Bradshaw (2014), using the naïve Chi-squared test, the authors concluded that “the DINA model cannot detect attribute hierarchies”. However, through our comprehensive simulation, by using parametric bootstrap, the attribute hierarchies can also be well detected in the DINA model.

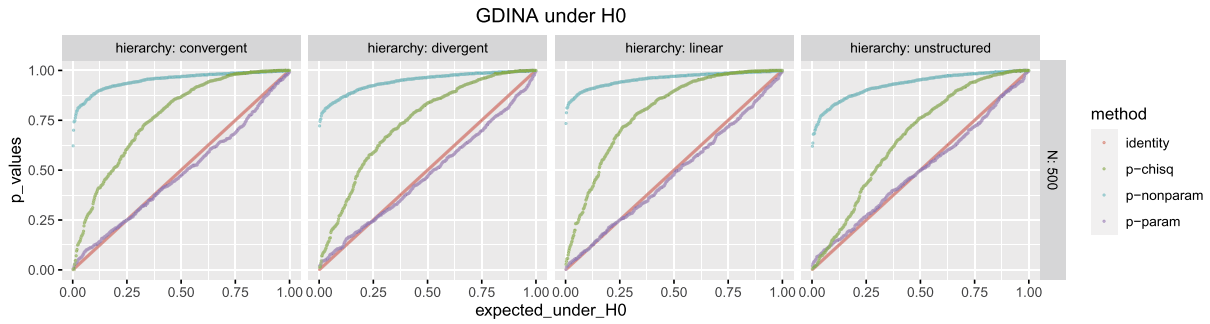


Figure 6: QQ-plots for the p-values of the GDINA model under the null hypothesis where  $\theta_j^+ = 0.8$ ,  $\theta_j^- = 0.2$  corresponding to the case with high noises.

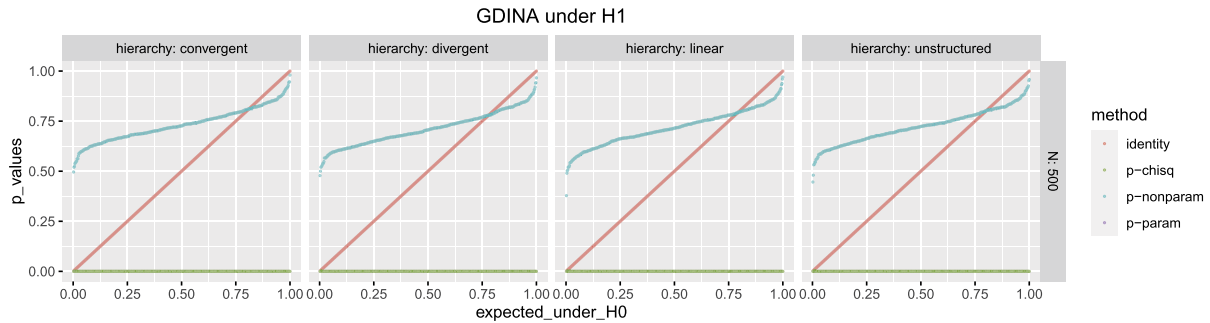


Figure 7: QQ-plots for the p-values of the GDINA model under the alternative hypothesis where  $\theta_j^+ = 0.8$ ,  $\theta_j^- = 0.2$  corresponding to high noises. The expected quantiles on the x-axis are the expected quantiles of the p-values under the null hypothesis, i.e., uniform distribution on  $[0,1]$ .

In summary, based on our discussions about the failure of the limiting distribution of LRT in Section 4.1 and the simulation results in Section 4.2, we recommend using parametric bootstrap to perform hypothesis testing for latent hierarchical structures in CDMs.

## 5 Real Data Analysis

In this section, we perform hypothesis testing procedures on the Examination for the Certificate of Proficiency in English (ECPE) data. The ECPE data is collected by the English Language Institute of the University of Michigan, in which there are 2,922 test takers and 28 ECPE test items. Three target attributes are considered, including morphosyntactic rules ( $\alpha_1$ ), cohesive rules ( $\alpha_2$ ) and lexical rules ( $\alpha_3$ ). The  $Q$ -matrix of the ECPE data is given in the supplementary material. Since the  $Q$ -matrix contains four identity submatrices, the testability conditions are satisfied. A linear hierarchical structure  $\mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  is often considered in literature such as Templin and Bradshaw (2014).

Under the linear hierarchy  $\mathcal{E}_0$ , the latent attribute profile set is  $\mathcal{A}_0 = \{(0, 0, 0), (0, 0, 1), (0, 1, 1), (1, 1, 1)\}$ . Under the null hypothesis, we fit a GDINA model with the profile set  $\mathcal{A}_0$ , and under the alternative, we fit a saturated GDINA model with all the possible attribute profiles. We generated bootstrap samples 1000 times in parametric bootstrap and nonparametric bootstrap

Table 1: p-values for different testing settings.

Setting	Para-boot	Nonpara-boot	Chi-squared
$\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$ vs. $\emptyset$	0.041	0.952	0.020
$\{\alpha_3 \rightarrow \alpha_2\}$ vs. $\emptyset$	0.052	0.511	0.072
$\{\alpha_2 \rightarrow \alpha_1\}$ vs. $\emptyset$	0.057	0.722	0.064
$\{\alpha_3 \rightarrow \alpha_1\}$ vs. $\emptyset$	0.169	0.954	0.066
$\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$ vs. $\{\alpha_3 \rightarrow \alpha_2\}$	0.098	0.962	0.047
$\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$ vs. $\{\alpha_2 \rightarrow \alpha_1\}$	0.073	0.906	0.052
$\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$ vs. $\{\alpha_3 \rightarrow \alpha_1\}$	0.026	0.625	0.051

respectively. The p-value obtained from parametric bootstrap is 0.041, while the p-value obtained from nonparametric bootstrap is 0.952. Moreover, we also calculate the p-value corresponding to the naïve test using the conventional Chi-squared limiting distribution and get the p-value 0.02. If we set the significance level to be 0.05, then by parametric bootstrap, we do reject the null hypothesis and conclude the linear hierarchy does not present in this data set; while if the significance level is set to be 0.01, we would not reject the null hypothesis and conclude there is such a linear attribute hierarchy. This conclusion is consistent with that in Templin and Bradshaw (2014). For both significance levels, the nonparametric bootstrap does not reject the null hypothesis.

To conduct a more comprehensive study of the linear hierarchies among the three target attributes, we further test each linear hierarchy relationship separately and examine which one is the strongest. In particular, we consider the following various test settings:

- $H_0 : \mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_2\}$  vs.  $H_1$ : no hierarchical structure  $\mathcal{E}_1 = \emptyset$ ;
- $H_0 : \mathcal{E}_0 = \{\alpha_2 \rightarrow \alpha_1\}$  vs.  $H_1$ : no hierarchical structure  $\mathcal{E}_1 = \emptyset$ ;
- $H_0 : \mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_1\}$  vs.  $H_1$ : no hierarchical structure  $\mathcal{E}_1 = \emptyset$ ;
- $H_0 : \mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $H_1 : \mathcal{E}_1 = \{\alpha_3 \rightarrow \alpha_2\}$ ;
- $H_0 : \mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $H_1 : \mathcal{E}_1 = \{\alpha_2 \rightarrow \alpha_1\}$ ;
- $H_0 : \mathcal{E}_0 = \{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $H_1 : \mathcal{E}_1 = \{\alpha_3 \rightarrow \alpha_1\}$ ;

The resulting p-values for parametric bootstrap, nonparametric bootstrap and the naïve Chi-squared test under different settings are presented in Table 1. From the table, we can see that among all the settings, the p-values of nonparametric bootstrap are very large and therefore we do not reject the null hypotheses. This is also consistent with our simulation study where we find nonparametric bootstrap is more conservative. The p-values of parametric bootstrap are much smaller than those of nonparametric bootstrap. If we set the significance level to be 0.05, parametric bootstrap does not reject the null except for the settings “ $\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $\emptyset$ ” and “ $\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $\{\alpha_3 \rightarrow \alpha_1\}$ ”. The p-values of the naïve Chi-squared test are of the similar scales of those in parametric bootstrap, and if the significance level is set to 0.05, the naïve Chi-squared test does not reject the null except for the settings “ $\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $\{\alpha_3 \rightarrow \alpha_2\}$ ” and “ $\{\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1\}$  vs.  $\emptyset$ ”. However, as we have shown in our simulation results, since the true limiting distribution of the LRT statistic is no longer the conventional Chi-squared distribution, the naïve test is not reliable.

For testing a single linear hierarchy relationship versus none hierarchical structure (i.e., the second to the fourth tests in Table 1), since the alternative hypothesis is the same while the null hypothesis varies, a larger p-value suggests a stronger pre-requisite relationship. Therefore

based on the results in Table 1 that the p-value for “ $\{\alpha_3 \rightarrow \alpha_1\}$  vs.  $\emptyset$ ” is the largest among these three tests, we can see that the prerequisite relationship between the third attribute and the first attribute is the strongest. Similarly, for testing the whole linear hierarchical  $\mathcal{E}_0$  versus a single linear hierarchy (i.e., the last three tests in Table 1), since now the null hypothesis is the same and the alternative hypothesis varies, a smaller p-value indicates a stronger pre-requisite relationship. In this case, the prerequisite relationship between the third attribute and the first attribute seems still the strongest.

## 6 Discussions

In this paper, we consider the hypothesis testing problem for latent hierarchical structures in CDMs. We first discuss the testability issues and present sufficient conditions for testability. Under the testability conditions, we study asymptotic properties of the likelihood ratio test and show the practical difficulties of directly using the limiting distribution of the LRT statistic to test latent hierarchies. We then compare two resampling-based testing procedures including parametric bootstrap and nonparametric bootstrap through comprehensive simulations under different settings and recommend using parametric bootstrap for testing latent hierarchical structures.

In this paper, we mainly focus on the hypothesis testing where the hierarchical structure is fully specified and all the latent attribute profiles that respect the hierarchy exist in the population. In many applications, the number of latent attributes  $K$  could be large, leading to a high-dimensional space for all the possible configurations of the attributes, where the number of potential attribute profiles can be even larger than the sample size. For scientific interpretability and practical use, it is often assumed that not all the possible attribute profiles exist in the population. In such cases, to test hierarchical structures, we may perform the selection of significant latent attribute profiles first and then conduct testing procedures.

This paper proposes to use parametric bootstrap, which is a resampling-based procedure and can be computationally expensive, especially for large-scale data sets. Therefore it would be useful to develop more efficient testing procedures. Moreover, further theoretical results are needed to better characterize the asymptotic distribution of the likelihood ratio test with the presence of latent variables and complex constraint structures as in CDMs.

## Supplementary Material

More comprehensive simulation results are presented in the supplementary material. Specifically, bootstrap results for DINA and GDINA models under both null hypothesis and alternative hypothesis with different sample sizes and noise levels are plotted there. We also include the codes for simulations and real data analysis.

## Acknowledgment

The authors thank Dr. Yuqi Gu for helpful discussions.

## Funding

This research is partially supported by National Science Foundation CAREER SES-1846747 and Institute of Education Sciences R305D200015.



## References

- Chen J (2017). On finite mixture models. *Statistical Theory and Related Fields*, 1(1): 15–27.
- Chen Y, Moustaki I, Zhang H (2020). A note on likelihood ratio tests for models with latent variables. *Psychometrika*, 85: 1–17.
- Dahlgren MA, Hult H, Dahlgren LO, af Segerstad HH, Johansson K (2006). From senior student to novice worker: Learning trajectories in political science, psychology and mechanical engineering. *Studies in Higher Education*, 31(5): 569–586.
- de la Torre J (2011). The generalized DINA model framework. *Psychometrika*, 76(2): 179–199.
- de la Torre J, van der Ark LA, Rossi G (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 51(4): 281–296.
- DiBello LV, Stout WF, Roussos LA (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In: Edited by Paul D. Nichols, Susan F. Chipman, Robert L. Brennan, *Cognitively Diagnostic Assessment*, 361–389. Routledge.
- Efron B (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, 7(1): 1–26.
- George AC, Robitzsch A (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology*, 11(3): 189–205.
- Gu Y, Xu G (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(2019): 1–58.
- Gu Y, Xu G (2020). Partial identifiability of restricted latent class models. *Annals of Statistics*, 48(4): 2082–2107.
- Gu Y, Xu G (2021). Identifiability of hierarchical latent attribute models. *arXiv preprint arXiv:1906.07869*.
- Haertel EH (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26(4): 301–321.
- Henson RA, Templin JL, Willse JT (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2): 191.
- Jimoyiannis A, Komis V (2001). Computer simulations in physics teaching and learning: A case study on students’ understanding of trajectory motion. *Computers & Education*, 36(2): 183–204.
- Junker BW, Sijtsma K (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3): 258–272.
- Leighton JP, Gierl MJ, Hunka SM (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka’s rule-space approach. *Journal of Educational Measurement*, 41(3): 205–237.
- Nylund KL, Asparouhov T, Muthén BO (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4): 535–569.
- O’Brien KL, Baggett HC, Brooks WA, Feikin DR, Hammitt LL, Higdon MM, et al. (2019). Causes of severe pneumonia requiring hospital admission in children without HIV infection from Africa and Asia: The PERCH multi-country case-control study. *The Lancet*, 394(10200): 757–779.
- Reckase MD (2009). Multidimensional item response theory models. In: *Multidimensional Item Response Theory*, 79–112. Springer.

- Self SG, Liang KY (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398): 605–610.
- Simon MA, Tzur R (2004). Explicating the role of mathematical tasks in conceptual learning: An elaboration of the hypothetical learning trajectory. *Mathematical Thinking and Learning*, 6(2): 91–104.
- Tatsuoka KK (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20: 345–354.
- Tatsuoka KK (1990). Toward an integration of item-response theory and cognitive error diagnosis. In: Edited by Norman Frederiksen, Robert Glaser, Alan Lesgold, and Michael G. Shafto, *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 453–488. Routledge.
- Templin J, Bradshaw L (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2): 317–339.
- Templin JL, Henson RA (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11(3): 287.
- von Davier M (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2): 1–35.
- Wang C, Gierl MJ (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.
- Wu Z, Deloria-Knoll M, Hammitt LL, Zeger SL (for Child Health Core Team PER) (2016a). Partially latent class models for case–control studies of childhood pneumonia etiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1): 97–114.
- Wu Z, Deloria-Knoll M, Zeger SL (2016b). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2): 200–213.
- Xu G (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45(2): 675–707.
- Xu G, Shang Z (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523): 1284–1295.
- Xu G, Zhang S (2016). Identifiability of diagnostic classification models. *Psychometrika*, 81(3): 625–649.