Statistica Sinica

Sufficient and Necessary Conditions for the Identifiability of the Q-matrix

Yuqi Gu and Gongjun Xu

University of Michigan

Abstract: Restricted latent class models (RLCMs) have recently gained prominence in educational assessment, psychiatric evaluation, and medical diagnosis. In contrast to conventional latent class models, the restrictions on RLCM parameters are imposed using a design matrix, in order to respect practitioners' scientific assumptions. The design matrix, called the Q-matrix in the cognitive diagnosis literature, is usually constructed by practitioners and domain experts; however, it remains subjective and can be misspecified. To address this problem, researchers have proposed estimating the Q-matrix from sample data. However, the fundamental learnability of the Q-matrix and the model parameters remains underexplored. As a result, studies often impose stronger than needed (or even impractical) conditions. Here, we propose sufficient and necessary conditions for the joint identifiability of the Q-matrix and the RLCM parameters under different types of RLCMs. The proposed identifiability conditions depend only on the design matrix, and are easy to verify in practice.

Key words and phrases: Identifiability; restricted latent class models; cognitive diagnosis.

1. Introduction

Latent class models are widely used as statistical tools in social and biological sciences to model the relationship between a set of observed responses and a set of discrete latent attributes of interest. This study focuses on a family of *restricted latent class models* (RLCMs), which play a key role in, for example, cognitive diagnosis in educational assessment (e.g., Junker and Sijtsma, 2001; Henson et al., 2009; Rupp et al., 2010; de la Torre, 2011), psychiatric evaluation (Templin and Henson, 2006; Jaeger et al., 2006; de la Torre et al., 2017), online testing and learning (Wang et al., 2016; Zhang and Chang, 2016; Xu et al., 2016), and disease etiology diagnosis and scientifically structured clustering of patients (Wu et al., 2017, 2018).

In contrast to conventional latent class models, the parameters of RLCMs are constrained using a design matrix, often called the Q-matrix in the cognitive diagnosis literature (Rupp et al., 2010). The Q-matrix encodes practitioners' understanding of how the responses depend on the underlying latent attributes. Various RLCMs have been developed, each with their own cognitive diagnostic assumptions (e.g., DiBello et al., 1995; de la Torre and Douglas, 2004; Templin and Henson, 2006; von Davier, 2008; Henson et al., 2009), including the basic deterministic input noisy output "And" gate (DINA) model (Junker and Sijtsma, 2001), which serves as a basic submodel for more general cognitive diagnostic models. See Section 2 for a review of these models.

Despite the popularity of RLCMs, the fundamental identifiability of such models

remains a challenge, as noted in the literature (DiBello et al., 1995; Maris and Bechger, 2009; Tatsuoka, 2009; DeCarlo, 2011; von Davier, 2014). Existing results related to the identifiability of unrestricted latent class models in statistics (Teicher, 1967; Goodman, 1974; Gyllenberg et al., 1994; Allman et al., 2009) cannot be applied directly to RLCMs, owing to the structural constraints induced by the Q-matrix. Recent works have examined the identifiability of RLCM parameters for the basic DINA model (Chen et al., 2015; Xu and Zhang, 2016; Gu and Xu, 2018) and general RLCMs (Xu, 2017; Gu and Xu, 2019, 2020), assuming that the Q-matrix is prespecified and correct.

However, the Q-matrix, specified by scientific experts when constructing the diagnostic items, can be misspecified. Moreover, in an exploratory analysis of newly designed items, much or all of the Q-matrix may not be available. Here, a misspecification of the Q-matrix could lead to a serious lack of fit for the model, and thus inaccurate inferences on the latent attribute profiles of the individuals. Therefore, it is desirable to estimate the Q-matrix and the model parameters jointly from the response data (e.g., de la Torre, 2008; DeCarlo, 2012; Liu et al., 2012; de la Torre and Chiu, 2016; Chen et al., 2018). A reliable and valid estimation and inference on the Q-matrix requires that we ensure the joint identifiability of the Q-matrix and the associated model parameters. Such joint identifiability has been studied recently by Liu et al. (2013) and Chen et al. (2015) under the DINA model, and by Xu and Shang (2018) under general RLCMs. Nevertheless, most of these works focus on developing sufficient conditions for joint

identifiability, and thus often impose stronger than needed or sometimes impractical constraints on the experimental design of a cognitive diagnosis.

Therefore, the necessary and sufficient conditions (or minimal requirements) for the joint identifiability of the Q-matrix and the model parameters remains an open problem. This study addresses this problem, contributing to the literature as follows.

First, under the DINA model, we derive the necessary and sufficient conditions for the joint identifiability of the Q-matrix and the associated DINA model parameters. Our necessary and sufficient conditions are succinctly and neatly written as three algebraic properties of the Q-matrix, which we summarize as *completeness* (Condition A), *distinctness* (Condition B), and *repetition* (Condition C); please see Theorem 1 for details. These three conditions require that the binary Q-matrix is *complete* by containing an identity submatrix, has all columns *distinct* other than the part of the identity submatrix, and *repeatedly* contains at least three entries of one in each column. In addition to guaranteeing identifiability, these conditions give the minimal requirements for the Q-matrix and DINA model parameters to be estimable from the observed responses. The identifiability result can be applied directly to the deterministic input noisy output "Or" gate (DINO) model (Templin and Henson, 2006), owing to the duality of the DINA and DINO models (Chen et al., 2015). The derived identifiability conditions also serve as necessary requirements for joint identifiability under general RLCMs, which include the DINA model as a submodel.

Second, we propose sufficient and necessary conditions for a weaker notation of identifiability, the so-called generic identifiability, under both the DINA model and general RLCMs. Generic identifiability implies that those parameters for which identifiability does not hold live in a set of Lebesgue measure zero (Allman et al., 2009). The motivation for studying generic identifiability is that the strict identifiability conditions are sometimes too restrictive in practice. For instance, it is known that unrestricted latent class models are not strictly identifiable (Gyllenberg et al., 1994), but are generically identifiable under certain conditions (Allman et al., 2009). In RLCMs, the model parameters are forced by the Q-matrix-induced constraints to fall in a measure-zero subset of the parameter space, and, thus, existing results for unrestricted models cannot be applied directly. Moreover, the generic identifiability conditions needed to jointly identify the Q-matrix and the model parameters are unknown. Therefore, in this work, we propose sufficient and necessary conditions for generic identifiability, and explicitly characterize the nonidentifiable measure-zero subset. Our mild sufficient conditions for generic identifiability under general RLCMs can be summarized as the following properties of the Q-matrix: *double generic completeness* (Condition D), and *generic repetition* (Condition E); see Theorem 4 for details. These two conditions require that the binary Q-matrix contains two *generically complete* square submatrices with all diagonal elements equal to one, and (*repeatedly*) contains at least one entry of "1" other than the part comprising these two submatrices.

The rest of the paper is organized as follows. Section 2 introduces RLCMs and reviews some popular models in cognitive diagnosis. Section 3 defines strict and generic identifiability for RLCMs, and presents an illustrative example. Sections 4 and 5 contain our main theoretical results for strict and generic identifiability for the DINA model and general RLCMs, respectively. Section 6 concludes the paper. The proofs of the theoretical results and additional simulation studies that verify the developed theory are included in the online Supplementary Material. The Matlab code used to check the proposed conditions is available at `https://github.com/yuqigu/Identify_Q`.

2. RLCMs for cognitive diagnosis

RLCMs are key statistical tools in cognitive diagnostic assessments that estimate individuals' attribute profiles based on their response data in the assessment. Specifically, consider a diagnostic test with J items. A subject (e.g., an examinee or a patient) provides a J-dimensional binary response vector $\boldsymbol{R} = (R_1, ..., R_J)^\top$ to the J items. These responses are assumed to be dependent in a certain way on K unobserved latent attributes. Under RLCMs, a complete set of K latent attributes is known as a latent class or an attribute profile, denoted by a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$, where $\alpha_k \in \{0, 1\}$ is a binary indicator of the absence or presence, respectively, of the kth attribute.

RLCMs assume a two-step data-generating process. The first step uses a population model for the attribute profile vector. We assume that the attribute profile follows a

categorical distribution with population proportions $\boldsymbol{p} := (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)^\top$, where $p_{\boldsymbol{\alpha}} > 0$, for all $\boldsymbol{\alpha} \in \{0,1\}^K$ and $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$.

The second step of the data-generating process follows a latent class model framework, incorporating constraints based on the underlying cognitive processes. Given a subject's attribute profile $\boldsymbol{\alpha}$, his/her responses to the J items $\{R_j : j = 1, \cdots, J\}$ are assumed to be conditionally independent, and each R_j follows a Bernoulli distribution with parameter $\theta_{j,\boldsymbol{\alpha}} = P(R_j = 1 \mid \boldsymbol{\alpha})$. Here, $\theta_{j,\boldsymbol{\alpha}}$ denotes the probability of a positive response, and is also called an item parameter of item j. The collection of all item parameters, denoted by the item parameter matrix $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$, is further constrained by the design matrix Q. The Q-matrix is the key structure that specifies the relationship between the J items and the K latent attributes. Specifically, the Q-matrix is a $J \times K$ binary matrix, with entries $q_{j,k} \in \{1, 0\}$ that indicate whether or not the jth item is linked to the kth latent attribute. When $q_{j,k} = 1$, we say attribute k is required by item j. The jth row vector $\boldsymbol{q}_j$ of Q gives the full attribute requirements of item j. Given an attribute profile $\boldsymbol{\alpha}$ and a matrix Q, we write $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ if $\alpha_k \geq q_{j,k},$ for all $k \in \{1, \ldots, K\}$, and $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j$ if there exists k such that $\alpha_k < q_{jk}$; similarly, we define the operations $\preceq$ and $\not\preceq$.

If $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, a subject with attribute pattern $\boldsymbol{\alpha}$ possesses all attributes required by item j specified by the Q-matrix, and is "capable" of answering item j correctly. On the other hand, if $\boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j$, a subject with $\boldsymbol{\alpha}'$ misses some required attribute of item j,

and is expected to have a smaller positive response probability than those of subjects with $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$. That is, the RLCMs we consider assume

$$\theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha} \succeq \boldsymbol{q}_j \text{ and } \boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j. \tag{2.1}$$

The *monotonicity assumption* in (2.1) is common to most RLCMs. Another common assumption of RLCMs is that mastering these nonrequired attributes of an item does not change the positive response probability to it; that is, $\theta_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}'}$ if $\boldsymbol{\alpha} \odot \boldsymbol{q}_j = \boldsymbol{\alpha}' \odot \boldsymbol{q}_j$, where "$\odot$" denotes the elementwise multiplication operator (Henson et al., 2009). Under the introduced setup, the response vector $\boldsymbol{R}$ has a probability mass function of the form

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-r_j}, \quad \boldsymbol{r} \in \{0,1\}^J, \tag{2.2}$$

where the constraints on $\theta_{j,\boldsymbol{\alpha}}$ imposed by Q are made implicit.

Next, we review several popular cognitive diagnosis models, showing where they fall within the family of RLCMs.

Example 1 (DINA model)**.** The DINA model is one of the basic cognitive diagnosis models (Junker and Sijtsma, 2001). The model assumes a conjunctive relationship among attributes, which means that providing a positive response to an item requires possessing all its required attributes, as indicated by the Q-matrix. For an item j and a subject with attribute profile $\boldsymbol{\alpha}$, an ideal response under the DINA model is

defined as $\Gamma_{j,\boldsymbol{\alpha}}^{DINA} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)$, which indicates whether the subject is capable of responding to item j. The uncertainty is incorporated at the item level in the slipping parameter $s_j = P(R_j = 0 \mid \Gamma_{j,\boldsymbol{\alpha}} = 1)$, denoting the probability that a capable subject slips the positive response, and the guessing parameter $g_j = P(R_j = 1 \mid \Gamma_{j,\boldsymbol{\alpha}} = 0)$, denoting the probability that a noncapable subject coincidentally gives the positive response by guessing. Then, the positive response probability for item j of class $\boldsymbol{\alpha}$ is $\theta_{j,\boldsymbol{\alpha}}^{DINA} = (1 - s_j)^{\Gamma_{j,\boldsymbol{\alpha}}} g_j^{1-\Gamma_{j,\boldsymbol{\alpha}}}$. The DINA model has only two parameters (i.e., s_j and g_j) for each item, regardless of the number of attributes required by the item. In the following discussion, we denote $\boldsymbol{s} = (s_1, \ldots, s_J)^\top$ and $\boldsymbol{g} = (s_1, \ldots, s_J)^\top$. Given a Q-matrix, the DINA model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ can be expressed equivalently by $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$. We further assume $\mathbf{1} - \boldsymbol{s} \succ \boldsymbol{g}$ (Xu and Zhang, 2016), which ensures that the DINA model satisfies the monotonicity assumption (2.1). The identifiability results for the basic DINA model are presented in Section 4.

Example 2 (GDINA model and General RLCMs)**.** de la Torre (2011) extended the DINA model to the generalized DINA (GDINA) model, which is formulated on the basis that $\theta_{j,\boldsymbol{\alpha}}$ can be decomposed into the sum of the effects caused by the presence of specific attributes and their interactions. Specifically, for an item j with $\boldsymbol{q}$-vector $\boldsymbol{q}_j = (q_{j,k} : k = 1, \cdots, K)$, the positive response probability is

$$\theta_{j,\boldsymbol{\alpha}}^{GDINA} = \sum_{\mathcal{S} \subseteq \{1,\ldots,K\}} \beta_{j,\mathcal{S}} \prod_{k \in \mathcal{S}} q_{j,k} \prod_{k \in \mathcal{S}} \alpha_k. \tag{2.3}$$

Note that not all β-coefficients in the above equation are included in the model. For a subset $\mathcal{S}$ of the K attributes $\{1, \ldots, K\}$, $\beta_{j,\mathcal{S}} \neq 0$ only if $\prod_{k \in \mathcal{S}} q_{j,k} = 1$. We interpret this as $\beta_{j,\varnothing}$ denoting the probability of a positive response when none of the required attributes are present in $\boldsymbol{\alpha}$; when $q_{j,k} = 1$, $\beta_{j,\{k\}}$ is included in the model, representing the change in the positive response probability resulting from the mastery of a single attribute k; when $q_{j,k} = q_{j,k'} = 1$, $\beta_{j,\{k,k'\}}$ is included in the model, representing the change in the positive response probability resulting from the interaction effect of mastering both k and k'. Under the GDINA model, each $\theta_{j,\boldsymbol{\alpha}}$ models the main effects and all interaction effects of the attributes measured by the item. We refer to these diagnostic models as *general* RLCMs. Other popular general RLCMs include the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and the general diagnostic model (GDM; von Davier, 2008). The identifiability results for general RLCMs are presented in Section 5.

3. Definitions and examples of strict and generic identifiability

This section introduces the definitions of joint strict identifiability and joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ for RLCMs, and gives an illustrative example.

Note that the monotonicity assumption stated in (2.1), is necessary for the identifiability of the Q-matrix, because, without it, $Q \neq \mathbf{1}_{J \times K}$ with parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ is distinguished from $\bar{Q} = \mathbf{1}_{J \times K}$ with the same parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ under the general

RLCM. The monotonicity constraints ensure that the constraints induced by $Q \neq \mathbf{1}_{J\times K}$ and $\bar{Q} = \mathbf{1}_{J\times K}$ cannot be the same and, therefore, Q can be identified under additional conditions; see Sections 4 and 5. In the following we assume the monotonicity assumption introduced in Section 2 is satisfied.

Another common issue with the identifiability of the Q-matrix is label swapping. In an RLCM setting, arbitrarily reordering the columns of a Q-matrix does not change the distribution of the responses. As a result, it is only possible to identify Q up to column permutation; thus, we write $\bar{Q} \sim Q$ if $\bar{Q}$ and Q have an identical set of column vectors, and write $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \sim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ if $\bar{Q} \sim Q$ and $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) = (\boldsymbol{\Theta}, \boldsymbol{p})$.

We first define the identifiability of the Q-matrix and the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$. We refer to this as *joint strict identifiability*.

Definition 1 (Joint Strict Identifiability)**.** Under an RLCM, the design matrix Q joint with the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are said to be strictly identifiable if for any $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$, there is no $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ such that

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \quad \text{for all} \quad \boldsymbol{r} \in \{0,1\}^J. \tag{3.4}$$

In the following discussion, we write (3.4) simply as $\mathbb{P}(\boldsymbol{R} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$.

Despite being the most stringent criterion for identifiability, strict identifiability can be too restrictive, ruling out many cases where $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are "almost surely" identifiable.

In the literature on unrestricted latent class models, Allman et al. (2011) proposed and studied the so-called *generic identifiability* of such models. Here, we introduce the concept of generic identifiability for RLCMs as follows.

Definition 2 (Joint Generic Identifiability)**.** Consider an RLCM with parameter space $\boldsymbol{\vartheta}_Q$, which is of full dimension in $\mathbb{R}^m$, with m corresponding to the number of free parameters in the model. The matrix Q joint with the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are said to be generically identifiable if the following set has Lebesgue measure zero in $\mathbb{R}^m$:

$$\boldsymbol{\vartheta}_{non} = \{(\boldsymbol{\Theta}, \boldsymbol{p}) : \exists (\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p}), \text{ such that } \mathbb{P}(\boldsymbol{R} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})\}.$$

3.1 Example of the generic identifiability phenomenon with $Q_{4\times 2}$

Here, we use an example to explain the difference between generic identifiability and strict identifiability. Consider the Q-matrix $Q_{4\times 2}$ in (3.5). Under the DINA model, we prove that this Q-matrix, joint with the associated model parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, is generically identifiable (by part (b.2) of Theorem 2), but not strictly identifiable (by Theorem 1).

$$Q_{4\times 2} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}^\top. \tag{3.5}$$

In particular, as long as the true proportions $\boldsymbol{p} = (p_{(00)}, p_{(01)}, p_{(10)}, p_{(11)})$ satisfy the following inequality constraint, $(Q_{4\times 2}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is identifiable (see the proof of Theorem

2 (b.2)):

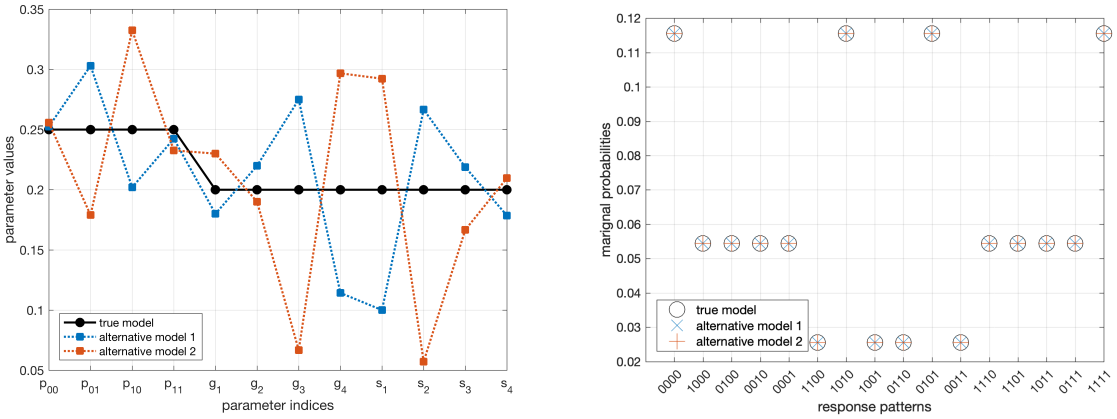
$$p_{(01)}p_{(10)} \neq p_{(00)}p_{(11)}. \tag{3.6}$$

On the other hand, when $p_{(01)}p_{(10)} = p_{(00)}p_{(11)}$, the model parameters are not identifiable, and there exist infinitely many sets of parameters that provide the same distribution of the observed response vector. Here, the parameter space $\boldsymbol{\vartheta}_Q = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \mathbf{1} - \boldsymbol{s} \succ \boldsymbol{g},\ \boldsymbol{p} \succ \mathbf{0},\ \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1\}$ is of full dimension in $\mathbb{R}^{11}$, where the nonidentifiable subset $\boldsymbol{\vartheta}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : p_{(01)}p_{(10)} = p_{(00)}p_{(11)}\}$ has Lebesgue measure zero in $\mathbb{R}^{11}$. We use a simulation study to illustrate the generic identifiability phenomenon. Under the $Q_{4\times 2}$ in (3.5), consider the following two simulation scenarios:

(a) the true model parameters are set as $g_j = s_j = 0.2$ for $j = 1, 2, 3, 4$, and $p_{(00)} = p_{(01)} = p_{(10)} = p_{(11)} = 0.25$, which violates (3.6);

(b) the true model parameters are generated randomly, which almost always satisfies (3.6). Specifically, we randomly generate 100 true parameter sets $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ using the following generating mechanism: $s_j \sim \mathcal{U}(0.1, 0.3)$, $g_j \sim \mathcal{U}(0.1, 0.3)$ for $j = 1, 2, 3, 4$, and $\boldsymbol{p} \sim$ Dirichlet$(3, 3, 3, 3)$. Here $\mathcal{U}(0.1, 0.3)$ denotes the uniform distribution on $[0.1, 0.3]$, and Dirichlet$(3, 3, 3, 3)$ denotes the Dirichlet distribution with parameter vector $(3, 3, 3, 3)$.

We show numerically that in scenario (a), there exist multiple sets of valid DINA parameters that give the same distribution of $\boldsymbol{R}$; in scenario (b), the model $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$



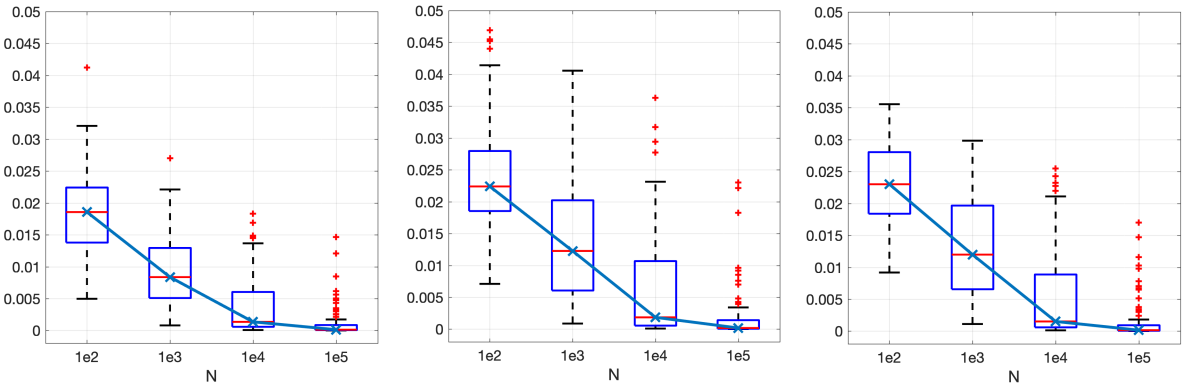
(a) (b)

Figure 1: Illustration of nonidentifiability under $Q_{4\times 2}$ in scenario (a).

is almost surely identifiable and estimable. In particular, corresponding to scenario (a), Figure 1 (a) plots the true model parameters and the other two sets of valid DINA model parameters (constructed based on the derivations in the proof of Theorem 2 (b.2)), and Figure 1 (b) plots the marginal probabilities of all $2^4 = 16$ response patterns under the three sets of model parameters. We can see that despite these three sets of parameters being quite different, they give the identical distribution of the four-dimensional binary response vector.

Corresponding to scenario (b), we randomly generate $B = 100$ sets of true parameters $(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, for $i = 1, \ldots, 100$. Then, for each $(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, we generate 200 independent data sets of size N, with $N = 10^2$, 10^3, 10^4, and 10^5, and then compute the mean square errors (MSEs) of the maximum likelihood estimators (MLEs) of the slipping, guessing and proportion parameters. To compute the MLEs of the model

parameters for each simulated data set, we run the EM algorithm with 10 random initializations, and choose the estimators that achieve the largest log-likelihood value of the 10 runs. Figure 2 shows the box plots of MSEs associated with the $B = 100$ true parameter sets for each sample size N. As N increases, we observe that the MSEs decrease to zero, indicating the (generic) identifiability of these randomly generated parameters.



(a) MSE of $\boldsymbol{p}$ (b) MSE of $\boldsymbol{s}$ (c) MSE of $\boldsymbol{g}$

Figure 2: Illustration of generic identifiability under $Q_{4\times 2}$, which corresponds to simulation scenario (b).

On the other hand, Figure 2 also shows that several parameter sets have MSEs that are "outliers" that converge to zero more slowly than others do as N increases. This happens because these sets of parameters fall near the nonidentifiability set $\mathcal{V}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \ p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$, making it more difficult to identify them. To illustrate this point, consider the scenario corresponding to the rightmost box plot in Figure 2(a), with sample size $N = 10^5$. For each of the 100 sets of true parameters

$(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, we plot $p^i_{(00)} \cdot p^i_{(11)}$ and $p^i_{(01)} \cdot p^i_{(01)}$ as the x-axis and y-axis coordinates, respectively (see Figure 3). Then, each point represents one set of true parameters used to generate the data. Specifically, we plot these parameter sets using a red "$*$" if their corresponding MSEs are the 20% largest outliers in the rightmost box plot in Figure 2(a); we plot the remaining 80% of the parameter sets using a blue "+". One can clearly see that as the true parameters become closer to the nonidentifiability set $\mathcal{V}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \ p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$ (represented by the straight reference line drawn from $(0, 0)$ to $(0.17, 0.17)$), the MSEs increase, and the MSEs converge more slowly. Thus, under generic identifiability, when the true model is close to the nonidentifiable set, the convergence of their MLEs becomes slow.

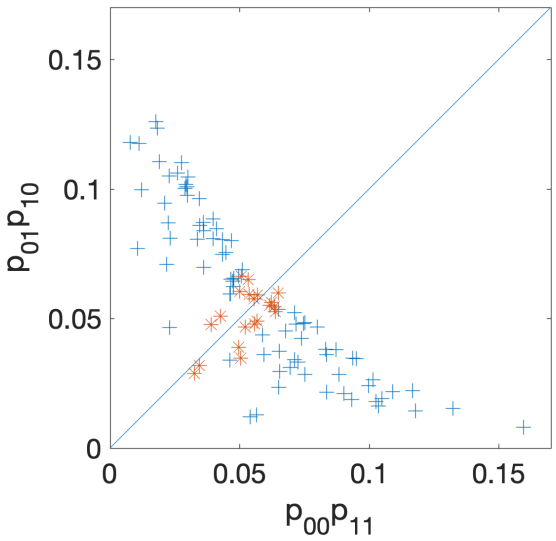


Figure 3: The effect of the generic identifiability constraint (3.6). Red "$*$"s represent parameter sets with the 20% largest MSEs in Figure 2(a), with $N = 10^5$; blue "+"s represent the remaining parameter sets.

Interestingly, the generic identifiability constraint (3.6) is equivalent to the statement that the two latent attributes are *not independent* of each other. To see this, view each subject's two-dimensional attribute profile as a random vector taking values in a 2×2 contingency table. Then, (3.6) states that the 2×2 matrix of joint probabilities of attributes mastery,

$$\begin{pmatrix} p_{(00)} & p_{(01)} \\ p_{(10)} & p_{(11)} \end{pmatrix},$$

has full rank, with nonzero determinant $p_{(00)}p_{(11)} - p_{(01)}p_{(10)}$. Therefore, one row (resp. column) of the matrix cannot be a multiple of the other row (resp. column), and hence the two binary attributes can not be independent. Intuitively, this implies that the DINA model essentially requires that each attribute is measured at least three times for identifiability (as shown in Condition B in Theorem 1). In particular, consider those attributes that are measured by only two items in the Q-matrix. If these attributes are independent, then, intuitively, they provide an independent source of information in which case the model is not identifiable. However, if these attributes are dependent, then the dependency instead helps to identify the model structure.

Before stating the strict and generic identifiability results on $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$, we show in the next proposition that any all-zero row vector in the Q-matrix can be dropped without affecting the identifiability conclusion.

Proposition 1. *Suppose the Q-matrix of size $J \times K$ takes the form $Q = ((Q')^\top, \mathbf{0}^\top)^\top$,*

where Q' is a $J' \times K$ submatrix containing J' nonzero $\boldsymbol{q}$-vectors, and $\mathbf{0}$ denotes a $(J - J') \times K$ submatrix containing these zero $\boldsymbol{q}$-vectors. Let $\boldsymbol{\Theta}'$ be the submatrix of $\boldsymbol{\Theta}$ containing its first J' rows. Then, for any RLCM, $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are jointly strictly (generically) identifiable if and only if $(Q', \boldsymbol{\Theta}', \boldsymbol{p})$ are jointly strictly (generically) identifiable.

Therefore, without loss of generality, from now on, we only consider Q-matrices without any zero $\boldsymbol{q}$-vectors when discussing joint identifiability. We examine various RLCMs that are popular in cognitive diagnosis assessment. In particular, in Section 4, we present the sufficient and necessary conditions for the strict and generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under the basic DINA model. These identifiability results can also be applied to the DINO model (Templin and Henson, 2006), owing to the duality between the two models (Chen et al., 2015). Section 5 presents the sufficient and necessary conditions for the generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under general RLCMs, which include the popular GDINA and LCDM models.

4. Identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under the DINA model

Under the DINA model, Liu et al. (2013) first studied the identifiability of the Q-matrix under the assumption that the guessing parameters $\boldsymbol{g}$ are known. Chen et al. (2015) and Xu and Shang (2018) proposed a further set of sufficient conditions without needing to assume known item parameters. An important requirement in these identifiability studies is the completeness of the Q-matrix (Chiu et al., 2009). Under the DINA model,

the Q-matrix is said to be complete if it contains a $K \times K$ identity submatrix I_K up to column permutation. Chen et al. (2015) and Xu and Shang (2018) require Q to contain at least two complete submatrices I_K for identifiability.

However, determining the minimal requirements on the Q-matrix for identifiability remains an open problem. In the next theorem, we solve this problem by providing the necessary and sufficient condition for the identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ under the earlier assumption that $p_{\boldsymbol{\alpha}} > 0$, for all $\boldsymbol{\alpha} \in \{0,1\}^K$ (Xu and Zhang, 2016; Gu and Xu, 2018).

Theorem 1. *Under the DINA model, the combination of Conditions A, B, and C is necessary and sufficient for the strict identifiability of* $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$*:*

A. The true Q-matrix is complete. Without loss of generality, assume the Q-matrix takes the following form:

$$Q = \begin{pmatrix} I_K \\ Q^\star \end{pmatrix}. \tag{4.7}$$

B. The column vectors of the submatrix $Q^\star$ *in* (4.7) *are distinct.*

C. Each column in Q contains at least three entries equal to one.

In the Supplementary Material, we provide simulations that verify Theorem 1. In particular, see simulation study I for the sufficiency of Conditions A, B, and C for joint identifiability; also see simulation studies III and IV for the necessity of the proposed conditions. Next, we compare our Theorem 1 with several existing results. First,

although the same set of conditions is proposed in Gu and Xu (2018), they assumed *a known* Q when examining the identifiability of the parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$. In contrast, Theorem 1 studies the joint identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, which is theoretically much more challenging, owing to the unknown Q-matrix, and therefore provides a much stronger result than that in Gu and Xu (2018). In terms of estimation, Theorem 1 implies that we can consistently estimate both Q and $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, without worrying that an incorrect Q-matrix is indistinguishable from the true Q. Second, Theorem 1 has much weaker requirements than those of the well-known identifiability conditions resulting from a three-way tensor decomposition (Kruskal, 1977; Allman et al., 2011). Specifically, these classical results require that the number of items $J \geq 2K + 1$ for (generic) identifiability. In contrast, the conditions in Theorem 1 imply that we need the number of items J to be at least $K + \lceil \log_2(K) \rceil + 1$ under the DINA model. This is because, other than the identity submatrix I_K, in order to satisfy Condition B of *distinctness*, the Q-matrix needs only contain a further $\log_2(K)$ items whose K-dimensional $\boldsymbol{q}$-vectors form a matrix with K distinct columns. For example, for $K = 8$, the conditions in Allman et al. (2011) require at least $2K + 1 = 17$ items, whereas our Theorem 1 guarantees that the following Q with $K + \log_2(K) + 1 = 12$ items suffices

for the strict identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ under DINA:

$$Q = \begin{pmatrix} & & & I_8 & & & & \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Conditions A, B, and C are the minimal requirements for joint strict identifiability. When the true Q fails to satisfy one or more of these, Theorem 1 implies that there must exist $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) \nsim (\bar{Q}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ such that (3.4) holds. In this scenario, there are still cases where the model is "almost surely" identifiable, though not strictly identifiable, as illustrated by the example under $Q_{4\times 2}$ in (3.5). On the other hand, there are also cases where the entire model is never identifiable, as shown in simulation studies III and IV in the Supplementary Material. Therefore, it is desirable to determine which conditions guarantee the generic identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$.

In the following, we discuss the necessity of Conditions A, B, and C under the weaker notion of generic identifiability. First, Condition A is necessary for the joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. If the true Q-matrix does not satisfy Condition A, then under the DINA model, certain latent classes would be equivalent given Q, and their separate proportion parameters can never be identified, not even generically (Gu

and Xu, 2020). In certain scenarios where Condition A fails, one can find a different $\bar{Q}$ that is not distinguishable from Q. Simulation study IV in the Supplementary Material illustrates the necessity of Condition A.

Second, Condition B is also difficult to relax, and serves as a necessary condition for generic identifiability when $K = 2$. Specifically, as shown in Gu and Xu (2018), when $K = 2$, the only possible structure of the Q-matrix that violates Condition B while satisfying Conditions A and C is

$$Q = \begin{pmatrix} 1 & 0 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \end{pmatrix}^\top .$$

In addition, Gu and Xu (2018) prove that for *any* valid DINA parameters associated with this Q, there exist infinitely many different sets of DINA parameters that lead to the same distribution of the responses. Therefore, the model is not generically identifiable.

Third, in contrast to Conditions A and B, for generic identifiability, Condition C can be relaxed to a certain extent. The next theorem characterizes how the Q-matrix structure in this case affects generic identifiability. For an empirical verification of Theorem 2, see simulation study II in the Supplementary Material.

Theorem 2. *Under the DINA model,* $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ *is not generically identifiable if some attribute is required by only one item.*

If some attribute is required by only two items, suppose the Q-matrix takes the following form, after some column and row permutations:

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & Q^\star \end{pmatrix}, \tag{4.8}$$

where $\boldsymbol{v}$ *is a vector of length* $K-1$*, and* $Q^\star$ *is a* $(J-2)\times(K-1)$ *submatrix.*

(a) If $\boldsymbol{v}=\mathbf{1}$*,* $(Q,\boldsymbol{s},\boldsymbol{g},\boldsymbol{p})$ *is not locally generically identifiable.*

(b) If $\boldsymbol{v}=\mathbf{0}$*,* $(Q,\boldsymbol{s},\boldsymbol{g},\boldsymbol{p})$ *is globally generically identifiable if either*

(b.1) the submatrix $Q^\star$ *satisfies Conditions A, B, and C in Theorem 1; or*

(b.2) the submatrix $Q^\star$ *has two submatrices* I_{K-1}*.*

(c) If $\boldsymbol{v}\neq\mathbf{0},\mathbf{1}$*,* $(Q,\boldsymbol{s},\boldsymbol{g},\boldsymbol{p})$ *is locally generically identifiable if* $Q^\star$ *satisfies Conditions A, B, and C in Theorem 1.*

Remark 1. We say $(Q,\boldsymbol{s},\boldsymbol{g},\boldsymbol{p})$ is locally identifiable if, in a neighborhood of the true parameters, there does not exist a different set of parameters that gives the same distribution of the responses. Local generic identifiability is a weaker notion than (global) generic identifiability. Therefore, the statement in part (a) of Theorem 2 also implies that $(Q,\boldsymbol{s},\boldsymbol{g},\boldsymbol{p})$ is not globally generically identifiable.

Remark 2. In scenario (b.1) of Theorem 2, the identifiable subset of the parameter space is $\big\{(\boldsymbol{s},\boldsymbol{g},\boldsymbol{p}) : \exists \boldsymbol{\alpha}^1 = (0,\alpha_2^1,\ldots,\alpha_K^1), \boldsymbol{\alpha}^2 = (0,\alpha_2^2,\ldots,\alpha_K^2) \in \{0\}\times\{0,1\}^{K-1}$, such that $p_{\boldsymbol{\alpha}^1}p_{\boldsymbol{\alpha}^2+\boldsymbol{e}_1} \neq p_{\boldsymbol{\alpha}^2}p_{\boldsymbol{\alpha}^1+\boldsymbol{e}_1}\big\}$, where $\boldsymbol{e}_j$ is a J-dimensional unit vector, with the jth element equal to one and all the others zero. In scenario (b.2) of Theorem 2, we can write $Q = (I_K, I_K, (Q^{\star\star})^\top)^\top$, in which case, the identifiable subset is $\big\{(\boldsymbol{s},\boldsymbol{g},\boldsymbol{p}) : \forall k \in \{1,\ldots,K\}, \exists \boldsymbol{\alpha}^{k,1}, \boldsymbol{\alpha}^{k,2} \in \{0,1\}^{k-1}\times\{0\}\times\{0,1\}^{K-k-1}$, such that $p_{\boldsymbol{\alpha}^{k,1}}p_{\boldsymbol{\alpha}^{k,2}+\boldsymbol{e}_k} \neq p_{\boldsymbol{\alpha}^{k,2}}p_{\boldsymbol{\alpha}^{k,1}+\boldsymbol{e}_k}\big\}$. The complements of these identifiable subsets in the parameter space give the nonidentifiable subsets, which are both of measure zero in the DINA model parameter space.

Next we discuss the generic identifiability of the DINA model in the special case of $K = 2$. We have the following proposition.

Proposition 2. *Under the DINA model with $K = 2$ attributes, $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is generically identifiable if and only if the conditions in Theorem 1 or 2(b) hold.*

Proposition 2 gives a full characterization of joint generic identifiability when $K = 2$, showing that the proposed generic identifiability conditions are necessary and sufficient in this case. The following example discusses all possible Q-matrices with $K = 2$, such that $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not strictly identifiable, which proves Proposition 2 automatically.

Example 3. When $K = 2$, the discussions on Conditions A and B before Theorem 2 show that $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not generically identifiable when A or B is violated. Therefore,

we need only focus on cases where Condition C is violated and Conditions A and B are satisfied. Specifically, when $J \leq 5$, the Q-matrix can only take the following forms up to column and row permutations:

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad Q_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

By Theorem 2, Q_1 falls in scenario (a); therefore, $(Q_1, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not locally generically identifiable; that is, even in a small neighborhood of the true parameters, there exist infinitely many different sets of parameters that give the same distribution of the responses. On the other hand, Q_2 falls in scenario (b.2) and Q_3 falls in scenario (b.1). Therefore, $(Q_2, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ and $(Q_3, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are both generically identifiable. In the case of $J > 5$, any Q satisfying A and B while violating C must contain one of the above Q_i as a submatrix and include additional row vectors of $(0, 1)$. By Theorem 2, any such Q extended from Q_1 is still not locally generically identifiable, and any such Q extended from Q_2 or Q_3 is globally generically identifiable.

5. Identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under general RLCMs

Because the DINA model is a submodel of general RLCMs, Conditions A, B, and C in Theorem 1 are also necessary for the strict identifiability of general RLCMs. For instance, our proposed Conditions A, B, and C are weaker than the sufficient conditions proposed by Xu and Shang (2018) for the strict identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under general RLCMs; and if their conditions are satisfied, the current conditions A, B, and C are also satisfied. However, these necessary requirements may be strong in practice, and cannot be applied to identify any Q that lacks some single-attribute items (i.e., lacks some unit vector as a row vector). A natural question is whether Conditions A, B, and C can be relaxed under the weaker notation of of generic identifiability. This section addresses this question.

Under general RLCMs, the next theorem shows that Condition C (each attribute is required by at least three items) is necessary for the generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$, contrary to the results for the DINA model, where Conditions A and B cannot be relaxed, but Condition C can. Simulation studies VI and VII in the Supplementary Material verify Theorem 3.

Theorem 3. *Under a general RLCM, Condition C in Theorem 1 is necessary for the generic identifiability of* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. *Specifically, when the true Q-matrix violates C, for any model parameters* $(\boldsymbol{\Theta}, \boldsymbol{p})$ *associated with Q, there exist infinitely many sets of*

$(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *such that equation* (3.4) *holds. Thus,* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *is not generically identifiable.*

Whereas Condition C is necessary, we next show that the other two conditions, A and B, can be relaxed further for the generic identifiability of general RLCMs. Before stating the result, we first introduce a new concept about the Q-matrix, called *generic completeness*.

Definition 3 (Generic Completeness)**.** A Q-matrix with K attributes is said to be generically complete if, after some column and row permutations, it has a $K \times K$ submatrix with all diagonal entries equal to one.

Generic completeness is a relaxation of the concept of completeness. In particular, a Q-matrix is generically complete if, up to column and row permutations, it contains a submatrix as follows:

$$\begin{pmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 1 \end{pmatrix},$$

where the off-diagonal entries "$*$" are left unspecified. Note that any complete Q-matrix is also generically complete, whereas a generically complete Q-matrix may not have any single-attribute items.

Using the concept of generic completeness, the next theorem gives sufficient conditions for joint generic identifiability, and shows that under general RLCMs, the necessary conditions A and B for strict identifiability are no longer necessary in the current setting.

Theorem 4. *Under a general RLCM, if the true Q-matrix satisfies the following Conditions D and E, then $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ is generically identifiable.*

D. The Q-matrix has two nonoverlapping generically complete $K \times K$ submatrices Q_1 and Q_2. Without loss of generality, assume the Q-matrix is in the following form:

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q^\star \end{pmatrix}_{J \times K} . \tag{5.9}$$

E. Each column of the submatrix $Q^\star$ in (5.9) contains at least one entry of one.

Remark 3. Under Theorem 4, the identifiable subset of the parameter space is $\{(\boldsymbol{\Theta}, \boldsymbol{p}) : \det(T(Q_1, \boldsymbol{\Theta}_{Q_1})) \neq 0,\ \det(T(Q_2, \boldsymbol{\Theta}_{Q_2})) \neq 0,$ and $T(Q^\star, \boldsymbol{\Theta}_{Q^\star}) \cdot \text{Diag}(\boldsymbol{p})$ has distinct column vectors$\}$. Its complement is the nonidentifiable subset, and it has measure zero in the parameter space $\boldsymbol{\vartheta}_Q$ when Q satisfies Conditions D and E. Please see the supplementary materials for the definition of the T-matrices ($T(Q_1, \boldsymbol{\Theta}_{Q_1})$, etc.).

Remark 4. The proof of Theorem 4 is based on the proof of Theorem 7 in Gu and Xu (2020), who proposed the same Conditions D and E as sufficient conditions for the generic identifiability of model parameters, given a known Q. We point out that though D and E serve as sufficient conditions for generic identifiability, both when Q is known and when Q is unknown, the generic identifiability results in these two scenarios are different. In particular, Theorem 8 in Gu and Xu (2020) shows that when Q is known, some attribute can be required by only two items for generic identifiability to hold (i.e., Condition C can be relaxed); in contrast, our current Theorem 3 shows that when Q is unknown, Condition C indeed becomes necessary.

The proposed sufficient Conditions D and E weaken the strong requirement of Conditions A and B, especially the identity submatrix requirement that may be difficult to satisfy in practice. Simulation study V in the Supplementary Material verifies Theorem 4. Note that Conditions D and E imply the necessary Condition C that each attribute is required by at least three items.

We next discuss the necessity of Conditions D and E. As shown in Section 3.2, under DINA, the completeness of Q is necessary for the joint strict identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$. For general RLCMs, we have an analogous conclusion that the generic completeness of Q, which is part of Condition D, is necessary for the joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. This is stated in the next theorem.

Theorem 5. *Under a general RLCM, generic completeness of the Q-matrix is necessary*

for the joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$.

Furthermore, we show that Conditions D and E themselves are in fact necessary when $K = 2$, indicating the difficulty of relaxing these further.

Proposition 3. *For a general RLCM with* $K = 2$*, Conditions* D *and* E *are necessary and sufficient for the generic identifiability of* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$.

We use the following example to illustrate the result of Proposition 3, which also gives a natural proof of the proposition.

Example 4. When $K = 2$, a Q-matrix that satisfies the necessary Condition C, but not Conditions D or E, can only take the following form Q_1 or Q_2, up to row permutations:

$$Q_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1 & * \\ * & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}; \qquad \bar{Q}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The "$*$"s in Q_2 are unspecified values, and can be either zero or one. For Q_1 with $J = 3$, $K = 2$, and any parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$, there are $2^J = 8$ constraints in (3.4) for solving $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ under Q_1 itself, whereas the number of free parameters of $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ is $|\{p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^2\} \cup \{\theta_{j,\boldsymbol{\alpha}} : j \in \{1,2\}, \boldsymbol{\alpha} \in \{0,1\}^2\}| = 2^K + 2^K \times J = 16 > 8$. For Q_2 with $J = 4$, $K = 2$, and any associated $(\boldsymbol{\Theta}, \boldsymbol{p})$, there are $2^J = 16$ constraints in (3.4) for

solving $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$, whereas the number of free parameters of $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ under the alternative $\bar{Q}_2$ is $2^K + J \times 2^K = 20 > 2^J = 16$. In both cases, there are infinitely many sets of solutions of (3.4) as alternative model parameters. Therefore, neither $(Q_1, \boldsymbol{\Theta}, \boldsymbol{p})$ nor $(Q_2, \boldsymbol{\Theta}, \boldsymbol{p})$ are generically identifiable.

6. Conclusion

In this work, we study the identifiability issue of RLCMs with unknown Q-matrices. For the basic DINA model, we derive the necessary and sufficient conditions for the strict joint identifiability of the Q-matrix and the associated model parameters. We also study a slightly weaker identifiability notion, called generic identifiability, and propose sufficient and necessary conditions for it under the DINA model and general RLCMs.

Statistical consequences of identifiability. In the setting of RLCMs, identifiability naturally leads to estimability, in different senses, under strict and generic identifiability. If the Q-matrix and the associated model parameters are strictly identifiable, then Q and the model parameters can consistently be jointly estimated from the data. If the Q-matrix and the model parameters are generically identifiable, then for true parameters ranging almost everywhere in the parameter space with respect to the Lebesgue measure, the Q-matrix and the model parameters can consistently be jointly estimated from the data.

As pointed out by one reviewer, the analysis of identifiability is under an ideal situation with an infinite sample size. Indeed, general identification problems assume the hypothetical exact knowledge of the distribution of the observed variables, and ask under what conditions one can recover the underlying parameters (Allman et al., 2009). Next, we discuss the finite-sample estimation issue under the proposed identifiability conditions for strict identifiability, following a similar argument to that in Proposition 1 in Xu and Shang (2018). Denote the true Q-matrix and model parameters by Q^0 and $\boldsymbol{\eta}^0 = (\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$, respectively. Consider a sample with N independent and identically distributed (i.i.d.) response vectors $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_N$, and denote the log-likelihood of the sample by $\ell(\boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{i=1}^{N} \log \mathbb{P}(\boldsymbol{R}_i \mid Q, \boldsymbol{\Theta}, \boldsymbol{p})$. Under a specified RLCM, a Q-matrix determines the structure of the item parameter matrix $\boldsymbol{\Theta}$ by specifying which entries are equal. For a given $\boldsymbol{\Theta}$, we can define an equivalent formulation of it, a sparse matrix $\boldsymbol{B}$, with the same size as $\boldsymbol{\Theta}$, as follows. Under a general RLCM, such as the GDINA model in Example 2, the item parameters can be parameterized as $\theta_{j,\boldsymbol{\alpha}} = \sum_{\mathcal{S} \subseteq \{1,\ldots,K\}} \beta_{j,\mathcal{S}} \prod_{k \in \mathcal{S}} \alpha_k$. Based on this, we define the jth row of $\boldsymbol{B}$ as a 2^K-dimensional vector collecting all of these β-coefficients; that is, $\boldsymbol{B}_j = (\beta_{j,0}, \beta_{j,1}, \ldots, \beta_{j,K}, \ldots, \beta_{j,12\cdots K})$. Then, as long as the $\boldsymbol{q}$-vector $\boldsymbol{q}_j \neq \boldsymbol{1}_K$, the vector $\boldsymbol{B}_j$ and the matrix $\boldsymbol{B}$ are both "sparse". For the true Q^0, we denote the corresponding $\boldsymbol{B}$-matrix by $\boldsymbol{B}^0$. Under a specified RLCM (e.g., DINA or GDINA), the identification of Q^0 is then implied by the identification of the indices of nonzero elements of $\boldsymbol{B}^0$. De-

note the support of the true $\boldsymbol{B}^0$ and any candidate $\boldsymbol{B}$ by S_0 and S, respectively. Define $C_{\min}(\boldsymbol{\eta}^0) = \inf_{\{S \neq S_0, |S| \leq |S_0|\}}(|S_0 \setminus S|)^{-1} h^2(\boldsymbol{\eta}^0, \boldsymbol{\eta})$, where $h^2(\boldsymbol{\eta}^0, \boldsymbol{\eta})$ denotes the Hellinger distance between the two distributions of $\boldsymbol{R}$, indexed by parameters $\boldsymbol{\eta}^0$ under the true $\boldsymbol{B}^0$, and by $\boldsymbol{\eta}$ under the candidate $\boldsymbol{B}$. Denote the Q-matrix and the model parameters that maximize the log-likelihood $\ell(\boldsymbol{\Theta}, \boldsymbol{p})$ subject to the L_0 constraint $|S| \leq |S_0|$ by $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$, and denote the "oracle" MLEs of the model parameters obtained, assuming Q^0 is known, by $\widehat{\boldsymbol{\eta}}^0 = (\widehat{\boldsymbol{\Theta}}^0, \widehat{\boldsymbol{p}}^0)$. Then, we have the following finite-sample error bound for the estimated Q-matrix and model parameters.

Proposition 4. *Suppose Q^0 satisfies the proposed sufficient conditions for joint strict identifiability; then, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \geq c_0$, for some positive constant c_0. Furthermore,*

$$\mathbb{P}(\widehat{Q} \not\sim Q^0) \leq \mathbb{P}(\widehat{\boldsymbol{\eta}} \neq \widehat{\boldsymbol{\eta}}^0) \leq c_2 \exp\{-c_1 N C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)\}, \tag{6.10}$$

where $c_1, c_2 > 0$ are some constants. That is, when the joint strict identifiability conditions hold, the finite-sample estimation error has an exponential bound.

Proposition 4 shows that the estimation error decreases exponentially in N if the model is identifiable. On the other hand, when the identifiability conditions fail to hold, there exist alternative models that are close to the true model in terms of the Hellinger distance. This would make the $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ in (6.10) equal to zero, instead of being bounded away from zero, as shown in Proposition 4. Therefore, the finite-sample

error bound in (6.10) becomes $O(1)$ in this nonidentifiable scenario. In particular, when the generic identifiability conditions are satisfied, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ depends on the distance between the true parameters and the nonidentifiable measure-zero subset of the parameter space; as the true parameters become closer to this measure-zero set, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ decreases to zero, and a larger sample size may be needed to achieve a prespecified level of estimation accuracy.

Potential extensions to other latent variable models. We briefly discuss potential extensions of the proposed theory to other latent variable models, such as RLCMs with ordinal polytomous attributes (von Davier, 2008; Ma and de la Torre, 2016; Chen and de la Torre, 2018), and multidimensional latent trait models (Embretson, 1991). First, an RLCM with ordinal polytomous attributes can be viewed as an RLCM with binary attributes and a constrained relationship among the binary attributes. For instance, consider an ordinal attribute γ that can take C different values $\{0, 1, \ldots, C-1\}$; then, γ can be equivalently viewed as a collection of $C-1$ binary random variables $\boldsymbol{\alpha}^\gamma := (\alpha_1, \ldots, \alpha_{C-1})$ with the following constraints. If $\alpha_i = 0$ for some $i < C-1$, then $\alpha_j = 0$, for all $j = i+1, \ldots, C-1$. In other words, any pattern $\boldsymbol{\alpha}^\gamma$ with $\alpha_i = 0$ and $\alpha_j = 1$, for some $i < j$ is "forbidden" and constrained to have proportion zero. The vector of polytomous attributes can be augmented to a longer vector of binary attributes using constraints in this fashion. Then, we can consider the RLCM with the augmented

proportion parameters by constraining the proportions of the "forbidden" binary attribute patterns to zero. In this scenario, it might be possible to extend the current theory and develop identifiability conditions for the case of polytomous attributes.

Second, if a multidimensional latent trait model includes both continuous and discrete latent traits, then the techniques used to establish the identifiability of the latent class models in this study would also be useful when treating discrete latent variables. For continuous latent variables, the techniques developed in Bai and Li (2012) for the identifiability of the factor analysis model and those developed for traditional multivariate analyses (Anderson, 2009) would be helpful.

In practice, the proposed identifiability theory can serve as a foundation for designing statistically guaranteed estimation procedures. Specifically, consider the set of all Q-matrices that satisfy our identifiability conditions (A, B, and C under the DINA model, or D and E under general RLCMs), and call it the "identifiable Q-set." Then, we can use likelihood-based approaches, such as that in Xu and Shang (2018), to jointly estimate Q and the model parameters by constraining Q to the identifiable Q-set; alternatively we can use Bayesian approaches to estimate Q, as in Chen et al. (2018). Additionally, if under the DINA model, the Q-matrix does not contain a submatrix I_K, then according to Gu and Xu (2020), certain attribute profiles would be equivalent and the strongest possible identifiability argument therein is the so-called $\boldsymbol{p}$-partial identifiability. In this scenario, it would be interesting to study the identifiability of

the incomplete Q-matrix under the notion of $\boldsymbol{p}$-partial identifiability. We leave this to future research.

Supplementary Material

The online Supplementary Material contains proofs of Propositions 1 and 4 and Theorems 1–5, as well as additional simulation results.

Acknowledgments

This work was supported in part by National Science Foundation grants CAREER SES-1846747, SES-1659328, and DMS-1712717.

References

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

Allman, E. S., Matias, C., and Rhodes, J. A. (2011). Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736.

Anderson, T. (2009). *An introduction to multivariate statistical analysis, 3rd Ed.* Wiley India Pvt. Limited.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Chen, J. and de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Frontiers in psychology*, 9:1474.

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

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

Chiu, C.-Y., Douglas, J. A., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika*, 74:633–665.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45:343–362.

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

de la Torre, J. and Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81:253–273.

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

de la Torre, J., van der Ark, L. A., and Rossi, G. (2017). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, pages 1–16.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: the DINA model, classification, class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35:8–26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a bayesian extension of the DINA model. *Applied Psychological Measurement*, 36(6):447–468.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In Nichols, P. D., Chipman, S. F., and Brennan, R. L., editors, *Cognitively diagnostic assessment*, pages 361–390. Erlbaum Associates, Hillsdale, NJ.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3):495–515.

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

Gu, Y. and Xu, G. (2018). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, 84(2):468–483.

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. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*. To appear.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548.

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

Jaeger, J., Tatsuoka, C., Berns, S. M., and Varadi, F. (2006). Distinguishing neurocognitive functions in schizophrenia using partially ordered classification models. *Schizophrenia bulletin*, 32(4):679–691.

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

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138.

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, 36:548–564.

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, 19(5A):1790–1817.

Ma, W. and de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3):253–275.

Maris, G. and Bechger, T. M. (2009). Equivalent diagnostic classification models. *Measurement*, 7:41–46.

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

Tatsuoka, C. (2009). Diagnostic models as partially ordered sets. *Measurement*, 7:49–53.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Statist.*, 38:1300–1302.

Templin, J. L. and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11:287–305.

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

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model

equivalency. *British Journal of Mathematical and Statistical Psychology*, 67(1):49–71.

Wang, S., Lin, H., Chang, H.-H., and Douglas, J. (2016). Hybrid computerized adaptive testing: from group sequential design to fully sequential design. *Journal of Educational Measurement*, 53(1):45–62.

Wu, Z., Casciola-Rosen, L., Rosen, A., and Zeger, S. L. (2018). A Bayesian approach to restricted latent class models for scientifically-structured clustering of multivariate binary outcomes. *arXiv preprint arXiv:1808.08326*.

Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2017). 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*, 45:675–707.

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

Xu, G., Wang, C., and Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 69(3):291–315.

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

Zhang, S. and Chang, H.-H. (2016). From smart testing to smart learning: how testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 1(1):67–92.

Yuqi Gu and Gongjun Xu

Department of Statistics, University of Michigan

E-mail: yuqigu, gongjun@umich.edu