

Statistica Sinica Preprint No: SS-2018-0410

Title	Sufficient and Necessary Conditions for the Identifiability of the Q-matrix
Manuscript ID	SS-2018-0410
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0410
Complete List of Authors	Yuqi Gu and Gongjun Xu
Corresponding Author	Gongjun Xu
E-mail	gongjun@umich.edu
Notice: Accepted version subject to English editing.	

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. Different from conventional latent class models, restrictions on the RLCM model parameters are imposed by a design matrix to respect practitioners' scientific assumptions. The design matrix, called Q -matrix in the cognitive diagnosis literature, is usually constructed by practitioners and domain experts, yet it is subjective and could be misspecified. To address this problem, researchers have proposed to estimate the Q -matrix from data. On the other hand, the fundamental learnability issue of the Q -matrix and model parameters remains underexplored and existing studies often impose stronger than needed or even impractical conditions. This paper proposes sufficient and necessary conditions for joint identifiability of the Q -matrix and the RLCM model parameters under different types of RLCMs. The developed identifiability conditions only depend 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 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 paper focuses on a family of *restricted latent class models* (RLCMs), which play a key role in various fields, including cognitive diagnosis in educational assessments (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), disease etiology diagnosis and scientifically-structured clustering of patients (Wu et al., 2017, 2018).

Different from conventional latent class models, the model parameters of RLCMs are constrained through 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 with different 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 issue is challeng-

ing to address. Identifiability of RLCMs has long been a concern in practice, as noted in the literature (DiBello et al., 1995; Maris and Bechger, 2009; Tatsuoka, 2009; DeCarlo, 2011; von Davier, 2014). Existing identifiability results of unrestricted latent class models in statistics (Teicher, 1967; Goodman, 1974; Gyllenberg et al., 1994; Allman et al., 2009) cannot be directly applied to RLCMs due to the structural constraints induced by the Q -matrix here. Recently, the identifiability of RLCM model parameters has been studied for the basic DINA model (Chen et al., 2015; Xu and Zhang, 2016; Gu and Xu, 2018b) and general RLCMs (Xu, 2017; Gu and Xu, 2018a), assuming that the Q -matrix is prespecified and correct.

However, the Q -matrix, specified by scientific experts upon construction of the diagnostic items, can be misspecified. Moreover, in an exploratory analysis of newly designed items, a large part or the whole Q -matrix may not be available. The misspecification of the Q -matrix could lead to a serious lack of fit of the model and consequently inaccurate inference 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). To achieve reliable and valid estimation and inference on the Q -matrix, a fundamental issue is to ensure joint identifiability of the Q -matrix and the associated model parameters. Such joint identifiability has been recently studied in Liu et al. (2013) and Chen et al. (2015) under the DINA model and Xu and Shang (2018)

under general RLCMs. Nevertheless, these existing works mostly focus on developing sufficient conditions for joint identifiability, so they often impose stronger than needed or sometimes impractical constraints on the experimental design of cognitive diagnosis.

It remains an open problem what would be the minimal requirements, i.e., the necessary and sufficient conditions, for joint identifiability of the Q -matrix and model parameters. This paper addresses this problem and has the following contributions.

First, under the DINA model, we derive the necessary and sufficient conditions for 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. In plain words, these three conditions require the binary Q -matrix to be *complete* by containing an identity submatrix, to have all columns *distinct* other than the part of the identity submatrix, and to *repeatedly* contain at least three entire rows of “1” in each column. The proposed conditions not only guarantee identifiability, but also give the minimal requirements for the Q -matrix and DINA model parameters to be estimable from the observed responses. The identifiability result can be directly applied to the Deterministic Input Noisy output “Or” gate (DINO) model (Templin and Henson, 2006), due 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 that cover the DINA as a submodel.

Second, we study a weaker notation of identifiability, the so-called generic identifiability, and propose sufficient and necessary conditions for it 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 strict identifiability conditions sometimes could be too restrictive in practice. For instance, it is known that unrestricted latent class models are not strictly identifiable (Gyllenberg et al., 1994), while they are generically identifiable under certain conditions (Allman et al., 2009). However, as to 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 directly applied. It is unknown what generic identifiability conditions are needed to jointly identify the Q -matrix as well as the model parameters. In this work, we propose sufficient and necessary conditions for generic identifiability, and explicitly characterize the non-identifiable 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. In plain words, these two conditions require the binary Q -matrix to contain two *generically complete* square submatrices with all diagonal elements equal to “1”, and to addition-

ally (*repeatedly*) contain at least one entry of “1” other than the part of these two submatrices.

The rest of the paper is organized as follows. Section 2 gives an introduction to RLCMs and reviews some popular models in cognitive diagnosis. Section 3 introduces the definitions of 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 DINA model and general RLCMs, respectively. Section 6 gives some discussions. All the proofs of the theoretical results and additional simulation studies that verify the developed theory are included in the Supplementary Material. The Matlab codes for checking all the proposed conditions are available at https://github.com/yuqigu/Identify_Q.

2. RLCMs for Cognitive Diagnosis

RLCMs are key statistical tools in cognitive diagnostic assessments with the aim to estimate individuals’ attribute profiles based on their response data in the assessment. Specifically, consider a diagnostic test with J items. A subject (such as an examinee or a patient) provides a J -dimensional binary response vector $\mathbf{R} = (R_1, \dots, R_J)^\top$ 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 $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$,

where $\alpha_k \in \{0, 1\}$ is a binary indicator of the absence or presence of the k th attribute, respectively.

RLCMs assume a two-step data generating process. The first step has a population model for the attribute profile vector. We assume that the attribute profile follows a categorical distribution with population proportions $\mathbf{p} := (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0, 1\}^K)^\top$ 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 $\{R_j : j = 1, \dots, J\}$, are assumed conditionally independent, and each R_j follows a Bernoulli distribution with parameter $\theta_{j,\boldsymbol{\alpha}} = P(R_j = 1 \mid \boldsymbol{\alpha})$. The $\theta_{j,\boldsymbol{\alpha}}$ denotes the probability of a positive response, and is also called an item parameter of item j . The collection of all the item parameters, denoted by the item parameter matrix $\Theta = (\theta_{j,\boldsymbol{\alpha}})_{J \times 2^K}$, are 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 $q_{j,k} \in \{1, 0\}$ indicating whether or not the j th item is linked to the k th latent attribute. When $q_{j,k} = 1$, we say attribute k is required by item j . The j th row vector \mathbf{q}_j of Q gives the full attribute requirements of item j . Given an attribute profile $\boldsymbol{\alpha}$ and a matrix Q , we write $\boldsymbol{\alpha} \succeq \mathbf{q}_j$ if $\alpha_k \geq q_{j,k}$ for all $k \in \{1, \dots, K\}$, and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$ if there exists k such that $\alpha_k < q_{j,k}$; similarly we define the operations \preceq

and $\not\geq$.

If $\alpha \succeq \mathbf{q}_j$, a subject having attribute pattern α possesses all the attributes required by item j specified by the Q -matrix, and would be “capable” of answering item j correctly. On the other hand, if $\alpha' \not\geq \mathbf{q}_j$, the subject with α' misses some required attribute of item j and is expected to have a smaller positive response probability than those subjects with $\alpha \succeq \mathbf{q}_j$. That is, the RLCMs we consider in this paper assume

$$\theta_{j,\alpha} > \theta_{j,\alpha'} \text{ for any } \alpha \succeq \mathbf{q}_j \text{ and } \alpha' \not\geq \mathbf{q}_j. \quad (2.1)$$

Such *monotonicity assumption* in (2.1) is common to most RLCMs. Another common assumption of RLCMs is that mastering those non-required attributes of an item will not change the positive response probability to it, i.e., $\theta_{j,\alpha} = \theta_{j,\alpha'}$ if $\alpha \odot \mathbf{q}_j = \alpha' \odot \mathbf{q}_j$, where “ \odot ” denotes the elementwise multiplication operator (Henson et al., 2009). Under the introduced setup, the response vector \mathbf{R} has probability mass function in the form

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

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

Next, we review some popular cognitive diagnosis models and illustrate how they fall into the family of RLCMs.

Example 1 (DINA model). One of the basic cognitive diagnosis models is the DINA model (Junker and Sijtsma, 2001). The DINA model assumes a conjunctive relationship among attributes, meaning that to be capable of providing a positive response to an item, it is necessary to possess all its required attributes 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 $\Gamma_{j,\boldsymbol{\alpha}}^{DINA} = I(\boldsymbol{\alpha} \succeq \mathbf{q}_j)$, which indicates whether the subject is capable of item j . The uncertainty is incorporated at the item level with 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 non-capable 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 s_j and g_j for each item regardless of the number of attributes required by the item. In the following discussion, we denote $\mathbf{s} = (s_1, \dots, s_J)^\top$ and $\mathbf{g} = (g_1, \dots, g_J)^\top$. Given the Q -matrix, the DINA model parameters $(\boldsymbol{\Theta}, \mathbf{p})$ can then be equivalently expressed by $(\mathbf{s}, \mathbf{g}, \mathbf{p})$. We further assume $\mathbf{1} - \mathbf{s} \succ \mathbf{g}$ (Xu and Zhang, 2016), which makes DINA satisfy the monotonicity assumption (2.1). Identifiability results of 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. The formulation of the GDINA

model based on $\theta_{j,\alpha}$ can be decomposed into the sum of the effects due to the presence of specific attributes and their interactions. Specifically, for an item j with \mathbf{q} -vector $\mathbf{q}_j = (q_{j,k} : k = 1, \dots, K)$, the positive response probability is

$$\theta_{j,\alpha}^{GDINA} = \sum_{\mathcal{S} \subseteq \{1, \dots, K\}} \beta_{j,\mathcal{S}} \prod_{k \in \mathcal{S}} q_{j,k} \prod_{k \in \mathcal{S}} \alpha_k. \quad (2.3)$$

Note that not all β -coefficients in the above equation are included in the model. For a subset \mathcal{S} of the K attributes $\{1, \dots, K\}$, the $\beta_{j,\mathcal{S}} \neq 0$ only if $\prod_{k \in \mathcal{S}} q_{j,k} = 1$. The interpretation is that, $\beta_{j,\emptyset}$ denotes the probability of a positive response when none of the required attributes are present in α ; when $q_{j,k} = 1$, $\beta_{j,\{k\}}$ is 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 in the model, representing the change in the positive response probability due to the interaction effect of mastery of both k and k' . Under the GDINA model, each $\theta_{j,\alpha}$ models the main effects and all the interaction effects of the attributes measured by the item. For such diagnostic models, we call them *general* RLCMs. Another popular general RLCM is the Log-linear Cognitive Diagnosis Model (LCDM; Henson et al., 2009) and the General Diagnostic Model (GDM; von Davier, 2008). Identifiability results of general RLCMs are presented in Section 5.

3. Definitions and Illustrations of strict and generic identifiability

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

We would also like to point out that the monotonicity assumption stated in (2.1), is necessary for the identifiability of the Q -matrix. Since otherwise any $Q \neq \mathbf{1}_{J \times K}$ with parameters (Θ, \mathbf{p}) can not be distinguished from $\bar{Q} = \mathbf{1}_{J \times K}$ with the same parameters (Θ, \mathbf{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 to be discussed in Sections 4 and 5. In the following we assume the monotonicity assumption introduced in Section 2 is satisfied.

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

We first introduce the definition of identifiability of Q -matrix as well as the model parameters (Θ, \mathbf{p}) , which we term as *joint strict identifiability*.

Definition 1 (Joint Strict Identifiability). Under an RLCM, the design matrix Q joint with the model parameters (Θ, \mathbf{p}) are said to be strictly identifiable if for any (Q, Θ, \mathbf{p}) ,

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

there is no $(\bar{Q}, \bar{\Theta}, \bar{p}) \approx (Q, \Theta, p)$ such that

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

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

Despite being the most stringent criterion for identifiability, strict identifiability could be too restrictive, ruling out many cases where the (Q, Θ, p) are “almost surely” identifiable. In the literature of unrestricted latent class models, Allman et al. (2011) proposed and studied the so-called *generic identifiability*. Here we also introduce the concept of generic identifiability for RLCMs as follows.

Definition 2 (Joint Generic Identifiability). Consider an RLCM with parameter space ϑ_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 (Θ, p) are said to be generically identifiable, if the following set has Lebesgue measure zero in \mathbb{R}^m :
 $\vartheta_{non} = \{(\Theta, p) : \exists(\bar{Q}, \bar{\Theta}, \bar{p}) \approx (Q, \Theta, p) \text{ such that } \mathbb{P}(\mathbf{R} \mid Q, \Theta, p) = \mathbb{P}(\mathbf{R} \mid \bar{Q}, \bar{\Theta}, \bar{p})\}.$

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

We use an example to show the difference between generic identifiability and strict identifiability. Consider the Q -matrix $Q_{4 \times 2}$ in (3.5). Under the DINA model, it will

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

be proved that this Q -matrix joint with the associated model parameters $(\mathbf{s}, \mathbf{g}, \mathbf{p})$ are generically identifiable (by part (b.2) of Theorem 2), but not strictly identifiable (by Theorem 1).

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

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

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

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 providing the same distribution of the observed response vector. Here the parameter space $\boldsymbol{\vartheta}_Q = \{(\mathbf{s}, \mathbf{g}, \mathbf{p}) : \mathbf{1} - \mathbf{s} \succ \mathbf{g}, \mathbf{p} \succ \mathbf{0}, \sum_{\alpha} p_{\alpha} = 1\}$ is of full dimension in \mathbb{R}^{11} , while the non-identifiable subset $\boldsymbol{\vartheta}_{non} = \{(\mathbf{s}, \mathbf{g}, \mathbf{p}) : p_{(01)}p_{(10)} = p_{(00)}p_{(11)}\}$ 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 to be $g_j = s_j = 0.2$ 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 randomly generated, which almost always satisfy

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

(3.6). Specifically, we randomly generate a total number of 100 true parameter sets $(\mathbf{s}, \mathbf{g}, \mathbf{p})$, with 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 $\mathbf{p} \sim \text{Dirichlet}(3, 3, 3, 3)$. Here $\mathcal{U}(0.1, 0.3)$ denotes the uniform distribution on $[0.1, 0.3]$, and $\text{Dirichlet}(3, 3, 3, 3)$ denotes the Dirichlet distribution with parameter vector $(3, 3, 3, 3)$.

We show numerically that in scenario (a), there exist multiple different sets of valid DINA parameters that give the same distribution of \mathbf{R} ; while in scenario (b), the model $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are almost surely identifiable and estimable. In particular, corresponding to scenario (a), Figure 1 (a) plots the true model parameters as well as 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 the $2^4 = 16$ response patterns under these three different sets of model parameters. We can see that despite these three sets of parameters are very different, they give the identical distribution of the 4-dimensional binary response vector.

Corresponding to scenario (b), we randomly generate $B = 100$ sets of true parameters $(\mathbf{s}^i, \mathbf{g}^i, \mathbf{p}^i)$ for $i = 1, \dots, 100$. Then for each $(\mathbf{s}^i, \mathbf{g}^i, \mathbf{p}^i)$, we generate 200 independent datasets of size N with $N = 10^2, 10^3, 10^4$ and 10^5 , and then compute the Mean Square Error (MSE) of the maximum likelihood estimators (MLE) of the slipping, guessing and proportion parameters, respectively. To compute the MLE of model parameters for each simulated dataset, we run the EM algorithm with 10 random initializations and

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

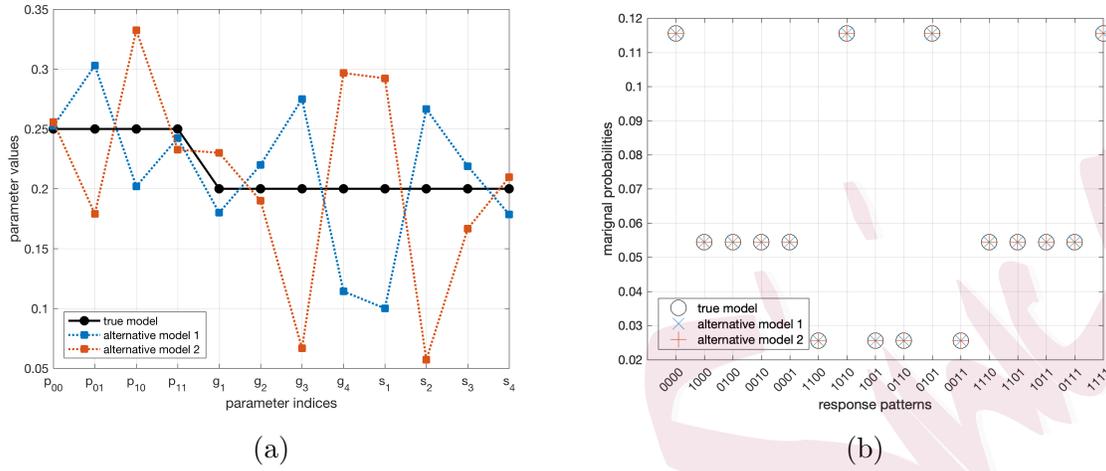


Figure 1: Illustration of non-identifiability under $Q_{4 \times 2}$ in scenario (a).

choose the estimators achieving the largest log-likelihood value out of the 10 runs. Figure 2 shows the boxplots of Mean Square Errors (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.

On the other hand, Figure 2 also shows that there do exist several parameter sets whose MSEs are “outliers” in the boxplots and converge to 0 much slower than others as N increases. This happens basically because these sets of parameters fall near the non-identifiability set, $\mathcal{V}_{non} = \{(\mathbf{s}, \mathbf{g}, \mathbf{p}) : p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$, and it becomes more difficult to identify them than others. To illustrate this point, we consider the scenario corresponding to the rightmost boxplot in Figure 2(a) with sample size $N = 10^5$. For each one of the 100 sets of true parameters $(\mathbf{s}^i, \mathbf{g}^i, \mathbf{p}^i)$, in Figure 3 we plot $p_{(00)}^i \cdot p_{(11)}^i$

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

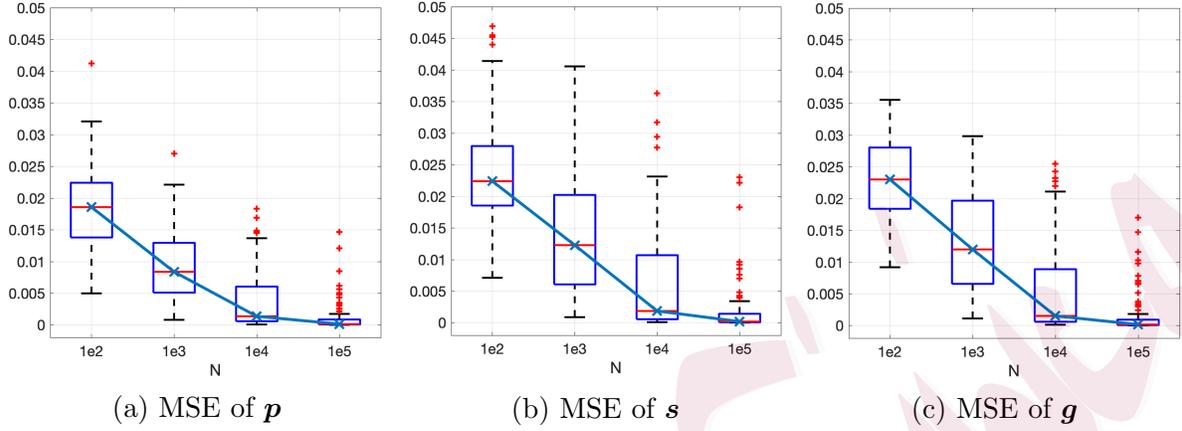


Figure 2: Illustration of generic identifiability under $Q_{4 \times 2}$, which corresponds to simulation scenario (b).

and $p_{(01)}^i \cdot p_{(01)}^i$ as the x -axis and y -axis coordinates, respectively. Then each point represents one set of true parameters used to generate the data. Specifically, we plot those parameter sets with red “*”s if their corresponding MSEs are the 20% largest outliers in the rightmost boxplot in Figure 2(a); and plot the remaining 80% parameter sets with blue “+”s. One can clearly see that the closer the true parameters lie to the non-identifiability set $\mathcal{V}_{non} = \{(\mathbf{s}, \mathbf{g}, \mathbf{p}) : p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$ (represented by the straight reference line drawn from $(0, 0)$ to $(0.17, 0.17)$), the larger the MSEs are, and the slower the convergence rate of the MLEs is. This indicates the phenomenon under generic identifiability that when the true model is close to the non-identifiable set, the convergence of their MLEs becomes slow.

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,

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

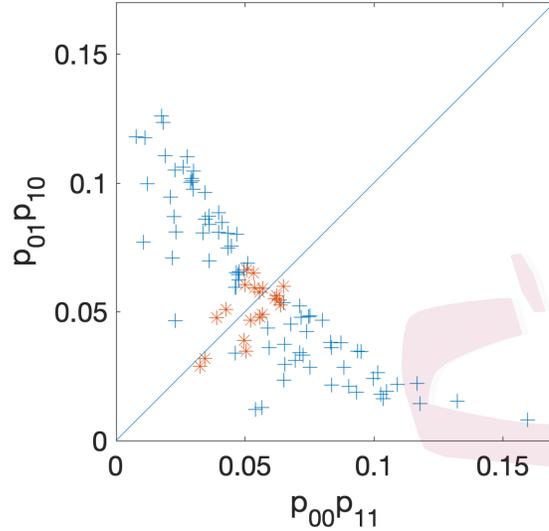


Figure 3: Illustration of impact of the generic identifiability constraint (3.6). Red “*”s represent parameter sets with the 20% largest MSEs in Figure 2(a) with $N = 10^5$; blue “+”s represent the remaining parameter sets.

view each subject’s 2-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 $p_{(00)}p_{(11)} - p_{(01)}p_{(10)}$. This means one row (resp. column) of the matrix can not 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 each attribute to be measured by at least three times for identifiability (as shown in Condition B in Theorem 1). In particular, consider those

3.1 Illustration of Generic Identifiability Phenomenon with $Q_{4 \times 2}$

attributes that are measured by only two items in the Q -matrix. If these attributes are independent, then intuitively they provide independent source of information, in which case the model is not identifiable. However, if these attributes are dependent, then the dependency instead helps with the identification of the model structure.

Before stating the strict and generic identifiability results on (Q, Θ, \mathbf{p}) , we show in the next proposition that any all-zero row vector in the Q -matrix can be dropped without impacting 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 the J' nonzero \mathbf{q} -vectors, and $\mathbf{0}$ denotes a $(J - J') \times K$ submatrix containing those zero \mathbf{q} -vectors. Let Θ' be the submatrix of Θ containing its first J' rows. Then for any RLCM, (Q, Θ, \mathbf{p}) are jointly strictly (generically) identifiable if and only if $(Q', \Theta', \mathbf{p})$ are jointly strictly (generically) identifiable.*

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

GDINA and LCDM models.

4. Identifiability of (Q, Θ, \mathbf{p}) under the DINA model

Under the DINA model, Liu et al. (2013) first studied identifiability of the Q -matrix under the assumption that the guessing parameters \mathbf{g} are known. Chen et al. (2015) and Xu and Shang (2018) further proposed a set of sufficient conditions without assuming 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. The previous studies in Chen et al. (2015) and Xu and Shang (2018) require Q to contain at least two complete submatrices I_K for identifiability.

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

Theorem 1. *Under the DINA model, the following Conditions A, B and C combined are necessary and sufficient for strict identifiability of $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$.*

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

takes the following form

$$Q = \begin{pmatrix} I_K \\ Q^* \end{pmatrix}. \quad (4.7)$$

B. The column vectors of the sub-matrix Q^ in (4.7) are distinct.*

C. Each column in Q contains at least three entries of “1”.

In the Supplementary Material, we perform simulations to verify Theorem 1. In particular, see simulation study I for the verification of the sufficiency of the conditions A , B and C for joint identifiability; also see simulation studies III and IV regarding the necessity of the proposed conditions. We provide comparisons of our Theorem 1 with some existing results. First, although the same set of conditions A , B and C were also proposed in Gu and Xu (2018b), they assumed a known Q and studied identifiability of parameters $(\mathbf{s}, \mathbf{g}, \mathbf{p})$; on the contrary, Theorem 1 studies the joint identifiability of $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$, which is theoretically much more challenging due to the unknown Q -matrix and therefore provides a much stronger result than that in Gu and Xu (2018b). In terms of estimation, Theorem 1 implies that one can consistently estimate both Q and $(\mathbf{s}, \mathbf{g}, \mathbf{p})$, without worrying about a wrong Q -matrix would be indistinguishable from the true Q . Second, Theorem 1 also has much weaker requirements than the celebrated identifiability conditions resulting from three-way tensor decomposition (Kruskal, 1977; Allman et al., 2011). Specifically, these classical results require the number of items

$J \geq 2K + 1$ for (generic) identifiability. In contrast, 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 only needs to contain another $\log_2(K)$ items whose K -dimensional \mathbf{q} -vectors form a matrix with K distinct columns. For example, for $K = 8$, conditions in Allman et al. (2011) require at least $2K + 1 = 17$ items while our Theorem 1 guarantees that the following Q with $K + \log_2(K) + 1 = 12$ items suffices for strict identifiability of $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ 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 any of them, Theorem 1 implies that there must exist $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p}) \approx (\bar{Q}, \bar{\mathbf{s}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ 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); and 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. It is therefore desirable to study conditions that guarantee the former case, i.e., generic identifiability of $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$.

In the following, we discuss necessity of Conditions A, B, C under the weaker notion of generic identifiability. First, Condition A is necessary for joint generic identifiability of (Q, Θ, \mathbf{p}) . If the true Q -matrix does not satisfy Condition A , then under 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, 2018a). In certain scenarios where Condition A fails, one can find a different \bar{Q} that is not distinguishable from Q . See simulation study IV in the Supplementary Material that illustrates the necessity of Condition A .

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

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

And it is proved in Gu and Xu (2018b) 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, differently from Conditions A and B , for generic identifiability, Condition C can be relaxed to certain extent. The next theorem characterizes how the Q -matrix structure in this case impacts generic identifiability. For empirical verification of Theorem 2, see simulation study II in the Supplementary Material.

Theorem 2. *Under the DINA model, $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are 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,

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{v}^\top \\ \mathbf{0} & Q^* \end{pmatrix}, \quad (4.8)$$

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

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

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

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

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

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

Remark 1. We say $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are 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, so the statement in part (a) of Theorem 2 also implies $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are not globally generically identifiable.

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

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

Proposition 2. Under the DINA model with $K = 2$ attributes, $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are generi-

ally 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 the possible Q -matrices with $K = 2$ such that $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are 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 $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are not generically identifiable when A or B is violated. So we only need to focus on the cases where Condition C is violated while Conditions A and B are satisfied. Specifically, when $J \leq 5$, the Q -matrix could only take the following forms up to column and row permutations,

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

By Theorem 2, Q_1 falls in scenario (a), so $(Q_1, \mathbf{s}, \mathbf{g}, \mathbf{p})$ are not locally generically identifiable, i.e., 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), so $(Q_2, \mathbf{s}, \mathbf{g}, \mathbf{p})$ and $(Q_3, \mathbf{s}, \mathbf{g}, \mathbf{p})$ 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 have some 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, Θ, \mathbf{p}) under general RLCMs

Since DINA is a submodel of the general RLCMs, Conditions A , B and C in Theorem 1 are also necessary for 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 strict identifiability of (Q, Θ, \mathbf{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 they can not be applied to identifying 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 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 generic identifiability of (Q, Θ, \mathbf{p}) , contrary to the results for the DINA model where Conditions A and B can not be

relaxed while Condition C can. See simulation studies VI and VII in the Supplementary Material for the verification of Theorem 3.

Theorem 3. *Under a general RLCM, Condition C in Theorem 1 is necessary for generic identifiability of (Q, Θ, \mathbf{p}) . Specifically, when the true Q -matrix violates C , for any model parameters (Θ, \mathbf{p}) associated with Q , there exist infinitely many sets of $(\bar{Q}, \bar{\Theta}, \bar{\mathbf{p}}) \approx (Q, \Theta, \mathbf{p})$ such that equation (3.4) holds, making (Q, Θ, \mathbf{p}) not generically identifiable.*

While Condition C is necessary, we next show that the other two conditions A and B can be further relaxed for generic identifiability of general RLCMs. Before stating the result, we first introduce a new concept about the Q -matrix, the *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 being “1”.

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, while a generically complete Q -matrix may not have any single attribute item.

With 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 not necessary any more in the current setting.

Theorem 4. *Under a general RLCM, if the true Q -matrix satisfies the following Conditions D and E , then (Q, Θ, \mathbf{p}) are 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

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q^* \end{pmatrix}_{J \times K} . \quad (5.9)$$

E. Each column of the submatrix Q^ in (5.9) contains at least one entry of “1”.*

Remark 3. Under Theorem 4, the identifiable subset of the parameter space is $\{(\Theta, \mathbf{p}) : \det(T(Q_1, \Theta_{Q_1})) \neq 0, \det(T(Q_2, \Theta_{Q_2})) \neq 0, \text{ and } T(Q^*, \Theta_{Q^*}) \cdot \text{Diag}(\mathbf{p}) \text{ has distinct column vectors}\}$. Its complement is the non-identifiable subset and it has measure

zero in the parameter space ϑ_Q , when Q satisfies Conditions D and E . Please see the supplementary materials for the definition of the T -matrices ($T(Q_1, \Theta_{Q_1})$, etc.).

Remark 4. The proof of Theorem 4 is based on the proof of Theorem 7 in Gu and Xu (2018a), which proposed the same Conditions D and E as sufficient conditions for 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 (2018a) shows that when Q is known, some attribute can be required by only two items for generic identifiability (i.e., Condition C can be relaxed); while 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. See simulation study V in the Supplementary Material for the verification of 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 joint strict identifiability of $(Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$. For general RLCMs, we have an analogous conclusion that the generic completeness of Q , which is part of Condition D , is necessary for joint generic identifiability of (Q, Θ, \mathbf{p}) .

This is stated in the next theorem.

Theorem 5. *Under a general RLCM, generic completeness of the Q -matrix is necessary for joint generic identifiability of (Q, Θ, \mathbf{p}) .*

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

Proposition 3. *For a general RLCM with $K = 2$, Conditions D and E are necessary and sufficient for generic identifiability of (Q, Θ, \mathbf{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 which 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,

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

The “*”s in Q_2 are unspecified values, either 0 or 1. For Q_1 with $J = 3$, $K = 2$ and any parameters (Θ, \mathbf{p}) , there are $2^J = 8$ constraints in (3.4) for solving $(\bar{\Theta}, \bar{\mathbf{p}})$ under

Q_1 itself, while the number of free parameters of $(\bar{\Theta}, \bar{\mathbf{p}})$ is $|\{p_\alpha : \alpha \in \{0, 1\}^2\} \cup \{\theta_{j,\alpha} : j \in \{1, 2\}, \alpha \in \{0, 1\}^2\}| = 2^K + 2^K \times J = 16 > 8$. For Q_2 with $J = 4$, $K = 2$ and any associated (Θ, \mathbf{p}) , there are $2^J = 16$ constraints in (3.4) for solving $(\bar{\Theta}, \bar{\mathbf{p}})$, while the number of free parameters of $(\bar{\Theta}, \bar{\mathbf{p}})$ 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, so neither $(Q_1, \Theta, \mathbf{p})$ nor $(Q_2, \Theta, \mathbf{p})$ are generically identifiable.

6. Discussion

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 strict joint identifiability of the Q -matrix and the associated model parameters. We also study a slightly weaker identifiability notion, generic identifiability, and propose sufficient and necessary conditions for it under the DINA model and general RLCMs, respectively.

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 model parameters can be jointly estimated from data consistently. If the Q -matrix and 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 model parameters can be jointly estimated from data consistently.

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 as Proposition 1 in Xu and Shang (2018). Denote the true Q -matrix and model parameters by Q^0 and $\eta^0 = (\Theta^0, \mathbf{p}^0)$. Consider a sample with N i.i.d. response vectors $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_N$, and denote the log-likelihood of the sample by $\ell(\Theta, \mathbf{p}) = \sum_{i=1}^N \log \mathbb{P}(\mathbf{R}_i \mid Q, \Theta, \mathbf{p})$. Under a specified RLCM, a Q -matrix determines the structure of the item parameter matrix Θ , by specifying which entries of it are equal. For a given Θ , we can define an equivalent formulation of it, a sparse matrix \mathbf{B} having the same size as Θ , in the following way. Under a general RLCM such as the GDINA model in Example 2, the item parameters can be parameterized as $\theta_{j,\alpha} = \sum_{\mathcal{S} \subseteq \{1, \dots, K\}} \beta_{j,\mathcal{S}} \prod_{k \in \mathcal{S}} \alpha_k$. Based on this, we define the j th row of \mathbf{B} as a 2^K -dimensional vector collecting all these β -coefficients; that is, $\mathbf{B}_j = (\beta_{j,0}, \beta_{j,1}, \dots, \beta_{j,K}, \dots, \beta_{j,12\dots K})$. Then as long as the \mathbf{q} -vector $\mathbf{q}_j \neq \mathbf{1}_K$, the vector \mathbf{B}_j would be “sparse” and so is the matrix \mathbf{B} . For the true Q^0 , we denote the corresponding \mathbf{B} -matrix by \mathbf{B}^0 . Under a specified RLCM (such as DINA or GDINA), the identification of Q^0 is then implied by the identification of the indices of

nonzero elements of \mathbf{B}^0 . Denote the support of the true \mathbf{B}^0 and any candidate \mathbf{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 \mathbf{R} indexed by parameters $\boldsymbol{\eta}^0$ under the true \mathbf{B}^0 and $\boldsymbol{\eta}$ under the candidate \mathbf{B} . Denote the Q -matrix and the model parameters that maximize the log-likelihood $\ell(\boldsymbol{\Theta}, \mathbf{p})$ subject to the L_0 constraint $|S| \leq |S_0|$ by $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\Theta}}, \hat{\mathbf{p}})$, and denote the ‘‘oracle’’ MLE of model parameters obtained assuming Q^0 is known by $\hat{\boldsymbol{\eta}}^0 = (\hat{\boldsymbol{\Theta}}^0, \hat{\mathbf{p}}^0)$. 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 $C_{\min}(\boldsymbol{\Theta}^0, \mathbf{p}^0) \geq c_0$ for some positive constant c_0 . Furthermore,*

$$\mathbb{P}(\hat{Q} \neq Q^0) \leq \mathbb{P}(\hat{\boldsymbol{\eta}} \neq \hat{\boldsymbol{\eta}}^0) \leq c_2 \exp\{-c_1 N C_{\min}(\boldsymbol{\Theta}^0, \mathbf{p}^0)\}, \quad (6.10)$$

where $c_1, c_2 > 0$ are some constants. Namely, when 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 $C_{\min}(\boldsymbol{\Theta}^0, \mathbf{p}^0)$ in (6.10) equal to zero, instead of bounded away from zero as shown in Proposition 4. Therefore, the finite sample

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

Potential extensions to other latent variable models. We briefly discuss the potential extensions of the developed theory to some other latent variable models, such as restricted latent class models 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 considered 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 $\{0, 1, \dots, C-1\}$, then γ can be equivalently viewed as a collection of $C-1$ binary random variables $\boldsymbol{\alpha}^\gamma := (\alpha_1, \dots, \alpha_{C-1})$ with the following constraints. If $\alpha_i = 0$ for some $i < C-1$, then $\alpha_j = 0$ for all $j = i+1, \dots, C-1$. 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 the polytomous attributes can be augmented to a longer vector of binary attributes with constraints in this fashion. Then we can consider the

restricted latent class model with the augmented proportion parameters, by constraining the proportions of those “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 of establishing identifiability for latent class models in this paper would also be useful when treating the discrete latent variables. For the continuous latent variables, the techniques developed in Bai and Li (2012) for identifiability of the factor analysis model and those developed in traditional multivariate analysis (Anderson, 2009) would be helpful.

In practice, the newly developed identifiability theory can serve as the 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 one can use likelihood-based approaches, such as that in Xu and Shang (2018), to jointly estimate Q and model parameters by constraining Q to the identifiable Q -set; or one can use Bayesian approaches to estimate Q such as that 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 (2018a), certain attribute profiles would be equivalent and the strongest possible identifiability argument therein is the so-called \mathbf{p} -partial

identifiability. In this scenario, it would be interesting to study the identifiability of the incomplete Q -matrix under the notion of \mathbf{p} -partial identifiability, and we leave this for future study.

Supplementary Materials

The online supplementary material contains the proofs of Proposition 1, 4, Theorems 1–5, and additional simulation results.

Acknowledgements

This work was supported in part by National Science Foundation grants 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.

REFERENCES

- 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.

REFERENCES

- 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. (2018a). Partial identifiability of restricted latent class models. *arXiv preprint arXiv:1803.04353*.
- Gu, Y. and Xu, G. (2018b). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*. 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

REFERENCES

- 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.

REFERENCES

- 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