PSYCHOMETRIKA—VOL. 87, NO. 4, 1343–1360
DECEMBER 2022
https://doi.org/10.1007/s11336-022-09852-y

IDENTIFIABILITY OF LATENT CLASS MODELS WITH COVARIATES

JING OUYANG AND GONGJUN XU

UNIVERSITY OF MICHIGAN

Latent class models with covariates are widely used for psychological, social, and educational research. Yet the fundamental identifiability issue of these models has not been fully addressed. Among the previous research on the identifiability of latent class models with covariates, Huang and Bandeen-Roche (Psychometrika 69:5–32, 2004) studied the local identifiability conditions. However, motivated by recent advances in the identifiability of the restricted latent class models, particularly cognitive diagnosis models (CDMs), we show in this work that the conditions in Huang and Bandeen-Roche (Psychometrika 69:5–32, 2004) are only necessary but not sufficient to determine the local identifiability of the model parameters. To address the open identifiability issue for latent class models with covariates, this work establishes conditions to ensure the global identifiability of the model parameters in both strict and generic sense. Moreover, our results extend to the polytomous-response CDMs with covariates, which generalizes the existing identifiability results for CDMs.

Key words: identifiability, latent class models, cognitive diagnosis models.

Latent class models are extensively applied in numerous scientific fields, including educational assessments, biological research, and psychological measurements, to infer the latent subgroups of a population as well as each subject's latent classification information. For instance, one application of latent class models in cognitive diagnosis is to classify individuals with different latent attributes based on their observed responses to items, for which reason they are key components in educational measurements (Junker & Sijtsma, 2001; von Davier Matthias, 2008), psychiatric evaluations (Templin & Henson, 2006), and disease detections (Wu et al., 2017). In addition to understanding the basic parameters in latent class models, researchers are also interested in studying the relations between latent class parameters with the observed covariates, such as subjects' gender, race, education level, and other characteristics (Formann, 1985; Collins & Lanza, 2009; Huang & Bandeen-Roche, 2004).

Latent class models with covariates can help to improve the classification accuracy of the latent classes and are useful in testing whether the covariates are related to the latent class membership probability or response probability. Such latent class models involving covariates have been studied in many works in psychometrics and statistics literature, where covariates were mostly constrained to be discrete at early stage (Clogg & Goodman, 1984; Formann, 1985), and further extended to be in general forms (Dayton & Macready, 1988; van der Heijden et al., 1996; Muthén & Muthén, 2017). The models have been popularly applied in educational, psychological, and behavioral sciences (Collins & Lanza, 2009; Muthén & Masyn, 2005; Reboussin et al., 2008; Bakk et al., 2013; Park et al., 2018). The related estimation problems have also received great interest from researchers in the psychometrics field, such as estimating the covariate coefficients (Petersen et al., 2012), adjusting for the bias in the estimation (Bakk et al., 2013), and estimating the number of latent classes (Huang, 2005; Pan & Huang, 2014).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11336-022-09852-y.

Correspondence should be made to Gongjun Xu, Department of Statistics, University of Michigan, 456 West Hall, 1085 South University, Ann Arbor, MI48109, USA. Email: gongjun@umich.edu

For latent class models with or without covariates, identifiability is one of the most fundamental issues as it is the prerequisite for parameter estimations and statistical inferences. Identifiability could be interpreted as the feasibility of recovering the model parameters based on observed responses, i.e., the parameters in identifiable models should be distinct given the probabilistic distribution of the observations. A rich body of literature works have studied identifiability issues, dating back to Koopmans (1950) and Koopmans and Reiersol (1950). Specifically, McHugh (1956) proposed conditions to determine the local identifiability for the binary-response latent class models, and Goodman (1974) further extended the local identifiability conditions to the polytomous-response models. In the sense of strict identifiability, Gyllenberg et al. (1994) found that the binary-response latent class models cannot be strictly identifiable. Nonetheless, Allman et al. (2009) considered the concept of generic identifiability and established sufficient conditions for the generic identifiability of latent class models, where a model is said to be generically identifiable if the model parameters are identifiable except for a measure-zero set of parameters. However, their generic identifiability conditions can only be applied to the unrestricted latent class models, but not directly to the restricted latent class models. To address this issue, Xu (2017) and Xu and Shang (2018) established the results for the identifiability of the Q-restricted binary-response latent class models. For the polytomous-response models, Culpepper (2019) and Fang, Liu, and Ying (2019) established strict identifiability conditions based on the algebraic theorems proposed by Kruskal (1977). Moreover, Gu and Xu (2020) studied the generic and partial identifiability of the restricted binary-response latent class models and extended their conditions to the polytomous-response models as well.

Among existing research, most focus on the identifiability of general or restricted latent class models without covariates, whereas few investigate the identifiability of latent class models with covariates. As the observed covariates represent characteristics of certain homogeneous groups, incorporating covariates into latent class models would help to explain the association of these characteristics with latent classes. The regression latent class models with covariates are general extensions of latent class models without covariates. In other words, the regular or restricted latent class models can be viewed as a special family of latent class models with covariates, where all covariates values are zero. Technically speaking, existing identifiability results for regular or restricted latent class models cannot be directly applied to the regression latent class models due to the existence of covariates, and new techniques are needed to establish the identifiability of the corresponding regression coefficients for those covariates, which do not exist in the regular or restricted latent class models. In the literature, Huang and Bandeen-Roche (2004) was among the first to study the identifiability of latent class models with covariates. The authors studied the local identifiability conditions for the model parameters, that is, the conditions to ensure that the model parameters are identifiable in a neighborhood of the true parameters.

However, as to be shown in the paper, the proposed identifiability conditions in Huang and Bandeen-Roche (2004) are only necessary but not sufficient for the local identifiability of latent class models. Our argument borrows ideas from the recent developments in the identifiability of cognitive diagnosis models (CDMs), a special family of the restricted latent class models. Besides, the results in Huang and Bandeen-Roche (2004) only concern the local identifiability but not the global identifiability. In light of these, our work establishes identifiability conditions to check the global identifiability for latent class models with covariates. Furthermore, we also establish the identifiability results for CDMs with covariates, which is a special family of the regression latent class models. Our results extend many identifiability conditions for the binary-response CDMs to the polytomous-response CDMs with covariates, and these conditions are beyond results in the existing literature related to CDMs identifiability (Xu, 2017; Culpepper, 2019; Gu & Xu, 2020).

The organization of this paper is as follows. Section 1 introduces the setup of the regression latent class models with covariates as well as the regression CDMs and reviews some existing identifiability results. Section 2 discusses the necessity and sufficiency of the existing identifiability conditions for the regression latent class models. Section 3 presents our main results for both strict and generic identifiability of the regression latent class models as well as the regression CDMs. Section 4 uses a Trends in Mathematics and Science Study (TIMSS) dataset as an example to illustrate the application of the identifiability results in educational assessments. Section 5 gives a discussion. The proofs for the main theorems and propositions are provided in the Supplementary Material.

1. Model Setup and Existing Works

1.1. Regression Latent Class Models (RegLCMs)

We start with the setup of latent class models without covariates. Suppose there are N subjects responding to J items. The response of subject i is denoted as $\boldsymbol{R}_i = (R_{ij};\ j = 1, \ldots, J)$, where R_{ij} denotes the response of subject i to item j, for $i = 1, \ldots, N$. And $R_{ij} \in \{0, \ldots, M_j - 1\}$, where M_j denotes the number of possible values for R_{ij}. Denote $\mathcal{S} = \times_{j=1}^{J}\{0, \ldots, M_j - 1\}$ as the set of all response patterns, and its cardinality is denoted as $S = |\mathcal{S}| = \prod_{j=1}^{J} M_j$. The case at $M_j = 2$ corresponds to the binary-response models. Consider there are C latent classes and denote L_i as the latent class membership for subject i. Assume the N subjects are independent; for $c = 0, \ldots, C - 1$, $L_i = c$ implies that the subject i is in the cth latent class category and $\eta_c = P(L_i = c)$ defines the latent class membership probability, i.e., the probability for subject i being in the cth latent class. The latent class membership probabilities are summarized as $\boldsymbol{\eta} = (\eta_c; c = 0, \ldots, C-1)$. For any $j = 1, \ldots, J, r = 0, \ldots, M_j - 1$, and $c = 0, \ldots, C-1$, we use $\theta_{jrc} = P(R_{ij} = r \mid L_i = c)$ to denote the conditional response probability, i.e., the probability of the response to item j being r given the subject i is in the cth latent class. Let the vector $\boldsymbol{\theta_{jc}} = (\theta_{j0c}, \ldots, \theta_{j(M_j-1)c})$ to denote the probability vector for item j given the latent class membership c. The conditional response probabilities are summarized as $\boldsymbol{\Theta} = (\boldsymbol{\theta_{jc}};\ j = 1, \ldots, J, c = 0, \ldots, C-1)$. The conditional probability mass function for R_{ij} is $P(R_{ij} \mid L_i = c, \boldsymbol{\theta_{jc}}) = \prod_{r=0}^{M_j-1} \theta_{jrc}^{\mathbb{I}\{R_{ij}=r\}}$, and the probability mass function of $\boldsymbol{R}_i$ is

$$P\left(\boldsymbol{R}_i \mid \boldsymbol{\eta}, \boldsymbol{\Theta}\right) = \sum_{c=0}^{C-1} \eta_c \prod_{j=1}^{J} P\left(R_{ij} \mid L_i = c, \boldsymbol{\theta_{jc}}\right) = \sum_{c=0}^{C-1} \eta_c \prod_{j=1}^{J} \prod_{r=0}^{M_j-1} \theta_{jrc}^{\mathbb{I}\{R_{ij}=r\}}.$$

To introduce the regression latent class models, following the model setting in Huang and Bandeen-Roche (2004), we let the latent class membership probability η_c's and the conditional response probability θ_{jrc}'s to be functionally dependent on covariates. Denote $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ to be the covariates of subject i, where $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^T_{(p+1)\times 1}$ are the primary covariates related to the latent class membership η_c for $c = 0, \ldots, C-1$, and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iJ})^T_{J\times q}$ with $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijq})^T_{q\times 1}$ are the secondary covariates associated with the conditional response probability θ_{jrc} for any $j = 1, \ldots, J$, $r = 0, \ldots, M_j - 1$, and $c = 0, \ldots, C-1$. The x_{it} and z_{ijs} can be categorical covariates representing gender, race, or marital status. They can also be continuous, such as the subject's age. As in some applications, we may have certain prior knowledge on the set of covariates related to η_c and that of the covariates related to θ_{jrc}, where the two sets may or may not contain the same covariates. Hence, we follow the general framework in Huang and Bandeen-Roche (2004) by applying different notations, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, to distinguish the covariates related to η_c and θ_{jrc}, while allowing the $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ to have some overlapped covariates.

Before presenting the generalized linear model framework, we need to clarify some notations. In models without covariates, e.g., latent class models or CDMs to be discussed in Sect. 1.2, we use η_c and θ_{jrc} to denote the corresponding latent class membership probability and conditional response probability, respectively. When covariates are involved in models, the parameters are dependent on the covariates. In this situation, we denote $\eta_c^i = P(L_i = c \mid \boldsymbol{x}_i, \boldsymbol{z}_i)$ to be the latent class membership probability for subject i and $\theta_{jrc}^i = P(R_{ij} = r \mid L_i = c, \boldsymbol{x}_i, \boldsymbol{z}_i)$ to be the conditional response probability for subject i, for $i = 1, \ldots, N$.

Under the framework of generalized linear model, we use logit link function to relate η_c^i's and θ_{jrc}^i's to covariates $(\boldsymbol{x}_i, \boldsymbol{z}_i)$. We let the log-odds be linearly dependent on the covariates and characterize the RegLCMs by the following equations:

$$\log\left(\frac{\eta_c^i}{\eta_0^i}\right) = \beta_{0c} + \beta_{1c}x_{i1} + \cdots + \beta_{pc}x_{ip}, \tag{1}$$

for $i = 1, \ldots, N$, $c = 1, \ldots, C-1$, and

$$\log\left(\frac{\theta_{jrc}^i}{\theta_{j0c}^i}\right) = \gamma_{jrc} + \lambda_{1jr}z_{ij1} + \cdots + \lambda_{qjr}z_{ijq}, \tag{2}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, J$, $r = 1, \cdots, M_j - 1$ and $c = 0, \ldots, C-1$, where β, γ, λ are regression coefficient parameters. We want to point out that the identifiability conditions to be shown in Sect. 3 still hold for RegLCMs when the logarithmic function in (1) and (2) is replaced with other monotonic functions. The key component in establishing the identifiability conditions for the coefficient parameters is the function monotonicity, which build the bijective mapping between identifiable $(\boldsymbol{\eta}, \boldsymbol{\Theta})$ and identifiable $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$. In this work, without loss of generality, we shall focus on the popularly used logit link function.

From (1) and (2), we equivalently express η_c^i and θ_{jrc}^i as

$$\eta_c^i = \frac{\exp(\beta_{0c} + \beta_{1c}x_{i1} + \cdots + \beta_{pc}x_{ip})}{1 + \sum_{l=1}^{C-1}\exp(\beta_{0l} + \beta_{1l}x_{i1} + \cdots + \beta_{pl}x_{ip})}, \tag{3}$$

for $i = 1, \ldots, N$, $c = 0, \ldots, C-1$, and

$$\theta_{jrc}^i = \frac{\exp(\gamma_{jrc} + \lambda_{1jr}z_{ij1} + \cdots + \lambda_{qjr}z_{ijq})}{1 + \sum_{s=1}^{M_j-1}\exp(\gamma_{jsc} + \lambda_{1js}z_{ij1} + \cdots + \lambda_{qjs}z_{ijq})}, \tag{4}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, J$, $r = 0, \cdots, M_j - 1$ and $c = 0, \ldots, C-1$. From the above expressions, we see that η_c^i and θ_{jrc}^i are functionally dependent on the linear functions $\boldsymbol{x}_i^T\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_{jc} + \boldsymbol{z}_{ij}^T\boldsymbol{\lambda}_j$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_c; c = 0, \ldots, C-1)_{(p+1)\times C}$ with $\boldsymbol{\beta}_c = (\beta_{lc}; l = 0, \ldots, p)^T_{(p+1)\times 1}$, $\boldsymbol{\gamma}_{jc} = (\gamma_{jrc}; r = 0, \ldots, M_j - 1)_{1\times M_j}$, and $\boldsymbol{\lambda}_j = (\boldsymbol{\lambda}_{jr}; r = 0, \ldots, M_j - 1)_{q\times M_j}$ with $\boldsymbol{\lambda}_{jr} = (\lambda_{ljr}; l = 1, \ldots, q)^T_{q\times 1}$.

Here following Huang and Bandeen-Roche (2004), in the conditional probability model (1), the regression parameters (β) are latent class specific. In the conditional probability model (2), we allow the intercept parameters (γ) dependent on the latent class, the response level, and the item index, while the regression coefficients parameters (λ) are dependent on the response level and the item index but not the latent class membership, which, as pointed in Huang and Bandeen-Roche (2004), is a logical assumption to prevent possible misclassification by adjusting for the associated covariates. The following two assumptions proposed by Huang and Bandeen-Roche (2004) hold for all RegLCMs.

1. The latent class membership probability η_c^i is dependent on $\boldsymbol{x}_i$ only and the conditional response probability θ_{jrc}^i is dependent on z_i only:

$$P\left(L_i = c \mid \boldsymbol{x}_i, \boldsymbol{z_i}\right) = P\left(L_i = c \mid \boldsymbol{x}_i\right);$$
$$P\left(R_{i1} = r_1, \ldots, R_{iJ} = r_J \mid L_i, \boldsymbol{x_i}, \boldsymbol{z_i}\right) = P\left(R_{i1} = r_1, \ldots, R_{iJ} = r_J \mid L_i, \boldsymbol{z_i}\right).$$

2. The measurements for different items are independent given the latent class and z_i (that is, the local independence assumption):

$$P\left(R_{i1} = r_1, \ldots, R_{iJ} = r_J \mid L_i, \boldsymbol{z_i}\right) = \prod_{j=1}^{J} P\left(R_{ij} = r_j \mid L_i, \boldsymbol{z_i}\right).$$

When the coefficients $\beta_{1c}, \ldots, \beta_{pc}$ in (1) and $\lambda_{1jr}, \ldots, \lambda_{qjr}$ in (2) are zeros, RegLCMs will be reduced to latent class models without covariates, which is a special case in the family of RegLCMs. Next, we will introduce a special family of RegLCMs, cognitive diagnosis models (CDMs), which is a family of the restricted latent class models and has been substantially studied in educational and psychological measurement. From there, we further introduce the regression CDMs. The two special RegLCMs (CDMs and regression CDMs) are important in the subsequent discussions about the identifiability conditions for RegLCMs.

1.2. Cognitive Diagnosis Models as Special RegLCMs

In CDMs, each latent class corresponds to a distinct vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \mathcal{A} = \{0, 1\}^K$ where $\alpha_1, \ldots, \alpha_K$ denote K binary latent attributes, respectively, and $\mathcal{A}$ denotes the attribute pattern space. The vector $\boldsymbol{\alpha}$ represents a unique latent profile with the kth entry $\alpha_k = 1$ implying the mastery of the subject on the kth latent attribute and $\alpha_k = 0$ implying his deficiency of it. The number of latent classes is $C = |\mathcal{A}| = 2^K$. For notational convenience, we follow the idea in Culpepper (2019) by introducing a tool vector $\boldsymbol{v} = (2^{K-1}, 2^{K-2}, \ldots, 1)^T$ and denote the latent class membership as $L = \boldsymbol{\alpha}^T \boldsymbol{v} = c \in \{0, \ldots, 2^K - 1\}$. The key characteristics of CDMs are their introduction of the latent attributes and let the combinations of mastery or deficiency of each attribute to represent the latent class memberships in the restricted latent class models.

The relationship between the response $\boldsymbol{R} = (R_1, \ldots, R_J)$ and the attribute profile $\boldsymbol{\alpha}$ for any subject could be summarized through a binary matrix $Q_{J \times K}$. Denote the jth row in Q-matrix to be $\boldsymbol{q}_j = (q_{j1}, \ldots, q_{jK})$, where $q_{jk} \in \{0, 1\}$ and $q_{jk} = 1$ means that the kth attribute is required for subjects to solve item j. Similar to RegLCMs, we consider the general polytomous responses $R_j \in \{0, \ldots, M_j - 1\}$. Given a subject's latent profile $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^T \boldsymbol{v} = c$, each R_j follows a categorical distribution with the probability vector to be $\boldsymbol{\theta_{jc}} = (\theta_{j0c}, \ldots, \theta_{j(M_j-1)c})$, where $\theta_{jrc} = P(R_j = r \mid \boldsymbol{\alpha}^T \boldsymbol{v} = c)$ is the probability for getting response value r in item j. The conditional probability mass function for R_j is $P(R_j \mid \boldsymbol{\alpha}^T \boldsymbol{v} = c,\ \boldsymbol{\theta_{jc}}) = \prod_{r=0}^{M_j-1} \theta_{jrc}^{\mathbb{I}\{R_j = r\}}$, and the probability mass function for $\boldsymbol{R}$ is

$$P\left(\boldsymbol{R} \mid \boldsymbol{\eta}, \boldsymbol{\Theta}\right) = \sum_{c=0}^{2^K-1} P\left(\boldsymbol{\alpha}^T \boldsymbol{v} = c\right) \prod_{j=1}^{J} P\left(R_j \mid \boldsymbol{\alpha}^T \boldsymbol{v} = c, \boldsymbol{\theta_{jc}}\right) = \sum_{c=0}^{2^K-1} \eta_c \prod_{j=1}^{J} \prod_{r=0}^{M_j-1} \theta_{jrc}^{\mathbb{I}\{R_j = r\}}.$$

Following the generalized DINA (G-DINA) model framework, we decompose the log-odds of θ_{jrc} into a sum of attribute effects as follows. This framework was introduced in Torre (2011) for G-DINA model with binary responses and extended to G-DINA with polytomous responses in J. Chen and de la Torre (2018). Specifically, given a latent profile $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, we have

$$\log\Big(\frac{\theta_{jrc}}{\theta_{j0c}}\Big) = b_{jr0} + \sum_{k=1}^{K} b_{jr1} q_{jk}\alpha_k + \sum_{k'=k+1}^{K} \sum_{k=1}^{K-1} b_{jrkk'}(q_{jk}\alpha_k)(q_{jk'}\alpha_{k'}) \\ + \cdots + b_{jr12\ldots K} \prod_{k=1}^{K} q_{jk}\alpha_k, \tag{5}$$

where $b_{jr0}, b_{jr1}, \ldots, b_{jrK}, b_{jr12}, \cdots, b_{jr(K-1)K}, \ldots, b_{jr12\ldots K}$ are the coefficients in the generalized linear regression of the log-odds of conditional response probability on all latent attribute mastery situations, that is, all the subsets of $\{q_{j1}\alpha_1, \ldots, q_{jK}\alpha_K\}$. Specifically, b_{jr0} is the intercept of the log-odds; $b_{jr1}, \ldots, b_{jrK}$ are the main effects of attributes, representing the change of log-odds due to the mastery of the single attribute of $\alpha_1, \ldots, \alpha_K$, respectively; $b_{jr12}, \ldots, b_{jr(K-1)K}, \ldots, b_{jr12\ldots K}$ are the interaction effects of attributes, representing the change of log-odds due to the mastery of the combination of two or more attributes of $\alpha_1, \ldots, \alpha_K$.

For subjects with covariates values being zeros, the log-odds in G-DINA model (5) is equivalent to general log-odds setting (2), which is the log-odds for RegLCMs and written as

$$\log\Big(\frac{\theta_{jrc}}{\theta_{j0c}}\Big) = \gamma_{jrc} + \lambda_{1jr} z_{j1} + \cdots + \lambda_{qjr} z_{jq} = \gamma_{jrc} + 0 + \cdots + 0 = \gamma_{jrc}, \tag{6}$$

which could be further expressed as

$$\theta_{jrc} = \frac{\exp(\gamma_{jrc})}{1 + \sum_{s=1}^{M_j-1} \exp(\gamma_{jsc})}.$$

for $j = 1, \ldots, J$, $r = 1, \ldots, M_j-1$ and $c = 0, \ldots, C-1$. We can show (5) and (6) are equivalent. Because in (5), the log-odds of conditional response probability are linear combinations of all the subsets of $\{q_{j1}\alpha_1, \ldots, q_{jK}\alpha_K\}$ and are dependent on latent profile $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ only, equivalently dependent on c at $c = \boldsymbol{\alpha}^T \boldsymbol{v}$. When covariates are zeros, the latent class category information is entirely captured by the intercept γ_{jrc} in (6), implying that for given j and r, each γ_{jrc} is bijectively corresponding to $\boldsymbol{\alpha} \in \mathcal{A}$, which further implies that there exists a bijective linear correspondence between $\{b_{jr0}, b_{jr1}, \ldots, b_{jrK}, b_{jr12}, \ldots, b_{jr(K-1)K}, \ldots, b_{jr12\ldots K}\}$ and $\{\gamma_{jrc} : c = 0, \ldots, C-1\}$.

When covariates are involved in CDMs, we introduce the regression CDMs (RegCDMs) by Eqs. (7) and (8) adapted from (1) and (2), with the additional characteristics of CDMs that each latent membership c is represented by a latent profile $\boldsymbol{\alpha}$. To make notations clear in this case, we denote the latent attributes of subject i as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})$ for $i = 1, \ldots, N$. And similarly as in RegLCMs, we use $\eta_c^i = P(\boldsymbol{\alpha}_i^T \boldsymbol{v} = c \mid \boldsymbol{x}_i, z_i)$ to denote the latent class membership probability for subject i and use $\theta_{jrc}^i = P(R_{ij} = r \mid \boldsymbol{\alpha}_i^T \boldsymbol{v} = c, \boldsymbol{x}_i, z_i)$ to denote the conditional response probability for subject i when these parameters are dependent on covariates.

Assuming that the latent membership $c = 0$ denotes the latent profile that the subject i does not master any of K attributes, i.e., $\boldsymbol{\alpha}_i = \mathbf{0}_{K\times 1}$, we model

$$\log\Big(\frac{\eta_c^i}{\eta_0^i}\Big) = \beta_{0c} + \beta_{1c} x_{i1} + \cdots + \beta_{pc} x_{ip}, \tag{7}$$

for $i = 1, \dots, N$ and $\boldsymbol{\alpha}_i^T \boldsymbol{v} = c$ with $\boldsymbol{\alpha}_i \in \{0, 1\}^K \setminus \mathbf{0}_{K\times 1}$, and

$$\log\Big(\frac{\theta^i_{jrc}}{\theta^i_{j0c}}\Big) = \gamma_{jrc} + \lambda_{1jr} z_{ij1} + \cdots + \lambda_{qjr} z_{ijq}, \tag{8}$$

for $i = 1, \dots, N$, $j = 1, \dots, J$, $r = 1, \dots, M_j - 1$, and $\boldsymbol{\alpha}_i^T \boldsymbol{v} = c$ with $\boldsymbol{\alpha}_i \in \{0, 1\}^K$. RegCDMs combine the regression setting on covariates from RegLCMs and the latent attribute representation from CDMs, which is to use binary latent profiles to represent latent classes. In addition, Assumptions 1 and 2 in Sect. 1.1 are also assumed for RegCDMs.

1.3. Identifiability Conditions in the Existing Literature

Before discussing our main results for the identifiability of the models introduced in Sects. 1.1 and 1.2, we give a review of the existing studies. The identifiability conditions for latent class models have been extensively investigated in the existing literature. In particular, McHugh (1956) studied the binary-response latent class models and proposed sufficient local identifiability conditions. Extending McHugh's work, Goodman (1974) presented a fundamental method to determine the local identifiability of the polytomous-response latent class models, stating that if the Jacobian matrix formed by the derivatives of response probability vector with respect to parameters has full column rank, then the parameters are locally identifiable. This condition is intuitively straightforward but empirically nontrivial to apply. When the number of latent class C or the number of possible responses to items M_j increases, the dimension of the Jacobian matrix would increase at a fast rate. Moreover, this method could only guarantee the local identifiability for latent class models but leave the global identifiability undiscussed.

To study global identifiability, Kruskal (1977) established algebraic results to ensure the uniqueness of factors in the decomposition of a three-way array. This work defined Kruskal rank which is analogous to the normal rank of a matrix. And it proved that if the Kruskal ranks of a triple product of matrices satisfy a certain arithmetic condition, the matrix decomposition will be unique. Based on Kruskal's theorems, Allman et al. (2009) extended the conditions to the decomposition into more than three variates and used them in the identifiability conditions for the latent class models with finite items. Besides, Allman et al. (2009) argued that even the parameters are not identifiable, the inference on parameters can be valid empirically when the model is generically identifiable, that is, the parameters are identifiable except for a zero-measure set of parameters. The generic identifiability results allow us to circumvent the complex calculation on the column rank of the Jacobian matrix.

In the recent literature, the identifiability of the restricted latent class models, such as CDMs, has also been studied. Related identifiability results on restricted models with binary responses were developed in Chen et al. (2015), Xu and Zhang (2016), Xu (2017), Xu and Shang (2018), Gu and Xu (2019), Gu and Xu (2020), etc. For the restricted latent class models with polytomous responses, Culpepper (2019), Fang et al. (2019), Chen et al. (2020), and Gu and Xu (2020) proposed the identifiability conditions dependent on the Q-matrix.

The above research focuses on the identifiability of the general or restricted latent class models without covariates. For the identifiability of latent class models with covariates, Huang and Bandeen-Roche (2004) generalized the result of Goodman (1974) and derived local identifiability conditions for RegLCMs. Under the setting of RegLCMs, denote $\mathcal{S}'$ as the response pattern space $\mathcal{S}$ with a reference pattern removed (e.g., $\mathbf{0}$), so the number of distinct response patterns in $\mathcal{S}'$ is then $S - 1$. Define

$$\boldsymbol{\Phi} = \big(\boldsymbol{\phi}_c; c = 0, \dots, C-1\big)_{(S-1)\times C},$$

where each column $\boldsymbol{\phi}_c$ is of dimension $S-1$ in which each element corresponds to a response pattern $\boldsymbol{r}=(r_1,\ldots,r_J)\in\mathcal{S}'$ and is defined as

$$\phi_{\boldsymbol{r}c}=P(\boldsymbol{R}=\boldsymbol{r}\mid L=c,\boldsymbol{z}=\boldsymbol{0})=\prod_{j=1}^{J}\frac{e^{\gamma_{jr_jc}}}{1+\sum_{s=1}^{M_j-1}e^{\gamma_{jsc}}},\tag{9}$$

where γ_{jr_jc} are defined as in (2) with $r=r_j$ and we set $\gamma_{j0c}=0$ for all $j=1,\ldots,j$ and $c=0,\ldots,C-1$. Huang and Bandeen-Roche (2004) proposed that RegLCMs are locally identifiable at free parameters of $(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda})=\{\beta_{dc},\gamma_{jrc},\lambda_{tjr}:j=1,\ldots,J,r=0,\ldots,M_j-1,c=0,\ldots,C-1,d=0,\ldots,p,t=1,\ldots,q\}$ if the following conditions are satisfied:

$(A1)$ $\prod_{j=1}^{J}M_j-1\geq C(\sum_{j=1}^{J}M_j-J)+C-1$;

$(A2)$ Free parameters $\gamma_{jrc},\lambda_{qjr},\beta_{pc}$ and covariate values x_{ip},z_{ijq} are all finite;

$(A3)$ The design matrix of the covariates

$$\boldsymbol{X}=\begin{pmatrix}\boldsymbol{x}_1^T\\ \vdots\\ \boldsymbol{x}_N^T\end{pmatrix}=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{N1} & \cdots & x_{Np}\end{pmatrix}$$

and

$$\boldsymbol{Z_j}=\begin{pmatrix}1 & \boldsymbol{z}_{1j}^T\\ \vdots & \vdots\\ 1 & \boldsymbol{z}_{Nj}^T\end{pmatrix}=\begin{pmatrix}1 & z_{1j1} & \cdots & z_{1jq}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & z_{Nj1} & \cdots & x_{Njq}\end{pmatrix},\quad j=1,\cdots,J$$

have full column rank;

$(A4)$ $\boldsymbol{\phi_0},\cdots,\boldsymbol{\phi_{C-1}}$ are linearly independent.

Remark 1. As in Huang and Bandeen-Roche (2004), if we consider F to be the number of pre-fixed conditional probabilities $\theta_{jrc}=0$ or 1, then Condition $(A1)$ should be extended to $\prod_{j=1}^{J}M_j-1\geq C(\sum_{j=1}^{J}M_j-J)+C-1-F$. For simplicity, we assume F = 0 throughout the paper.

Remark 2. Condition $(A1)$ implies that the number of independent response probabilities

$$\prod_{j=1}^{J}M_j-1=\text{card}\left(\left\{P(R_1=r_1,\ldots,R_J=r_J):r_j=0,\ldots,M_j-1,j=1,\ldots,J\right\}\right)$$

exceeds the number of independent parameters in $(\boldsymbol{\eta},\boldsymbol{\Theta})$,

$$\begin{aligned}&C\left(\sum_{j=1}^{J}M_j-J\right)+C-1\\ &\quad=\text{card}\left(\left\{\eta_c,\theta_{jrc}:j=1,\ldots,J,r=1,\ldots,M_j-1,c=0,\ldots,C-1\right\}\right).\end{aligned}$$

Condition $(A1)$ is necessary, without which the observed response information, may produce infinite parameter solutions and lead the model to be not identifiable. For technical rigorousness, Condition $(A2)$ as proposed in Huang and Bandeen-Roche (2004) specifies the model parameters and covariates x_{ip}, z_{ijq} are finite. In practice, the observed covariates are documented as finite values, and thus the finite condition on x_{ip} and z_{ijq} is automatically satisfied.

For RegLCMs without covariates, which are equivalent to RegLCMs with $\boldsymbol{x}_i = (1, \mathbf{0}_{1\times p})^T$ and $\boldsymbol{z}_i = \mathbf{0}_{J\times q}$, Huang and Bandeen-Roche (2004) gave a reduced form of identifiability conditions. They claimed an equivalence between the full column rank condition on the Jacobian matrix and linear independence condition on the columns of marginal probability matrix $\boldsymbol{\Psi}$ defined as

$$\boldsymbol{\Psi} = \left(\boldsymbol{\psi}_c; c = 0, \ldots, C-1\right)_{(S-1)\times C},$$

where each column $\boldsymbol{\psi}_c$ is of dimension $S-1$ in which each element corresponds to a distinct response pattern $\boldsymbol{r} = (r_1, \ldots, r_J) \in \mathcal{S}'$ and

$$\psi_{\boldsymbol{r}c} = P(\boldsymbol{R} = \boldsymbol{r} \mid L = c) = \prod_{j=1}^{J} \prod_{r=0}^{M_j-1} \theta_{jrc}^{\mathbb{I}\{r_j=r\}} = \prod_{j=1}^{J} \theta_{jr_jc}. \tag{10}$$

Here for notational convenience, we let θ_{jr_jc} to denote θ_{jrc} defined in Sect. 1.1 with $r = r_j$. Under the particular covariate latent class models with $\boldsymbol{x}_i = (1, \mathbf{0}_{1\times p})^T$ and $\boldsymbol{z}_i = \mathbf{0}_{J\times q}$, Huang and Bandeen-Roche (2004) proposed that $(\boldsymbol{\eta}, \boldsymbol{\Theta}) = \{\eta_c, \theta_{jrc} : j = 1, \ldots, J, r = 0, \ldots, M_j - 1, c = 0, \ldots, C-1\}$ are locally identifiable if Condition $(A1)$ and the following conditions are satisfied:

- $(A2^*)$ For all free parameters, $\theta_{jrc} > 0$ and $\eta_c = P(L = c) > 0$;
- $(A3^*)$ $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_{C-1}$ are linearly independent.

We see that Conditions $(A1)$–$(A3)$ and Condition $(A2^*)$, are necessary for the respective latent class models. The necessity of Conditions $(A1)$ and $(A2)$ are discussed in Remark 2. Condition $(A2^*)$ guarantees that the latent class membership probabilities and conditional response probabilities are nonzero. Condition $(A3)$ ensures $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are uniquely identifiable when $\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ are identifiable. As for Condition $(A3^*)$, it is related to the condition that the Jacobian matrix has full column rank. In the next section, we show under the assumption that Conditions $(A1)$ and $(A2^*)$ hold, Condition $(A3^*)$ is necessary for the local identifiability of the special RegLCMs without covariates, but is actually not sufficient. Similarly, for RegLCMs with covariates, under Conditions $(A1)$–$(A3)$, Condition $(A4)$ is a necessary identifiability condition but not a sufficient condition.

2. Necessity but Insufficiency of Huang and Bandeen-Roche (2004)

In this section, we show that the identifiability conditions in Huang and Bandeen-Roche (2004) are not sufficient. Following the discussion in Sect. 1.3, we first present the necessity of Condition $(A4)$ for RegLCMs and that of Condition $(A3^*)$ for RegLCMs without covariates, respectively.

Proposition 1. *For RegLCMs, Condition* $(A4)$ *is necessary for the identifiability of* $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ *under Conditions* $(A1)$–$(A3)$*. For RegLCMs without covariates, Condition* $(A3^*)$ *is necessary for the identifiability of* $(\boldsymbol{\eta}, \boldsymbol{\Theta})$ *under Conditions* $(A1)$ *and* $(A2^*)$.

Despite the necessary results, we next show that satisfying Conditions $(A1)$, $(A2^{*})$, and $(A3^{*})$ or satisfying Conditions $(A1)$–$(A4)$ is not sufficient to guarantee the local identifiability of RegLCMs without or with covariates, respectively. Our non-sufficient results are motivated by the existing works in the literature related to the identifiability of CDMs, which are a special family of RegLCMs as shown in Sect. 1.2. Specifically, we next present a proposition to show Conditions $(A1)$, $(A2^{*})$, and $(A3^{*})$ are not sufficient for CDMs without covariates and thus not sufficient for the identifiability of RegLCMs without covariates. Further, we show Conditions $(A1)$–$(A4)$ are not sufficient for RegCDMs, and thus not sufficient for the identifiability of RegLCMs in general.

Proposition 2. *Consider the setting of CDMs with polytomous responses. We assume Conditions $(A1)$–$(A3)$ hold for RegCDMs, and Conditions $(A1)$ and $(A2^{*})$ hold for RegCDMs without covariates, i.e., CDMs. If the following conditions hold:*

- $(P1)$ *Some latent attribute is required by only one item;*
- $(P2)$ *After rows permutation, the Q-matrix contains an identity matrix $\mathcal{I}_K$.*

Then, we have

- (i) *For CDMs, the matrix $\boldsymbol{\Psi}$ in Condition $(A3^{*})$ has full column rank but $(\boldsymbol{\eta}, \boldsymbol{\Theta})$ are not identifiable;*
- (ii) *For RegCDMs, the matrix $\boldsymbol{\Phi}$ in Condition $(A4)$ has full column rank but $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ are not identifiable.*

According to Proposition 2, the Q-matrix as shown in the following form satisfies Conditions $(P1)$ and $(P2)$,

$$Q = \left(\begin{array}{c|c} \multicolumn{2}{c}{\mathcal{I}_K} \\ \hline 0 & \\ \vdots & Q^{*} \\ 0 & \end{array} \right).$$

The above Q-matrix is complete as the top $K \times K$ block is an identity matrix $\mathcal{I}_K$. From the $(K+1)$th row to the Jth row, the entries in the first column are $\mathbf{0}_{J-K}$ and the entries in the remaining columns are denoted as a submatrix Q^{*}. The first result (i) in Proposition 2 is derived by extending a similar conclusion for CDMs with binary responses in Gu and Xu (2020) to CDMs with polytomous responses.

With a complete Q-matrix, the matrix $\boldsymbol{\Psi}$ in Condition $(A3^{*})$ can be shown to have full column rank, or equivalently, $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_{C-1}$ are linearly independent. And further, we can show that for RegCDMs, the matrix $\boldsymbol{\Phi}$ in Condition $(A4)$ has full column rank, that is, $\boldsymbol{\phi}_0, \ldots, \boldsymbol{\phi}_{C-1}$ are linearly independent.

With Proposition 2, we see that RegLCMs without covariates may not be identifiable when Conditions $(A1)$, $(A2^{*})$, and $(A3^{*})$ are satisfied. Specifically, consider CDMs without covariates, given Conditions $(P1)$–$(P2)$ of Proposition 2 are satisfied, Condition $(A3^{*})$ will be true since $\boldsymbol{\psi}_0, \ldots, \boldsymbol{\psi}_{C-1}$ are linearly independent. However, Proposition 2(i) shows that such CDMs are not identifiable. Therefore, Conditions $(A1)$, $(A2^{*})$, and $(A3^{*})$ are not sufficient for the identifiability of CDMs.

Similarly, RegLCMs may not be identifiable provided that Conditions $(A1)$–$(A4)$ hold. For RegCDMs, given Conditions $(A1)$–$(A3)$ and Conditions $(P1)$–$(P2)$ of Proposition 2 are met, Condition $(A4)$ will be true since $\boldsymbol{\phi}_{\mathbf{0}}, \ldots, \boldsymbol{\phi}_{\boldsymbol{C-1}}$ are linearly independent, but such RegCDMs are not identifiable according to Proposition 2(ii). Therefore, Conditions $(A1)$–$(A4)$ are not sufficient for the identifiability of RegCDMs.

3. Sufficient and Practical Identifiability Conditions

As shown in Sect. 2, Conditions $(A1)$–$(A4)$ are necessary but not sufficient for the identifiability of RegLCMs. To address the issue, this section provides sufficient conditions to determine the identifiability of RegLCMs. In addition, we also establish sufficient identifiability conditions for RegCDMs, which are of great importance in cognitive diagnosis.

For completeness, we first review the fundamental method to check the local identifiability before discussing the strict and generic identifiability. In Sect. 1.3, we have introduced the results of the local identifiability conditions proposed by Goodman (1974). The conditions can be generalized to finite many items and under the setting of RegLCMs.

We first consider RegLCMs without covariates. The definitions of conditional response probabilities θ_{jrc} follow from Sect. 1.1. For $\boldsymbol{r} = (r_1, \ldots, r_J) \in \mathcal{S}$, recall that we denote the response probability as

$$P(\boldsymbol{R} = \boldsymbol{r}) = \sum_{c=0}^{C-1} \eta_c P(\boldsymbol{R} = \boldsymbol{r} \mid L = c) = \sum_{c=0}^{C-1} \eta_c \prod_{j=1}^{J} \theta_{jr_jc}.$$

The local identifiability condition proposed by Goodman is associated with the Jacobian matrix

$$\mathbf{J} = \left(\boldsymbol{J}_{\eta_1}, \ldots, \boldsymbol{J}_{\eta_{C-1}}, \boldsymbol{J}_{\theta_{110}}, \ldots, \boldsymbol{J}_{\theta_{1(M_1-1)0}}, \ldots, \boldsymbol{J}_{\theta_{J1(C-1)}}, \ldots, \boldsymbol{J}_{\theta_{J(M_J-1)(C-1)}}\right).$$

The row dimension of $\mathbf{J}$ is $S - 1$ and the column dimension is $C(\sum_{j=1}^{J} M_j - J) + C - 1$, where each row index corresponds to one response probability $P(\boldsymbol{R} = \boldsymbol{r})$ for $\boldsymbol{r} \in \mathcal{S}'$ and each column index corresponds to one free parameter from $\{\eta_1, \ldots, \eta_{C-1}, \theta_{110}, \ldots, \theta_{1(M_1-1)0}, \cdots, \theta_{J1(C-1)}, \cdots, \theta_{J(M_J-1)(C-1)}\}$. For $c = 1, \ldots, C-1$, $\boldsymbol{J}_{\eta_c}$ is a vector of dimension $S - 1$. Each entry is a partial derivative of the response probability $P(\boldsymbol{R} = \boldsymbol{r})$ with respect to η_c at true value of η_c, which is computed to be

$$\frac{\partial P(\boldsymbol{R} = \boldsymbol{r})}{\partial \eta_c} = \prod_{j=1}^{J} \theta_{jr_jc} - \prod_{j=1}^{J} \theta_{jr_j0}.$$

And for $j = 1, \ldots, J$, $r = 1, \ldots, M_j - 1$ and $c = 0, \ldots, C-1$, $\boldsymbol{J}_{\theta_{jrc}}$ is a vector of dimension $S - 1$. Each entry is a partial derivative of the response probability $P(\boldsymbol{R} = \boldsymbol{r})$ with respect to θ_{jrc} at true value of θ_{jrc}, which is computed to be

$$\frac{\partial P(\boldsymbol{R} = \boldsymbol{r})}{\partial \theta_{jrc}} = \begin{cases} \eta_c \prod_{d \neq j} \theta_{dr_dc}, & \text{if} \quad r_j = r; \\ -\eta_c \prod_{d \neq j} \theta_{dr_dc}, & \text{if} \quad r_j = 0; \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 1. (Local Identifiability for LCMs and CDMs) *Consider RegLCMs without covariates or CDMs. Under Conditions $(A1)$ and $(A2^*)$, $(\boldsymbol{\eta}, \boldsymbol{\Theta})$ are locally identifiable if and only if the following condition holds.*

$(A3^{**})$ *The Jacobian matrix* $\mathbf{J}$ *formed above has full column rank.*

To better present the following local identifiability theorem for RegLCMs and RegCDMs, we consider a “hypothetical” subject with all covariates being zeros, that is, $\boldsymbol{x} = (1, \mathbf{0}_{1\times p})^T$ and $\boldsymbol{z} = \mathbf{0}_{J\times q}$. Denote the parameters of this particular subject to be $\boldsymbol{\eta}^0$ and $\boldsymbol{\Theta}^0$. The Jacobian matrix $\mathbf{J}^0$ formed by the derivatives of conditional response probabilities with respect to parameters $\boldsymbol{\eta}^0$ and $\boldsymbol{\Theta}^0$ is equivalent to the computation of Jacobian matrix $\mathbf{J}$ of general restricted latent class models shown in Theorem 1. Next, we present a theorem to associate the $\mathbf{J}^0$ with the local identifiability of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$.

Theorem 2. (Local Identifiability for RegLCMs and RegCDMs). *Consider RegLCMs or RegCDMs. Under Conditions* $(A1)$–$(A3)$, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ *are locally identifiable if and only if the following condition holds.*

- $(A4')$ *The Jacobian matrix* $\mathbf{J}^0$ *formed from the hypothetical subject with covariates being zeros has full column rank.*

Theorems 1 and 2 are intuitively straightforward but nontrivial to apply in practice. When the number of latent classes C and the number of item responses M_j increase, the dimension of the Jacobian matrix would increase, making it challenging to compute the rank of the Jacobian matrix.

Moreover, the conditions introduced in Theorems 1 and 2 only guarantee the local identifiability, while the global identifiability is not discussed. To ensure the sufficiency for global strict identifiability, we combine Goodman’s idea with the algebraic results from Kruskal theorem to establish our conditions. Recall that $\boldsymbol{\Phi} = (\boldsymbol{\phi}_c; c = 0, \dots, C-1)$ defined in (9) is a matrix of dimension $(S-1)\times C$. And $\boldsymbol{\phi}_c$ is a vector where each element corresponds to one response pattern and is denoted as $\phi_{\boldsymbol{r}c} = P(\boldsymbol{R} = \boldsymbol{r} \mid L = c, \boldsymbol{z} = \mathbf{0})$. To apply Kruskal theorem and to establish the strict identifiability conditions, we consider a three-way decomposition of $\boldsymbol{\Phi}$ and propose the linear independence condition regarding the decomposed matrices instead of $\boldsymbol{\Phi}$. We divide the total of J items of $\boldsymbol{\Phi}$ into three mutually exclusive item sets $\mathcal{J}_1$, $\mathcal{J}_2$, and $\mathcal{J}_3$ containing J_1, J_2, and J_3 items, respectively, with $J_1 + J_2 + J_3 = J$. For $t = 1, 2,$ and 3, each set $\mathcal{J}_t$ can be viewed as one polytomous variable T_t taking on values in $\{1, \dots, \kappa_t\}$ with cardinality $\kappa_t = \prod_{j\in\mathcal{J}_t} M_j$ to be the number of response patterns for this set. And each variable T_t is used to construct a $\kappa_t \times C$ submatrix $\boldsymbol{\Phi}_t$, where its row indices arise from the response patterns corresponding to T_t. The linear independence condition is then regarding to the Kruskal ranks of $\boldsymbol{\Phi}_t$ rather than normal column rank of $\boldsymbol{\Phi}$, where for any matrix $\boldsymbol{\Phi}_t$, its Kruskal rank I_t is the smallest number of columns of $\boldsymbol{\Phi}_t$ that are linearly dependent.

Theorem 3. (Strict Identifiability for RegLCMs). *Continue with the notation definitions in Sect. 1.3. For RegLCMs, under Conditions* $(A1)$–$(A3)$ *and the following condition,* $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ *are strictly identifiable.*

- $(C4)$ *The matrix* $\boldsymbol{\Phi}$ *can be decomposed into* $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$, *and* $\boldsymbol{\Phi}_3$ *with Kruskal ranks* I_1, I_2, *and* I_3 *satisfying* $I_1 + I_2 + I_3 \geq 2C + 2$.

Theorem 3 is sufficient to guarantee the strict identifiability for RegLCMs, including RegCDMs. Compared with the local identifiability conditions in Huang and Bandeen-Roche (2004), Theorem 3 keeps Conditions $(A1)$–$(A3)$ and replaces Condition $(A4)$ concerning the column rank of $\boldsymbol{\Phi}$ with a stronger Condition $(C4)$ concerning the Kruskal ranks of the decomposed matrices from $\boldsymbol{\Phi}$. This condition is based on the algebraic result in Kruskal (1977). We next present identifiability conditions tailored to RegCDMs.

Proposition 3. (Strict Identifiability for RegCDMs). *For RegCDMs with polytomous responses, under Conditions* $(A1)$–$(A3)$ *and the following condition,* $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ *are strictly identifiable.*

($C4^*$) *After rows permutation, Q-matrix takes the form $Q = (\mathcal{I}_K, \mathcal{I}_K, Q^*)^T$ containing two identity matrices $\mathcal{I}_K$ and one submatrix $Q^*_{(J-2K)\times K}$. And for any different latent classes c and c', there exist at least one item $j > 2K$ such that $(\theta_{j0c}, \ldots, \theta_{j(M_j-1)c})^T \neq (\theta_{j0c'}, \ldots, \theta_{j(M_j-1)c'})^T$.*

It has been established that Condition ($C4^*$) itself is a sufficient condition for the identifiability of general restricted latent class models with binary responses (Xu, 2017). In addition, Xu and Shang (2018) showed that the Q-matrix is also identifiable under Condition ($C4^*$). This condition is further extended to the restricted latent class models with polytomous responses in Culpepper (2019). Compared to the previous literature, the major contribution of Proposition 3 is to extend this constraint to the polytomous-response RegCDMs that the Q-matrix contains two identity matrices and the conditional response probability $(\theta_{j0c}, \ldots, \theta_{j(M_j-1)c})^T$ is distinct among different latent classes.

In practice, the theoretical results in Theorem 3 and Proposition 3 may need further adjustments to accommodate the empirical needs. As previously discussed, generic identifiability is commonly used in practice as it guarantees the identifiability of most parameters other than a measure-zero set of parameters (Allman et al., 2009). The following theorem and proposition will provide us with an easy way to determine the generic identifiability of RegLCMs and RegCDMs.

Theorem 4. (Generic Identifiability for RegLCMs). *For RegLCMs, under Conditions ($A1$)–($A3$) and the following condition, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ are generically identifiable.*

($C4'$) *The matrix $\boldsymbol{\Phi}$ can be decomposed into $\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$, and $\boldsymbol{\Phi}_3$ with row dimensions κ_1, κ_2, and κ_3 satisfying $\min\{C, \kappa_1\} + \min\{C, \kappa_2\} + \min\{C, \kappa_3\} \geq 2C + 2$.*

Remark 3. Under the special case that the number of possible responses to each item are identical, $M_1 = \cdots = M_J$, we have a reduced form of Condition ($C4'$) in Theorem 4. This finding is based on Corollary 5 and its related discussions from Allman et al. (2009). They show that for these special cases, the decomposition can be carefully chosen to maximize $\min\{C, \kappa_1\} + \min\{C, \kappa_2\} + \min\{C, \kappa_3\}$, which results in a simpler form of identifiability condition.

Consider RegLCMs with binary responses $M_j = 2$ for $j = 1, \ldots, J$, we have $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ to be generically identifiable if we replace Condition ($C4'$) with the condition $J \geq 2\left\lceil \log_2 C \right\rceil + 1$. More generally, for the RegLCMs with $M_j = M$ for $j = 1, \ldots, J$, we have $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ to be generically identifiable if Condition ($C4'$) is replaced with the condition $J \geq 2\left\lceil \log_M C \right\rceil + 1$. For these special models, the reduced conditions provide researchers with simpler ways to determine the generic identifiability compared with Condition ($C4'$) as they only concern the number of items J and the number of latent classes C.

Compared with the strict identifiability conditions in Theorem 3, Theorem 4 makes it more practical to check the identifiability of RegLCMs as the variables in Condition ($C4'$) are row dimensions rather than the Kruskal ranks of the decomposed matrices. But Theorem 4 does not apply to all latent class models. For instance, the parameter space of restricted latent class models may lie in the nonidentifiable measure-zero set from the parameter space of general latent class models. Therefore, Theorem 4 does not apply to restricted latent class models with covariates such as RegCDMs. To address this issue, Proposition 4 is established to determine the generic identifiability for RegCDMs with polytomous responses.

Proposition 4. (Generic Identifiability for RegCDMs). *For RegCDMs with polytomous responses, under Conditions ($A1$)–($A3$) and the following condition, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ are generically identifiable.*

($C4''$) *After rows permutation, Q-matrix takes the form $Q = (Q_1, Q_2, Q^*)^T$ containing one submatrix $Q^*_{(J-2K)\times K}$ in which each attribute is required by at least one item, and two submatrices Q_1 and Q_2 in the following form,*

$$Q_i = \begin{pmatrix} 1 & * & \cdots & * \\ * & 1 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & 1 \end{pmatrix}, \quad i = 1, 2 \tag{11}$$

where "$*$" indicates the entry is either 1 or 0.

Condition ($C4''$) was first proposed by Gu and Xu (2020) to determine the generic identifiability of CDMs. For RegCDMs, Proposition 4 gives more flexible conditions than Proposition 3 as Condition ($C4''$) puts less constraints on the Q-matrix than Condition ($C4^*$) does. Condition ($C4^*$) requires the Q-matrix to contain two identity submatrices, whereas in the Q-matrix form required by ($C4''$), the two identity matrices are replaced by two matrices as shown in (11), which allows more flexibility on the off-diagonal entries. Under this new condition, the parameters may not be strictly identifiable but are identifiable in the generic sense.

Proposition 4 provides sufficient conditions to guarantee the generic identifiability of RegCDMs. Under certain special cases, we can show that those conditions are also necessary. Next, we introduce a particular example where the conditions in Proposition 4 are not only sufficient, but also necessary for the generic identifiability of the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$.

Example 1. Consider a special RegCDM with binary responses and two latent attributes, i.e., $K = 2$ and $M_j = 2$. Under Conditions ($A1$)–($A3$), Condition ($C4''$) in Proposition 4 is necessary and sufficient for the generic identifiability of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$. For instance, after rows permutation, the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & * \\ * & 1 \\ 1 & * \\ * & 1 \\ \hline \multicolumn{2}{c}{Q'} \end{pmatrix}, \tag{12}$$

where "$*$" is either zero or one and Q' is a matrix with at least one entry to be 1 in each column. Proposition 3 in Gu and Xu (2021) shows that Condition ($C4''$) is necessary and sufficient condition for generic identifiability for Q-matrix, $\boldsymbol{\Theta}$ and $\boldsymbol{\eta}$. Hence for RegCDMs, we have ($\boldsymbol{\eta}^i$, $\boldsymbol{\Theta}^i$) identifiable. As for the identifiability of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ in RegCDMs, under Condition ($A3$) that $\boldsymbol{X}$ and $\boldsymbol{Z_j}$'s have full column rank, ($\boldsymbol{\eta}^i$, $\boldsymbol{\Theta}^i$) are identifiable if and only if $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ are identifiable, which can be seen from *Steps 2–3* of the *Proof of Theorem* 2 in Supplementary Material. Therefore, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$ are identifiable for the considered RegCDMs with two attributes if and only if Condition ($C4''$) in Proposition 4 holds.

4. Data Example

In this section, we use a real dataset to demonstrate an application of the proposed identifiability conditions in educational assessments. Trends in Mathematics and Science Study (TIMSS) is an international and large-scale assessment to evaluate the mathematics skills and science knowledge of students in different grades. We consider a TIMSS 2007 4th Grade dataset, which was studied in Park and Lee (2014) and is accessible from the R package "CDM" (George et al., 2016; Robitzsch et al., 2020). The dataset contains $N = 698$ Austrian 4th grade students' binary responses ($M_j = 2$) to $J = 25$ items together with their gender information. The gender is denoted as a binary variable with $g_i = 1$ for female students and $g_i = 0$ for male students.

We model the TIMSS 2007 dataset using RegCDMs and study their identifiability. We consider gender g_i as covariates with $\boldsymbol{x}_i = (1, g_i)^T$ and $z_{ij} = (g_i)$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, under the assumption that both $\boldsymbol{\eta}$ and $\boldsymbol{\Theta}$ can be associated with the gender. Following Park and Lee (2014), the test assesses $K = 7$ latent attributes in the domains of (α_1) Whole numbers; (α_2) Fractions and Decimals; (α_3) Number Sentences, Patterns, & Relationships; (α_4) Lines and Angles; (α_5) Two- and Three-Dimensional Shapes; (α_6) Location and Movement; (α_7) Reading, Interpreting, Organizing, & Representing. As shown in Park and Lee (2014), the seven latent attributes can be further aggregated into $K' = 3$ general domains: (α'_1) Number; (α'_2) Geometric Shapes and Measures; (α'_3) Data Display.

We first show that the RegCDM with $K = 7$ attributes is generically identifiable by Proposition 4. As there are $C = 2^7 = 128$ latent classes, Condition $(A1)$ holds as $\prod_{j=1}^{J} M_j - 1 - C(\sum_{j=1}^{J} M_j - J) - C + 1 = 2^{25} - 2^7 \times 25 - 2^7 > 0$. Condition $(A2)$ holds as the binary covariates are finite and coefficient parameters are free since we have no constraint on coefficients. Condition $(A3)$ holds as the design matrices

$$\boldsymbol{X} = \boldsymbol{Z_j} = \begin{pmatrix} 1 & g_1 \\ 1 & g_2 \\ \vdots & \vdots \\ 1 & g_N \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}, \quad \text{for } j = 1, \ldots, J,$$

have full column rank given the sample has both female and male students. Lastly for Condition $(C4'')$, the Q-matrix after rows permutation from Park and Lee (2014) is presented in Table 1. The Q-matrix implies that Condition $(C4'')$ holds as the matrices Q_1 and Q_2 have diagonal entries to be ones and each column of the submatrix Q^* contains the value one for at least once. According to Proposition 4, the RegCDM is generically identifiable. However, the Q-matrix is not complete, so the RegCDM is not strictly identifiable.

We next show that the RegCDM with $K' = 3$ attributes is generically identifiable as well by Proposition 4. As there are $C = 2^3 = 8$ latent classes, Condition $(A1)$ holds because $\prod_{j=1}^{J} M_j - 1 - C(\sum_{j=1}^{J} M_j - J) - C + 1 = 2^{25} - 2^3 \times 25 - 2^3 > 0$. As the items, the students' responses, and the covariates are unchanged, we have Conditions $(A2)$–$(A3)$ hold by the same arguments as in the RegCDM with $K = 7$. In assessing the three general attributes, the Q-matrix used in Park and Lee (2014) is given in Table 2 after rows permutation. This Q-matrix contains Q_1 and Q_2 with diagonal entries to be ones and the submatrix Q^* with each attribute column containing the value one for at least one entry. Therefore, Condition $(C4'')$ holds and Proposition 4 shows that the RegCDM with $K' = 3$ is generically identifiable. However, the Q-matrix does not contain an identity matrix as the α'_3 is not singularly required by any item. So the RegCDM with $K' = 3$ is not strictly identifiable.

TABLE 1.
The Q-matrix for TIMSS 2007 data at $K = 7$

	Item No.	α_1	α_2	α_3	α_4	α_5	α_6	α_7
Q_1	1	1	0	0	0	0	0	0
	3	1	1	0	0	0	0	0
	5	1	0	1	0	0	0	0
	10	0	0	0	1	1	0	0
	9	0	0	0	0	1	0	0
	6	0	0	0	0	1	1	0
	12	1	0	0	0	0	0	1
Q_2	15	1	0	0	0	0	0	0
	4	1	1	0	0	0	0	0
	17	1	0	1	0	0	0	0
	11	1	0	0	1	0	0	0
	24	0	0	0	0	1	0	0
	22	0	0	0	0	1	1	0
	13	1	0	0	0	0	0	1
Q^*	2	0	1	0	0	0	0	0
	8	1	0	0	0	1	0	0
	7	0	0	0	1	1	1	0
	14	1	1	0	0	0	0	1
	16, 23	1	0	0	0	0	0	0
	18, 20	1	0	1	0	0	0	0
	19, 25	1	0	0	0	0	0	1
	21	1	0	1	0	0	0	0

TABLE 2.
The Q-matrix for TIMSS 2007 data at $K = 3$

	Item No.	α'_1	α'_2	α'_3
Q_1	1	1	0	0
	6	0	1	0
	12	1	0	1
Q_2	2	1	0	0
	7	0	1	0
	13	1	0	1
Q^*	3–5, 15–18, 21, 23	1	0	0
	9, 10, 22, 24	0	1	0
	14, 19, 20, 25	1	0	1
	8, 11	1	1	0

5. Discussion

This paper studies latent class models with covariates, in particular RegLCMs. Under the setup of RegLCMs and its special family RegCDMs, we focus on the identifiability conditions for the coefficient parameters of the covariates. We show that Huang and Bandeen-Roche (2004) presented necessary but not sufficient conditions for the local identifiability of RegLCMs. Then, we establish conditions for the local and global identifiability of RegLCMs and RegCDMs.

The classical and fundamental method for local identifiability is based on Goodman's results, which is to ensure the full column rank of the Jacobian matrix formed by the derivatives of general response probabilities with respect to parameters. We propose sufficient and practical conditions based on Huang and Bandeen-Roche (2004) to replace the previous linear independence condition on the marginal probability matrix with the linear independence condition concerning three decomposed probability matrices. Noticing the empirical convenience of the generic identifiability, we present specific conditions to ensure the generic identifiability as well. The conditions for generic identifiability involve more accessible variables from decomposed submatrices. In addition to the global identifiability of general RegLCMs, the conditions for the global identifiability of RegCDMs are dependent on the Q-matrix, and these conditions are extended from the binary-response CDMs to the polytomous-response CDMs.

Regarding the consistency of estimation, Gu and Xu (2020) proved that for general restricted latent class models, the latent class membership probability and conditional response probability can be consistently estimated with maximum likelihood estimators. The estimation consistency is retained for the parameters in RegLCMs because the parameters are linearly related to the log-odds and the design matrices of covariates have full column ranks. The proposed conditions are sufficient and practical, but may not be necessary in strict identifiability cases. For generic identifiability, we discuss the sufficient and necessary conditions for the binary-response CDMs with binary attributes in Example 1, except which the necessary side of identifiability conditions is still under research. For future works, we plan to investigate the sufficient and necessary conditions for the identifiability of latent class models with covariates.

Acknowledgments

The authors are grateful to the Editor-in-Chief Professor Matthias von Davier, an Associate Editor, and a referee for their valuable comments and suggestions. This research is partially supported by NSF CAREER SES-1846747 and Institute of Education Sciences R305D200015.

References

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics, 37*(6A), 3099–3132.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology, 43*(1), 272–311.

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

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika, 85*(1), 121–153.

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

Clogg, C. C., & Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association, 79*(388), 762–771.

Collins, L. M., & Lanza, S. T. (2009). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences (Vol. 718). Wiley.

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

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83*(401), 173–178.

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

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

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38*(1), 87–111.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24.

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

Gu, Y., & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika, 84*(2), 468–483.

Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *Annals of Statistics, 48*(4), 2082–2107.

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica, 31*, 449–472.

Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability, 31*(2), 542–548.

Huang, G.-H. (2005). Selecting the number of classes under latent class regression: A factor analytic analogue. *Psychometrika, 70*(2), 325–345.

Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69*(1), 5–32.

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

Koopmans, T. C. (1950). *Statistical inference in dynamic economic models*. Wiley.

Koopmans, T. C., & Reiersol, O. (1950). The identification of structural characteristics. *Annals of Mathematical Statistics, 21*(2), 165–181.

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.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21*(4), 331–347.

Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics, 30*(1), 27–58.

Muthén, L., & Muthén, B. (2017). Mplus user's guide: Statistical analysis with latent variables, user's guide. Muthén & Muthén.

Pan, J.-C., & Huang, G.-H. (2014). Bayesian inferences of latent class models with an unknown number of classes. *Psychometrika, 79*(4), 621–646.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38*(5), 376–390.

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied Psychological Measurement, 42*(5), 376–392.

Petersen, J., Bandeen-Roche, K., Budtz-Jørgensen, E., & Groes Larsen, K. (2012). Predicting latent class scores for subsequent analysis. *Psychometrika, 77*(2), 244–262.

Reboussin, B. A., Ip, E. H., & Wolfson, M. (2008). Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 171*(4), 877–897.

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2020). CDM: Cognitive diagnosis modeling. Retrieved from https://CRAN.R-project.org/package=CDM (R package version 7.5-15)

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

van der Heijden, P. G. M., Dessens, J., & Bockenholt, U. (1996). Estimating the concomitant-variable latent-class model with the EM algorithm. *Journal of Educational and Behavioral Statistics, 21*(3), 215–229.

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

Wu, Z., Deloria-Knoll, M., & 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. *Annals of Statistics, 45*(2), 675–707.

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

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

Manuscript Received: 19 MAR 2021
Final Version Received: 16 OCT 2021
Published Online Date: 7 MAR 2022