# Supplementary Material for "New Paradigm of Identifiable General-response Cognitive Diagnostic Models: Beyond Categorical Data"

Seunghyun Lee and Yuqi Gu

Department of Statistics, Columbia University

This Supplementary Material is organized as follows. Section S.1 provides a measure-theoretic definition of general-response CDMs. Section S.2 introduces additional examples of parametric CDMs, such as the exponential family general diagnostic models (ExpGDMs) for general responses and negative-binomial-based CDMs. Section S.3 presents proofs to all theoretical results in the main manuscript. Section S.4 illustrates how our EM algorithms for estimating ExpCDMs can be modified to estimate negative-binomial-based CDMs. Section S.5 presents additional simulation details, and a simulation study showing empirical consistency under our generic identifiability conditions. Finally, we present our real data pre-processing details and an analysis of the response accuracy data in Section S.6.

## S.1 Rigorous definition of the general-response CDMs

In this section, we define the sample space for the responses, $\mathcal{Y}_j$, in a more general and rigorous manner. Here, $\mathcal{Y}_j$'s only need to be separable metric spaces for our main Theorems to hold. We adopt the following measure-theoretic terms to introduce our modeling framework and identifiability results in full generality; these terms follow from classical probability theory textbooks such as Durrett (2019).

Let $\mathcal{F}_j$ and $m_j$ be the Borel sigma-algebra and a base-measure on $\mathcal{Y}_j$, respectively. Thus, for each $\mathcal{Y}_j$, we have defined the probability triplet $(\mathcal{Y}_j, \mathcal{F}_j, m_j)$. In many concrete model examples, $\mathcal{F}_j$ and $m_j$ are naturally defined upon specifying the response type of $Y_j \in \mathcal{Y}_j$. For instance, as illustrated in the main text, when $\mathcal{Y}_j = \mathbb{R}$ with $Y_j$ being a continuous response,

$\mathcal{F}_j$ is the collection of Borel sets in $\mathbb{R}$ and $m_j$ is the Lebesgue measure on $\mathbb{R}$.

Under this definition of $\mathcal{F}_j$, equation (6) in the main paper (for defining parametric CDMs) can be generalized as:

$$\mathbb{P}_{j,\boldsymbol{\alpha}}(Y_j \in S_j \mid \mathbf{A} = \boldsymbol{\alpha}) = \int_{S_j} g(y; \boldsymbol{\eta}_{j,\boldsymbol{\alpha}}) \mathrm{d}m_j(y). \tag{S.1}$$

Consequently, the parametric CDMs and ExpCDMs can be defined for such $\mathcal{Y}_j$'s.

Our above general assumption on the response types covers binary, polytomous, continuous, count responses, etc. As mentioned in the main paper, we introduce the above abstract measure-theoretic definitions to illustrate that the model (Definition 1) and identifiability results (Theorem 1 and 2) can be stated in the most general setting.

## S.2 More examples of parametric CDMs

### S.2.1 The ExpGDM model for general responses

In this section, we define the exponential family-based general diagnostic models for general responses, which we call ExpGDMs. ExpGDMs can be viewed as general-response extensions of the categorical-response GDM proposed by von Davier (2008). In categorical-response GDM and similar models such as the generalized DINA model (GDINA, de la Torre, 2011) and the loglinear CDM (LCDM, Henson et al., 2009), the response probabilities depend not only on the main effects but also on all the interaction effects of the required attributes of an item. In an ExpGDM, we define the parameter $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}}$ in (6) as

$$\begin{aligned}\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} &= \boldsymbol{h}\Big(\beta_{j,\varnothing} + \sum\nolimits_{k=1}^{K} \beta_{j,k} \{q_{j,k}\alpha_k\} + \sum\nolimits_{1\le k_1<k_2\le K} \beta_{j,k_1k_2} \{q_{j,k_1}\alpha_{k_1}\} \{q_{j,k_2}\alpha_{k_2}\} \\ &\qquad + \cdots + \beta_{j,12\cdots K} \prod\nolimits_{k=1}^{K} \{q_{j,k}\alpha_k\}\,,\ \gamma_j\Big) \\ &= \boldsymbol{h}\Big(\sum_{S\subseteq \mathrm{pa}(j)} \beta_{j,S} \prod_{k\in S} \alpha_k,\ \gamma_j\Big).\end{aligned} \tag{S.2}$$

Here, similar to ExpACDMs, not all the $\beta$-coefficients are needed to specify the model; instead, only those corresponding to the main and interaction effects of the required attributes in $\mathrm{pa}(j)$ enter the definition of $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}}$. Therefore, the $\mathbf{Q}$-matrix constraints in (7) are satisfied.

The mapping $\boldsymbol{h}$ is a link function that maps the main- and interaction-effects of attributes to the natural parameters $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}}$, and $\gamma_j$ is the additional parameter that does not depend on $\boldsymbol{\alpha}$. With these notations at hand, we define the ExpGDM as follows.

**Definition 1** (ExpGDM)**.** *The ExpGDM is an ExpCDM for which the natural parameters* $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}}$ *satisfy* (S.2)*. The parameters in an ExpGDM model with link* $\boldsymbol{h}$ *include* $\{\beta_{j,S} : j \in [J], S \subseteq \mathrm{pa}(j)\}, \{\gamma_j : j \in [J]\}$ *and proportion parameters* $\boldsymbol{p}$*.*

Next, we define the lognormal-GDM and Poisson-GDM and provide examples of the ExpGDM model.

**Example 1** (lognormal-GDM)**.** *Define the lognormal-GDM by letting* $g$ *be the lognormal density with*

$$\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} = \Big(\frac{\mu_{j,\boldsymbol{\alpha}}}{\sigma^2_{j,\boldsymbol{\alpha}}}, \ -\frac{1}{2\sigma^2_{j,\boldsymbol{\alpha}}}\Big), \quad \textit{where } \mu_{j,\boldsymbol{\alpha}} = \sum_{S\subseteq \mathrm{pa}(j)} \beta_{j,S} \prod_{k\in S} \alpha_k, \quad \sigma^2_{j,\boldsymbol{\alpha}} = \gamma_j$$

*with the same link* $\boldsymbol{h}$ *as in the lognormal-ACDM. This lognormal-GDM is a sub-model of the C-G-DINA model in Minchen and de la Torre (2018). The C-G-DINA model additionally considers all effects for the variance parameter as well, i.e.*

$$\mu_{j,\boldsymbol{\alpha}} = \sum_{S\subseteq \mathrm{pa}(j)} \beta_{j,S} \prod_{k\in S} \alpha_k, \qquad \sigma^2_{j,\boldsymbol{\alpha}} = \sum_{S\subseteq \mathrm{pa}(j)} \beta'_{j,S} \prod_{k\in S} \alpha_k.$$

*This model can also be viewed as an ExpGDM by considering the interaction-effect coefficients for* $S \subseteq \mathrm{pa}(j)$ *as a two-dimensional vector* $(\beta_{j,S}, \beta'_{j,S})^\top$*.*

**Example 2** (Poisson-GDM)**.** *Define the Poisson-GDM by letting* $g$ *be the Poisson density with* $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} = \log \lambda_{j,\boldsymbol{\alpha}}$*,* $\boldsymbol{h}(\lambda_{j,\boldsymbol{\alpha}}) = \log \lambda_{j,\boldsymbol{\alpha}}$*, and writing*

$$\lambda_{j,\boldsymbol{\alpha}} = \sum_{S\subseteq \mathrm{pa}(j)} \beta_{j,S} \prod_{k\in S} \alpha_k.$$

*Here,* $\beta_{j,S} \geq 0$*. This Poisson-GDM model can be viewed as a reparametrization of the Poisson Diagnostic Classification Model (PDCM; Liu et al., 2022).*

### S.2.2 Parametric CDMs with non-exponential family distributions

In this section, we illustrate how the general-response DINA model and ACDM can be defined when the parametric family $\mathcal{P}$ is not an exponential family. In particular, we focus on the negative binomial distribution and define negative binomial-based DINA model (NegBin-DINA) and negative binomial-based ACDM (NegBin-ACDM). In addition to the Poisson, the negative binomial is another popular distribution for modeling count data. The negative binomial distribution with parameters $(r,\ \pi)$ has the following probability mass function:

$$\mathbb{P}(X = x) = \binom{x + r - 1}{x}(1 - \pi)^x \pi^r, \quad x = 0, 1, 2, \ldots$$

where $r > 0$ and $\pi \in (0, 1)$.

**Example 3** (NegBin-DINA)**.** For each item $j \in [J]$, we define different values of $(r,\ \pi)$ for the two classes of $\boldsymbol{\alpha}$ depending on the value of $\Gamma_{j,\boldsymbol{\alpha}} = \prod_{k=1}^{K} \alpha_k^{q_{j,k}}$. By letting $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} = (r_{j,\Gamma_{j,\boldsymbol{\alpha}}},\ \pi_{j,\Gamma_{j,\boldsymbol{\alpha}}})$, the distribution $Y_j \mid \mathbf{A}$ can be written as

$$Y_j \mid \mathbf{A} = \boldsymbol{\alpha} \sim \begin{cases} \text{NegBin}(r_{j,1},\ \pi_{j,1}), & \text{if } \Gamma_{j,\boldsymbol{\alpha}} = 1, \\ \text{NegBin}(r_{j,0},\ \pi_{j,0}), & \text{if } \Gamma_{j,\boldsymbol{\alpha}} = 0. \end{cases} \tag{S.3}$$

This can be viewed as a reparametrized version of the CDCM-DINA model proposed in Liu et al. (2023).

**Example 4** (NegBin-ACDM)**.** For count data, we believe that it is unreasonable to model the variance as a constant and the variance may have a linear relationship with the mean. Hence, we define $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} = (r_{j,\boldsymbol{\alpha}}, \pi_{j,\boldsymbol{\alpha}})$ by the link function

$$\boldsymbol{\eta}_{j,\boldsymbol{\alpha}} = \boldsymbol{h}\Big(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{j,k} \alpha_k, \gamma_j\Big) = \left(\frac{\gamma_j}{1 - \gamma_j}(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{j,k} \alpha_k),\ \gamma_j\right). \tag{S.4}$$

With this parametrization in (6), $\mathbb{P}_{j,\boldsymbol{\alpha}}$ can be written as

$$Y_j \mid \mathbf{A} = \boldsymbol{\alpha} \sim \text{NegBin}\left(\frac{\gamma_j}{1 - \gamma_j}(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{j,k} \alpha_k),\ \gamma_j\right),$$

and in particular, we get the desired linear mean and variance with repsect to the required

attributes:

$$\mathbb{E}(Y_j \mid \boldsymbol{\alpha}) = \frac{r_{j,\boldsymbol{\alpha}}(1-\pi_{j,\boldsymbol{\alpha}})}{\pi_{j,\boldsymbol{\alpha}}} = \beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{j,k} \alpha_k,$$
$$\mathrm{Var}(Y_j \mid \boldsymbol{\alpha}) = \frac{r_{j,\boldsymbol{\alpha}}(1-\pi_{j,\boldsymbol{\alpha}})}{\pi_{j,\boldsymbol{\alpha}}^2} = \frac{\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{j,k} \alpha_k}{\gamma_j}.$$

We remark that for the negative binomial link, some studies have adopted different ways to model the variance (e.g. Man and Harring, 2019; Liu et al., 2023). These models can all be covered in our general-response ACDM framework by defining a different link function $\boldsymbol{h}$.

## S.3 Proof of Theorems 1, 2 and Propositions 1, 2, 3, 4

### S.3.1 Proof of Theorem 1 and Proposition 1

Our proof is motivated by the proof of Theorem 8 in Allman et al. (2009), which proved the identifiability of nonparametric mixture models with independent marginals by discretizing the sample space $\mathbb{R}$ into bins. After the discretization, the marginal probability distribution of the observed variables can be written as a tensor, and identifiability is established by proving that there exists a unique decomposition of that tensor. Here, we adopt a similar discretization trick and exploit the uniqueness of the resulting tensor decomposition. However, technical difficulties arise as (1) we work with general responses rather than the continuous responses in Allman et al. (2009) and (2) our conditional distributions $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s are subject to the equality constraints induced by the $\mathbf{Q}$-matrix while Allman et al. (2009) works with linearly independent conditional distributions. Our proof mainly consists of the following four steps:

Step 1. We reduce estimating the nonparametric distribution $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s into estimating a finite number of values $\{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_j) : S_j \in \mathcal{D}_j\}$'s. Here, $\mathcal{D}_j$ is a finite set that is defined later in the proof.

Step 2. We write the marginal probability $\mathbb{P}(\mathbf{Y} \in \times_{j=1}^{J} S_j)$ as a three-way tensor decomposition. This decomposition naturally arises from (4) by combining (unfolding) the row-indices of $\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}^\star$.

Step 3. We apply Kruskal's theorem and show that the three-way tensor can be uniquely decomposed.

Step 4. We show that the model components $\boldsymbol{p}$ and $\{\mathbb{P}_{j,\boldsymbol{\alpha}}\}$ can be recovered from the decomposed tensor.

In order to provide rigorous proof that applies to general responses, we adopt measure-theoretic terminology and notations.

*Proof.* Suppose that the $\mathbf{Q}$-matrix and the true model components $(\boldsymbol{p}, \{\mathbb{P}_{j,\boldsymbol{\alpha}}\})$ satisfy conditions A and B. We next show that there exists no other choice of model components that satisfy (4) in four steps.

**Step 1: Reducing the problem to identifying a finite number of probabilities** We claim that for all $j$, there exist countable separating classes[1] $\mathcal{C}_j$'s that determine the probability measure $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s. First, because $\mathcal{Y}_j$ is a separable metric space, there exists a countable basis $\mathcal{B}_j$ of $\mathcal{Y}_j$. For example, one can simply consider open balls centered at a countable dense set. Let $\mathcal{C}_j$ be the finite intersections of $\mathcal{B}_j$. Then, $\mathcal{C}_j$ is a countable $\pi$-system such that $\sigma(\mathcal{C}_j)$ contains all open sets in $\mathcal{Y}_j$. Because $\mathcal{F}_j$ is the Borel $\sigma$-algebra (i.e. the $\sigma$-algebra generated by all open sets), $\mathcal{F}_j \subseteq \sigma(\mathcal{C}_j)$. Hence, $\mathcal{C}_j$ is a separating class by Theorem 3.3 in (Billingsley, 2013, page 42).

Since $\mathbb{P}_{j,\boldsymbol{\alpha}}$ is determined by its values evaluated at $S_j \in \mathcal{C}_j$, it suffices to show that $(\boldsymbol{p},\ \{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_j) : S_j \in \mathcal{D}_j\})$ are identifiable for all finite subsets $\mathcal{D}_j$'s of $\mathcal{C}_j$ that fully distinguish different $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s[2]. To this extent, we consider an increasing collection of sets $\mathcal{D}_j^{(t)}, t \geq 1$ (i.e. $\mathcal{D}_j^{(t)} \subsetneq \mathcal{D}_j^{(t+1)}$, $\cup_{t\leq\infty}\mathcal{D}_j^{(t)} = \mathcal{C}_j$) and prove that $(\boldsymbol{p}, \{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_j) : S_j \in \mathcal{D}_j^{(t)}\})$ is uniquely determined, and well-defined across $t$. For notational simplicity, we drop the superscript $(t)$ by writing

$$\mathcal{D}_j = \mathcal{D}_j^{(t)} = \{S_{1,j}, ..., S_{\kappa_j,j}\}$$

when the meaning is clear. Here, $\kappa_j = |\mathcal{D}_j|$. Without the loss of generality, we assume that $S_{\kappa_j,j} = \mathcal{Y}_j$, i.e. the last element in $\mathcal{D}_j$ is the entire range.

[1] $\mathcal{C}_j \subset \mathcal{F}_j$ is a separating class of $\mathcal{F}_j$ if for any two measures $\mathbb{P}_1, \mathbb{P}_2$ on $(\mathcal{Y}_j, \mathcal{F}_j)$ with $\mathbb{P}_1(E) = \mathbb{P}_2(E)$ for all $E \in \mathcal{C}_j$, we have $\mathbb{P}_1 = \mathbb{P}_2$.

[2] $\mathcal{D}_j$ fully distinguishes different $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s when, for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ such that $\mathbb{P}_{j,\boldsymbol{\alpha}} \neq \mathbb{P}_{j,\boldsymbol{\alpha}'}$, there exists $S_j \in \mathcal{D}_j$ with $\mathbb{P}_{j,\boldsymbol{\alpha}}(S_j) \neq \mathbb{P}_{j,\boldsymbol{\alpha}'}(S_j)$

**Step 2: Writing the marginal probability distribution as a tensor decomposition.** We start by introducing new notations that are necessary to write the conditional probabilities evaluated at $S_{l,j} \in \mathcal{C}_j$. For $j = 1, \ldots, 2K$, define $\Lambda^j$ to be a $\kappa_j \times 2$ matrix where $\Lambda^j(l, \alpha_j) = \mathbb{P}_{j,\alpha_j}(S_{l,j})$. For $j > 2K$, let $\Psi^j$ be a $\kappa_j \times 2^K$ matrix where $\Psi^j(l, \boldsymbol{\alpha}) = \mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l,j})$. Also, define $\mathbf{N}_1$ to be a $\kappa_1 ... \kappa_K \times 2^K$ matrix whose $((l_1, ..., l_K), \boldsymbol{\alpha})$-th entry is $\mathbb{P}(Y_1 \in S_{l_1,1}, ..., Y_K \in S_{l_K,K} \mid \boldsymbol{\alpha})$. Here, we index the $2^K$ columns of $\Psi^j$ and $\mathbf{N}_1$ using the binary vector $\boldsymbol{\alpha} \in \{0,1\}^K$. For $\Lambda^j$, the columns are indexed by $\alpha_j \in \{0,1\}$. Similarly, we index the rows of $\mathbf{N}_1$ by $(l_1, \ldots, l_K)$, where $l_j \in [\kappa_j]$.

Using the assumptions (2), (3), and condition A, we can write

$$\mathbb{P}(Y_1 \in S_{l_1,1}, ..., Y_K \in S_{l_K,K} \mid \boldsymbol{\alpha}) = \prod_{j=1}^{K} \mathbb{P}(Y_j \in S_{l_j,j} \mid \boldsymbol{\alpha}_{\mathrm{pa}(j)}) = \prod_{j=1}^{K} \mathbb{P}(Y_j \in S_{l_j,j} \mid \alpha_j).$$

Hence, $\mathbf{N}_1$ can be decomposed as $\bigotimes_{j=1}^{K} \Lambda^j$, where $\bigotimes$ denotes the Kronecker product of matrices. Similarly, define $\mathbf{N}_2$ be a $\kappa_{K+1} \ldots \kappa_{2K} \times 2^K$ matrix whose $((l_{K+1}, \ldots, l_{2K}), \boldsymbol{\alpha})$-th entry is $\mathbb{P}(Y_{K+1} \in S_{l_{K+1},K+1}, ..., Y_{2K} \in S_{l_{2K},2K} \mid \boldsymbol{\alpha})$, and $\mathbf{N}_3$ be a $\kappa_{2k+1} \ldots \kappa_J \times 2^K$ matrix whose $((l_{2K+1}, \ldots, l_J), \boldsymbol{\alpha})$-th entry is $\mathbb{P}(Y_{2K+1} \in S_{l_{2K+1},1}, ..., Y_J \in S_{l_J,J} \mid \boldsymbol{\alpha})$. Then, we can write $\mathbf{N}_2 = \bigotimes_{j=K+1}^{2K} \Lambda^j$ and $\mathbf{N}_3 = \odot_{j=2K+1}^{J} \Psi^j$, where $\odot$ denotes the columnwise Khatri-Rao product. For notational simplicity, let $\eta_1 = \prod_{k=1}^{K} \kappa_k, \eta_2 = \prod_{k=K+1}^{2K} \kappa_k, \eta_3 = \prod_{k=2K+1}^{J} \kappa_k$. Similar to $\mathbf{N}_1$, we index the rows of $\mathbf{N}_2$ and $\mathbf{N}_3$ by $(l_{K+1}, \ldots, l_{2K})$ and $(l_{2K+1}, \ldots, l_J)$, respectively.

Next, we define the marginal probability tensor $\mathbf{P}$ of size $\kappa_1 \times \ldots \times \kappa_J$ by setting $\mathbf{P}(l_1, \ldots, l_J) = \mathbb{P}(\mathbf{Y} \in \times_{j=1}^{J} S_{l_j,j})$. Using (4), we can write

$$\begin{aligned}
\mathbb{P}(\mathbf{Y} \in \times_{j=1}^{J} S_{l_j,j}) = & \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \mathbb{P}(\mathbf{Y}_{1:K} \in \times_{j=1}^{K} S_{l_j,j} \mid \boldsymbol{\alpha}) \mathbb{P}(\mathbf{Y}_{K+1:2K} \in \times_{j=K+1}^{2K} S_{l_j,j} \mid \boldsymbol{\alpha}) \\
& \qquad \mathbb{P}(\mathbf{Y}_{2K+1:J} \in \times_{j=2K+1}^{J} S_{l_j,j} \mid \boldsymbol{\alpha}) \\
= & \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \mathbf{N}_1((l_1, \ldots, l_K), \boldsymbol{\alpha}) \mathbf{N}_2((l_{K+1}, \ldots, l_{2K}), \boldsymbol{\alpha}) \mathbf{N}_3((l_{2K+1}, \ldots, l_J), \boldsymbol{\alpha})
\end{aligned}$$

Viewing $\mathbf{P}_0$ as the unfolded 3-way tensor of $\mathbf{P}$ (with size $\eta_1 \times \eta_2 \times \eta_3$), $\mathbf{P}_0$ can be written as the tensor product:

$$\mathbf{P}_0 = [\mathbf{N}_1 \mathrm{Diag}(\boldsymbol{p}), \mathbf{N}_2, \mathbf{N}_3].$$

**Step 3: Applying Kruskal's Theorem for three-way tensor decomposition.** Let $rk_k$ denote the Kruskal rank of a matrix, defined as the maximum value of $r$ such that any $r$ columns of the matrix are independent (Kruskal, 1977; Derksen, 2013). Now we claim that under our assumptions A and B,

$$rk_k(\mathbf{N}_1\text{Diag}(\boldsymbol{p})) + rk_k(\mathbf{N}_2) + rk_k(\mathbf{N}_3) \geq 2^{K+1} + 2 \tag{S.5}$$

holds. Assuming (S.5), we can apply Kruskal's theorem (this theorem guarantees the uniqueness of a three-way tensor decomposition; Kruskal, 1977), and the components $\mathbf{N}_1\text{Diag}(\boldsymbol{p}), \mathbf{N}_2, \mathbf{N}_3$ are *uniquely determined* up to a column permutation and scaling (i.e. multiplying each component by a constant).

To show (S.5), we prove that

$$rk_k(\mathbf{N}_1) = 2^K, \quad rk_k(\mathbf{N}_2) = 2^K, \quad rk_k(\mathbf{N}_3) \geq 2. \tag{S.6}$$

Equivalently, we show that $\mathbf{N}_1$ and $\mathbf{N}_2$ have full column rank of $2^K$, and any two columns of $\mathbf{N}_3$ are linearly independent.

To compute the rank of $\mathbf{N}_1$, recall that $\mathbf{N}_1 = \bigotimes_{j=1}^{K} \Lambda^j$. Here, $\Lambda^j$ is a $\kappa_j \times 2$ matrix with full column rank, due to the "fully distinguishable" assumption of $\mathcal{D}_j$. This is because we assume that $\mathbb{P}(Y_j \in S_{l_j,j} \mid \alpha_j = 1) \neq \mathbb{P}(Y_j \in S_{l_j,j} \mid \alpha_j = 0)$ for some $l_j \in [\kappa_j]$. Hence, using the multiplicative property of ranks under the Kronecker product,

$$rk(\mathbf{N}_1) = rk\left(\bigotimes_{j=1}^{K} \Lambda^j\right) = \prod_{j=1}^{K} rk(\Lambda^j) = 2^K$$

and $\mathbf{N}_1$ has full rank. Hence, $rk_k(\mathbf{N}_1) = 2^K$. Similarly, $rk_k(\mathbf{N}_2) = 2^K$.

Now, we show that $rk_k(\mathbf{N}_3) \geq 2$ by checking that any two columns in $\mathbf{N}_3$ are linearly independent. Because the last row of $\mathbf{N}_3$ is the all-one vector (since $S_{\kappa_j,j} = \mathcal{Y}_j$), it suffices to show that all two columns are distinct. Take any two columns indexed by $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. Assumption B implies that there exist $j > 2K$ and $l_j$ such that $\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j}) \neq \mathbb{P}_{j,\boldsymbol{\alpha}'}(S_{l_j,j})$. Then,

$$\mathbf{N}_3((\kappa_{2K+1}, \ldots \kappa_{j-1}, l_j, \kappa_{j+1}, \ldots, \kappa_J), \boldsymbol{\alpha}) = \mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j})$$

$$\neq \mathbb{P}_{j,\boldsymbol{\alpha}'}(S_{l_j,j}) = \mathbf{N}_3((\kappa_{2K+1}, \ldots \kappa_{j-1}, l_j, \kappa_{j+1}, \ldots, \kappa_J), \boldsymbol{\alpha}')$$

and the $\boldsymbol{\alpha}$-th column and $\boldsymbol{\alpha}'$-th column are distinct.

**Step 4: Using the decomposed components to identify model parameters.** We denote $\mathbf{P}_0 = [\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3]$ to be one tensor decomposition, and prove that we can recover the true components $\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3$. Here, $\mathbf{M}_i$'s are matrices of size $\boldsymbol{\eta}_i \times 2^K$, and are identical to $\mathbf{N}_i$'s up to a column permutation and scaling.

Because we assume $\Lambda^j(\alpha_j, \kappa_j) = \mathbb{P}_{j,\alpha_j}(S_{\kappa_j,j}) = 1$ for all $j$, the last row of $\mathbf{N}_i$'s is the one vector for all $i = 1, 2, 3$. Consequently, we can set the last rows of $\mathbf{M}_2$ and $\mathbf{M}_3$ as the one vector, and the scaling issue is resolved. In other words, $\mathbf{M}_i$'s are identified up to a permutation of the $2^K$ columns. It remains to identify the model parameters from $\mathbf{M}_i$'s. As we are proving different conclusions, we present separate proofs under the setting in Theorem 1 and Proposition 1.

**Under the setting of Theorem 1:** In order to identify the model parameters from $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$, we need to represent each column as a binary vector of length $K$. One subtlety here comes from the omitted notation $(t)$. Recall that we are working with an increasing collection of sets $\mathcal{D}_j^{(t)}$, and we desire all estimated $\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j})$ and $p_{\boldsymbol{\alpha}}^{(t)}$'s to be the same for all $t$. Otherwise, the values are not well-defined. This argument is unnecessary for binary response CDMs (because $\mathcal{D}_j = \{\{0\}, \{0, 1\}\}$ and considering a single $t$ is enough), but crucial for our proof strategy for general-response CDMs.

To assign a distinct binary vector of length $K$ for each column, we focus on the $\boldsymbol{\eta}_2^{(t)} \times 2^K$ matrix $\mathbf{M}_2^{(t)}$, where the rows are indexed by $(l_{K+1}, \ldots, l_{2K})$ and the columns are indexed by $r \in [2^K]$. Let

$$v_{j,r,l_j}^{(t)} := \sum_{l_{j'} \in [\kappa_{j'}],} \quad \sum_{\forall j' \text{ s.t. } K+1 \leq j' \neq j \leq 2K} \mathbf{M}_2^{(t)}((l_{K+1}, \ldots, l_{2K}), r)$$

for all $j \in \{K+1, \ldots, 2K\}, r \in [2^K]$, and $l_j \in [\kappa_j^{(t)}]$. For fixed $j, l_j$, the set $\{v_{j,r,l_j}^{(t)} : r = 1, \ldots, 2^K\}$ can be interpreted as the set of desired probability values $\{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{j,l_j}) : \boldsymbol{\alpha} \in \{0, 1\}^K\}$. Since the true $\mathbf{Q}$-matrix corresponding to the items considered in $\mathbf{M}_2^{(t)}$ is the

identity matrix $\mathbf{I}_K$, $Y_j$ only measures $\alpha_{j-K}$ for $j = K+1, \ldots, 2K$. Therefore, the vector

$$\mathbf{v}_{j,r}^{(t)} = \left(v_{j,r,1}^{(t)}, \ldots, v_{j,r,\kappa_j}^{(t)}\right)$$

only depends on the value of the $j-K$th skill, and $\{\mathbf{v}_{j,r}^{(t)}\}_r$ can take only two values. Based on this observation, we assign a distinct binary vector of length $K$ to each row of $\mathbf{M}_2^{(t)}$, denoted as $\mathbf{a}^{(t)} = (a_1^{(t)}, \ldots, a_K^{(t)})$, as follows:

1. If $t = 1$, for each $k \in [K]$, cluster the $2^K$ rows into two disjoint sets $V_{k,0}^{(1)}, V_{k,1}^{(1)}$, where each group corresponds to $r \in [2^K]$'s with equal $\mathbf{v}_{K+k,r}^{(1)}$.

2. If $t \geq 2$, for each $k \in [K]$, let $M_{k+K}^{(t)}$ be the set of all indices $m_{k+K} \in [\kappa_{k+K}^{(t)}]$ such that $S_{m_{k+K},k+K}^{(t)} \in \mathcal{D}_{k+K}^{(t-1)}$. Define $V_{k,0}^{(t)} = \{s \in [2^K] : \mathbf{v}_{k+K,s,M_{k+K}^{(t)}}^{(t)} = \mathbf{v}_{k+K,r}^{(t-1)}, \forall r \in V_{k,0}^{(t-1)}\}$ and similarly define $V_{k,1}^{(t)}$. (Here, $\mathbf{v}_{k+K,s,M_{k+K}^{(t)}}^{(t)}$ is the sub-vector of $\mathbf{v}_{k+K,s}^{(t)}$ consisting of elements indexed by $M_{k+K}^{(t)}$.)

3. For each $k \in [K]$, let $a_k^{(t)} = 0$ for the rows indexed by $V_{k,0}^{(t)}$, and 1 for the rows indexed by $V_{k,1}^{(t)}$.

Now, for the vectors $\mathbf{v}_{j,\cdot}^{(t)}$ and matrices $\mathbf{M}_{\cdot}^{(t)}$, we use the (column-)index $\mathbf{a}^{(t)}$. By construction, $\mathbf{a}^{(t)}$ is consistent across $t$ in the sense that $v_{k+K,\mathbf{a},l_{k+K}^{(t)}}^{(t)} = v_{k+K,\mathbf{a},l_{k+K}^{(t')}}^{(t')}$ when $S_{l_{k+K}^{(t)},k+K}^{(t)} = S_{l_{k+K}^{(t')},k+K}^{(t')}$.

Compared to the true latent skill pattern $\boldsymbol{\alpha}$, which we used to denote the column index of $\mathbf{N}_2$, $\mathbf{a}$ is identical up to a sign flip for each coordinate. Hence, there exists permutations $\tau_k \in S_{\{0,1\}}$ such that

$$\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j}^{(t)}) = v_{j,(\tau_1(\alpha_1),\ldots,\tau_K(\alpha_K)),l_j}^{(t)}$$

for all $j \in [K+1, 2K], \boldsymbol{\alpha} \in \{0,1\}^K$, $t \geq 1$. So, we have identified $\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j}^{(t)})$ for $j \in [K+1, 2K]$. Then, $\boldsymbol{p}$ is also identified up to the permutations $\tau_k$'s, as this is exactly the last row of $\mathbf{M}_1$. For $j \notin [K+1, 2K]$, $\{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j}^{(t)}) : \ l_j \in [\kappa_j]\}$ is also determined by marginalizing out all other rows in the $\mathbf{M}_1, \mathbf{M}_3$. Hence, for all $t$, the model components $(\boldsymbol{p}, \{\mathbb{P}_{j,\boldsymbol{\alpha}}(S_{l_j,j}) : S_{l_j,j} \in \mathcal{D}_j^{(t)}\})$ can be uniquely determined up to the permutation $\tau_k$'s.

**Under the setting of Proposition 1:** Here, we assume that the responses are real-valued with $\mathcal{Y}_j \subseteq \mathbb{R}$. By the conclusion of Theorem 1, $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s are determined up to a sign flip. For

$k = 1, \ldots, K$, the $k$th item is measured only by the $k$th attribute, and (17) can be simplified as

$$\mathbb{E}(Y_k \mid A_k = 1) > \mathbb{E}(Y_k \mid A_k = 0).$$

This assumption enables us to distinguish between $A_k = 1$ and 0, and fixes the sign of $A_k$. Hence, the columns of $\mathbf{M}_1$'s uniquely determine the latent configuration $\boldsymbol{\alpha} \in \{0, 1\}^K$. Now, the proportion parameter $p_{\boldsymbol{\alpha}}$'s can be determined by observing the last row of $\mathbf{M}_1$. Also, for any $j \in [J], l_j \in [\kappa_j]$, $\mathbb{P}_{j,\boldsymbol{\alpha}}(S_j)$ can be obtained by marginalizing out the other rows. Hence, all model components can be identified. □

### S.3.2 Proof of Proposition 2

*Proof.* Suppose that $\mathbf{Q}^*$ contains $I_K$. Without the loss of generality, assume that

$$\mathbf{Q}^* = [\mathbf{I}_K, \mathbf{Q}^\dagger]^\top.$$

For any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$, there exists $k \in [K]$ such that $\alpha_k \neq \alpha'_k$. By our assumption, the $2K + k$th item only measures $A_k$, and the $(2K + k)$th row of the $\mathbf{Q}$-matrix is the standard basis vector $\boldsymbol{e}_k$. Then, the definition of the $\mathbf{Q}$-matrix gives $\mathbb{P}_{2K+k,\boldsymbol{\alpha}} \neq \mathbb{P}_{2K+k,\boldsymbol{\alpha}'}$ and condition B holds. □

### S.3.3 Proof of Proposition 3

*Proof.* Recall that $\mathbb{P}_{j,\boldsymbol{\alpha}}$ for the ExpACDM can be written as a parametric density $g(y; \beta_{j,0} + \sum_{k=1}^K \beta_{j,k}\alpha_k, \gamma_j)$. For any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$, $\mathbb{P}_{j,\boldsymbol{\alpha}} \neq \mathbb{P}_{j,\boldsymbol{\alpha}'}$ if and only if $\sum_{k=1}^K \beta_{j,k}\alpha_k \neq \sum_{k=1}^K \beta_{j,k}\alpha'_k$, and condition B reduces to B2. □

### S.3.4 Proof of Theorem 2

*Proof.* In this proof, we use the notations introduced in the main manuscript and the proof of Theorem 1. Recall that in the proof of Theorem 1, $\mathbf{N}_1$ is defined as a $(\kappa_1 ... \kappa_K) \times 2^K$ matrix whose $((l_1, ..., l_K), \boldsymbol{\alpha})$-th entry is $\mathbb{P}(Y_1 \in S_{l_1,1}, ..., Y_K \in S_{l_K,K} \mid \boldsymbol{\alpha}) = \prod_{k=1}^K \mathbb{P}(Y_k \in S_{l_k,k} \mid \boldsymbol{\alpha})$. In this proof, we further assume that $\mathbb{P}_{j,\boldsymbol{\alpha}}$'s follow an ExpACDM with parameters $(\boldsymbol{\beta}, \gamma)$, and that conditions $A^\star$ and $B^\star$ hold. We claim that it suffices to show that $\mathbf{N}_1, \mathbf{N}_2, \mathbf{N}_3$ satisfy (S.6) for generic parameters. Assuming this, (S.5) holds, and Step 3 in the proof of Theorem

1 holds almost surely. Consequently, we can apply Step 4 in the proof of Theorem 1, and the model components are identifiable up to a measure zero set.

Now, we present the proof of (S.6) for generic parameters. We separately show that $rk_k(\mathbf{N}_1) = 2^K$ and $rk_k(N_3) \geq 2$. This proof is motivated by existing works that show generic identifiability for binary CDMs (Gu and Xu, 2020; Chen et al., 2020).

**Proof of $rk_k(\mathbf{N}_1) = 2^K$.** Note that $\mathbf{N}_1$ can be seen as a function of $\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}$ and we abuse the notation and also write $\mathbf{N}_1$ in the functional form $\mathbf{N}_1(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K})$. We prove that if the first $K$ rows of the $\mathbf{Q}$-matrix are given as $\mathbf{Q}_1$ in Theorem 2, $\mathbf{N}_1$ generically has full column rank. Because $\mathbf{N}_1$ has full column rank if and only if $\det(\mathbf{N}_1^\top \mathbf{N}_1) \neq 0$, it suffices to show that

$$\{\boldsymbol{\beta}_{1:K,0:K} \in \Omega(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}; \mathbf{Q}_1) : \det(\mathbf{N}_1^\top \mathbf{N}_1) \neq 0\}$$

is a measure 0 set in $\Omega(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}; \mathbf{Q}_1)$.

First, we consider the case when $\mathbf{Q}_1 = \mathbf{I}_K$. We can use the calculation of the previous theorem to see that $\mathbf{N}_1(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K})$ has full column rank for all $\boldsymbol{\beta}_{1:K,0:K} \in \Omega(\boldsymbol{\beta}_{1:K,0:K}; \mathbf{Q}_1 = \mathbf{I}_K) = \{\boldsymbol{\beta}_{1:K,0:K} : \beta_{j,j} \neq 0, \beta_{j,k} = 0 \text{ for all } 1 \leq j \neq k \leq K\}$ and any $\gamma_{1:K}$.

Next, for any $\mathbf{Q}^*$ whose diagonal entries are ones, suppose that the true $\mathbf{Q}_1$ is $\mathbf{Q}^*$. Consider the mapping

$$(\det(\mathbf{N}_1^\top \mathbf{N}_1))(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}) : \Omega(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}; \mathbf{Q}_1 = \mathbf{Q}^*) \to \mathbb{R}. \tag{S.7}$$

Observe that $\det(\mathbf{N}_1^\top \mathbf{N}_1)$ defined in (S.7) is a polynomial of entries of $\mathbf{N}_1(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K})$. Because $g$ is a density of an exponential family, $g(Y; \boldsymbol{\eta})$ is an analytic function in $\boldsymbol{\eta}$. Also, recall that we assume $\boldsymbol{h}$ is also analytic. Therefore, all entries of $\mathbf{N}_1(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K})$ can be written as a composition of analytic functions, and is also analytic. Consequently, the map $\det(\mathbf{N}_1^\top \mathbf{N}_1)$ is a polynomial of analytic functions, and is also analytic. Also note that the domain of (S.7) is an open, connected subset of a Euclidean space of appropriate dimension.

Finally, we claim that $\det(\mathbf{N}_1^\top \mathbf{N}_1)$ is not identically zero, i.e. there exists $(\boldsymbol{\beta}^\star_{1:K,0:K}, \gamma^\star_{1:K})$ such that $\det(\mathbf{N}_1^\top \mathbf{N}_1)(\boldsymbol{\beta}^\star_{1:K,0:K}, \gamma^\star_{1:K}) \neq 0$. This follows from noting that there exists $(\boldsymbol{\beta}^\dagger_{1:K,0:K}, \gamma^\dagger_{1:K}) \in \Omega(\boldsymbol{\beta}, \gamma; \mathbf{Q}_1 = \mathbf{I}_K)$ that has a nonzero determinant, and we can find $(\boldsymbol{\beta}^\star_{1:K,0:K}, \gamma^\star_{1:K})$ arbitrarily close to $(\boldsymbol{\beta}^\dagger_{1:K,0:K}, \gamma^\dagger_{1:K})$ because $\det(\mathbf{N}_1^\top \mathbf{N}_1)$ is a continuous mapping. Hence, we conclude

that

$$\{\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K} \in \Omega(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}; \mathbf{Q}_1 = \mathbf{Q}^*) : (\det(\mathbf{N}_1^\top \mathbf{N}_1))(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}) = 0\}$$

is a null set in $\Omega(\boldsymbol{\beta}_{1:K,0:K}, \gamma_{1:K}, \mathbf{Q}_1 = \mathbf{Q}^*)$ by the following Lemma. This is true for any $\mathbf{Q}_1 = \mathbf{Q}^*$, so the proof is complete. We can apply the same argument to show $rk_k(\mathbf{N}_2) = 2^K$ as well.

**Lemma 1** (Lemma 5 of Chen et al. (2020); see Mityagin (2020) for a proof)**.** *Let $f : \Omega \to \mathbb{R}$ be a real analytic function defined on a open, connected domain $\Omega \in \mathbb{R}^d$ that is not identically zero. Then, the set of zeros of $f$ has Lebesgue measure 0.*

**Proof of $rk_k(\mathbf{N}_3) \geq 2$.** In the proof of the previous theorem, we have proved $rk_k(\mathbf{N}_3) \geq 2$ under assuming condition B. Hence, it suffices to show that condition B2 (in proposition 3) holds under generic parameters. Recall that $\mathbb{P}_{j,\boldsymbol{\alpha}}$ can be written as a parametric density $g(y; \beta_{j,0} + \sum_{k=1}^K \beta_{j,k}\alpha_k, \gamma_j)$. For any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$, we prove that there exists $j > 2K$ such that $\sum_{k=1}^K \beta_{j,k}\alpha_k \neq \sum_{k=1}^K \beta_{j,k}\alpha'_k$ for generic parameters in $\Omega(\boldsymbol{\beta}_{2K+1:J,1:K}; \mathbf{Q}_3)$.

Fix $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. Then, there exists $k$ such that $\alpha_k \neq \alpha'_k$. By condition B', there exists $j$ such that $q_{j,k} = 1$. Then,

$$\{\boldsymbol{\beta}_{j,1:K} \in \Omega(\boldsymbol{\beta}_{j,1:K}; \mathbf{Q}_3) : \sum_{l=1}^K \beta_{j,l} q_{j,l} (\alpha_l - \alpha'_l) = 0\}$$

is a $\sum_{l \neq k} q_{j,l}$-dimensional (or lower dimensional) Euclidean subset of the $\sum_{1 \leq l \leq K} q_{j,l}$-dimensional Euclidean space $\Omega(\boldsymbol{\beta}_{j,1:K}; \mathbf{Q}_3)$. Hence, this is a measure 0 set in $\Omega(\boldsymbol{\beta}_{2K+1:J,1:K}; \mathbf{Q}_3)$. The finite union of measure 0 sets also has measure 0, and condition B holds almost surely. □

**Remark 1.** In the proof, we use the exponential family assumption only to deduce that $g$ is an analytic function. Hence, the statement can be relaxed to general-response ACDMs without the exponential family assumption as long as the probability mass/density function $g(Y; \boldsymbol{\eta})$ is analytic in $\boldsymbol{\eta}$.

### S.3.5 Proof of Proposition 4

*Proof.* Write $\theta = (\boldsymbol{\eta}, \boldsymbol{p})$. By Theorem 10.1.6 in Casella and Berger (2021), the MLE $\widehat{\theta} = (\widehat{\boldsymbol{\eta}}, \widehat{\boldsymbol{p}})$ is consistent under the following regularity conditions:

(C1) The parameter $\theta$ is identifiable.

(C2) The densities $\mathbb{P}(\mathbf{Y} \mid \theta)$ have common support and is differentiable in $\theta$.

(C3) The parameter space contains an open set of which the true parameter value $\theta_0$ is an interior point.

It suffices to check the three regularity conditions hold. Condition C1 follows from Theorems 1 (under the first set of assumptions) or 2 (under the second set of assumptions). Condition C2 holds because $g(\cdot \mid \boldsymbol{\eta})$ is an exponential family distribution. Condition C3 is exactly the assumption we impose on $\theta_0 = (\boldsymbol{\eta}_0, \boldsymbol{p}_0)$, and the proof is complete. □

## S.4 EM algorithm for the negative binomial-based CDMs

As we have stated the EM Algorithms 1 and 2 assuming exponential family distributions for modeling $Y_j \mid \mathbf{A} = \boldsymbol{\alpha}$, we briefly sketch how they can be modified to non-exponential family distributions. In particular, we display the algorithms for the negative binomial distribution. Recall that we have defined NegBin-DINA and NegBin-ACDM in Section A.2 of this Supplementary Material.

**EM algorithm for the NegBin-DINA Model** The algorithm is similar to the one for exponential family-based DINA model. The main difference is the specific form of the parametrization and the updates for the item parameters. For the sake of completeness, we display the steps in Algorithm 1. Here, we write

$$\mathbb{P}(Y \mid r, \pi) = \binom{Y + r - 1}{Y} (1 - \pi)^Y \pi^r$$

to simplify the presentation. Note that there are no closed updates for the item parameters in the M step and here we solve a two-dimensional equation. One may choose to directly apply an optimization software instead in the M step.

**Algorithm 1:** EM algorithm for the NegBin-DINA

**Data:** response $\mathbf{Y}$, $\mathbf{Q}$-matrix $\mathbf{Q}$

Initialize $r_j, \pi_j, p_{\boldsymbol{\alpha}}$'s.

**while** *log-likelihood has not converged* **do**

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

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

$$\begin{aligned}\varphi_{i,\boldsymbol{\alpha}}^{(t+1)} &= \mathbb{P}(A_i = \boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{r}^{(t)}, \boldsymbol{p}^{(t)}, \pi^{(t)}) \\ &= \frac{p_{\boldsymbol{\alpha}}^{(t)} \prod_j \mathbb{P}(Y_{i,j} \mid r_{j,0}, \pi_{j,0})^{(1-\Gamma_{j,\boldsymbol{\alpha}})} \mathbb{P}(Y_{i,j} \mid r_{j,1}, \pi_{j,1})^{\Gamma_{j,\boldsymbol{\alpha}}}}{\sum_{\boldsymbol{\alpha}'} p_{\boldsymbol{\alpha}'}^{(t)} \prod_j \mathbb{P}(Y_{i,j} \mid r_{j,0}, \pi_{j,0})^{(1-\Gamma_{j,\boldsymbol{\alpha}'})} \mathbb{P}(Y_{i,j} \mid r_{j,1}, \pi_{j,1})^{\Gamma_{j,\boldsymbol{\alpha}'}}};\end{aligned}$$

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

$$p_{\boldsymbol{\alpha}}^{(t+1)} = \frac{\sum_i \varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_{i,\boldsymbol{\alpha}'} \varphi_{i,\boldsymbol{\alpha}'}^{(t+1)}};$$

**for** $j \in [J], h \in \{0,1\}$ **do**

Solve the series of equations for $(r_{j,h},\ \pi_{j,h})$:

$$\begin{aligned}&\frac{\sum_{i,\boldsymbol{\alpha}} \varphi_{i,\boldsymbol{\alpha}} (1-\Gamma_{j,\boldsymbol{\alpha}})^{1-h} \Gamma_{j,\boldsymbol{\alpha}}^h}{\sum_{i,\boldsymbol{\alpha}} \varphi_{i,\boldsymbol{\alpha}} (1-\Gamma_{j,\boldsymbol{\alpha}})^{1-h} \Gamma_{j,\boldsymbol{\alpha}}^h (Y_{i,j} + r_{j,h})} = \pi_{j,h}, \\ &\sum_{i,\boldsymbol{\alpha}} \varphi_{i,\boldsymbol{\alpha}} (1-\Gamma_{j,\boldsymbol{\alpha}})^{1-h} \Gamma_{j,\boldsymbol{\alpha}}^h \left( \log(\pi_{j,h}) + \sum_{m=0}^{Y_{i,j}-1} \frac{1}{r_{j,h}+m} \right) = 0.\end{aligned}$$

**Output:** $r_j, \pi_j, p_{\boldsymbol{\alpha}}$'s.

**EM algorithm for the NegBin-ACDM** For the NegBin-ACDM, we continue working with the parametrization of $(\boldsymbol{\beta}_j, \gamma_j)$. Recall that we have defined the negative binomial parameters in terms of $(\boldsymbol{\beta}_j, \gamma_j)$ in (S.4). Even though the negative binomial is not an exponential family, $\boldsymbol{h}$ defined in (S.4) plays the exact same role as $\boldsymbol{h}$ in the context of exponential families. Hence, $(r_{j,\boldsymbol{\alpha}}, \pi_{j,\boldsymbol{\alpha}})$ in (S.4) are linear combinations of the latent skills, and plays an almost identical role as the natural parameter $\boldsymbol{\eta}_{j,\boldsymbol{\alpha}}$ in Algorithm 2. The other parts of the algorithm are almost identical, and we present the details in the following Algorithm 2.

**Algorithm 2:** EM algorithm for the main-effect model with negative binomial link

**Data:** response $\mathbf{Y}$, $\mathbf{Q}$-matrix $\mathbf{Q}$

Initialize $\boldsymbol{\beta}_j, \gamma_j, p_{\boldsymbol{\alpha}}$'s.

**while** *log-likelihood has not converged* **do**

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

**for** $j \in [J], \boldsymbol{\alpha} \in \{0,1\}^K$ **do**

$$r_{j,\boldsymbol{\alpha}}^{(t)} = h(\beta_{j,0} + \textstyle\sum_k \beta_{j,k} q_{j,k} \alpha_k, \gamma_j^{(t)}) = \frac{\gamma_j^{(t)}}{1-\gamma_j^{(t)}} (\beta_{j,0}^{(t)} + \textstyle\sum_k \beta_{j,k}^{(t)} q_{j,k} \alpha_k)$$

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

$$\varphi_{i,\boldsymbol{\alpha}}^{(t+1)} = \mathbb{P}(A_i = \boldsymbol{\alpha} \mid \mathbf{Y}, \boldsymbol{\beta}_j^{(t)}, \gamma_j^{(t)}, p_{\boldsymbol{\alpha}}^{(t)}) \;=\; \frac{p_{\boldsymbol{\alpha}} \prod_j \mathbb{P}(Y_{i,j} \mid r_{j,\boldsymbol{\alpha}}^{(t)}, \gamma_j^{(t)})}{\sum_{\boldsymbol{\alpha}'} p_{\boldsymbol{\alpha}'} \prod_j \mathbb{P}(Y_{i,j} \mid r_{j,\boldsymbol{\alpha}'}^{(t)}, \gamma_j^{(t)})};$$

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

$$p_{\boldsymbol{\alpha}}^{(t+1)} = \frac{\sum_i \varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_{i,\boldsymbol{\alpha}'} \varphi_{i,\boldsymbol{\alpha}'}^{(t+1)}};$$

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

$$(\boldsymbol{\beta}_j, \gamma_j)^{(t+1)} = \underset{\boldsymbol{\beta}_j, \gamma_j}{\operatorname{argmax}} \sum_{i,\boldsymbol{\alpha}} \Big[ \log \binom{Y_{i,j} + r_{j,\boldsymbol{\alpha}} - 1}{Y_{i,j}} + r_{j,\boldsymbol{\alpha}} \log \gamma \Big] \varphi_{i,\boldsymbol{\alpha}}^{(t+1)} + \sum_i Y_{i,j} \log(1-\gamma)$$

subject to $\beta_{j,k} = 0$ if and only if $q_{j,k} = 0$.

Here, $r_{j,\boldsymbol{\alpha}} = h(\beta_{j,0} + \sum_k \beta_{j,k} q_{j,k} \alpha_k, \gamma)$.

**Output:** $\boldsymbol{\beta}_j, \gamma_j, p_{\boldsymbol{\alpha}}$'s.

# S.5 Simulation details and additional simulations

## S.5.1 Simulation details

In Tables S.1, S.2, S.3, we display the average root mean squared error (RMSE) values in all simulation studies described in Section 5 of the main manuscript. In all simulations, the proportion parameters $\boldsymbol{p}$ are initialized as Dirichlet random variables with parameters $(1, 1, \ldots, 1)$. The item parameters are initialized by adding a uniform random noise to the true parameters. We set the convergence criterion of the EM algorithm by terminating it when the increment of the log-likelihood is smaller than 0.05. Typically, this criterion is met before 10 EM iterations, regardless of which model is considered.

Also, we report the number of iterations and computation time under the lognormal

| Link function | Model | $N$ | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{\mu}}$ or $\widehat{\boldsymbol{\beta}}$) | RMSE($\widehat{\boldsymbol{\sigma}}$ or $\widehat{\boldsymbol{\gamma}}$) |
|---|---|---|---|---|---|
| Transformed-Normal | ExpDINA | 100 | 0.0196 | 0.230 | 0.357 |
| | | 500 | 0.0080 | 0.071 | 0.104 |
| | | 1000 | 0.0056 | 0.050 | 0.075 |
| | | 1500 | 0.0045 | 0.041 | 0.059 |
| | | 2000 | 0.0039 | 0.036 | 0.053 |
| | ExpACDM | 100 | 0.0177 | 0.185 | 0.146 |
| | | 500 | 0.0080 | 0.083 | 0.065 |
| | | 1000 | 0.0057 | 0.057 | 0.045 |
| | | 1500 | 0.0045 | 0.046 | 0.036 |
| | | 2000 | 0.0039 | 0.041 | 0.032 |

Table S.1: RMSE for the estimated transformed-Normal parameters, $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\sigma}})$ under ExpDINA and $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\beta}}, \widehat{\gamma})$ under ExpACDM.

| Link | Model | $N$ | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{\lambda}}$ or $\widehat{\boldsymbol{\beta}}$) |
|---|---|---|---|---|
| Poisson | ExpDINA | 100 | 0.0252 | 0.281 |
| | | 500 | 0.0115 | 0.122 |
| | | 1000 | 0.0081 | 0.085 |
| | | 1500 | 0.0065 | 0.069 |
| | | 2000 | 0.0055 | 0.061 |
| | ExpACDM | 100 | 0.0278 | 0.311 |
| | | 500 | 0.0124 | 0.127 |
| | | 1000 | 0.0086 | 0.089 |
| | | 1500 | 0.0070 | 0.072 |
| | | 2000 | 0.0055 | 0.062 |

Table S.2: RMSE for the estimated Poisson parameters, $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\lambda}})$ under ExpDINA and $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\beta}})$ under ExpACDM.

link in Table S.4. One can see that the computation time increases with the sample size $N$, whereas the number of iterations does not depend on $N$. The estimation time for the ExpACDM is larger compared to the ExpDINA because we utilize the optimization toolbox in Matlab for the M-step of Algorithm 2. Nonetheless, the average computation time is still less than 15 seconds for each iteration, even when the sample size is as large as $N = 2000$.

| Link | Model | $N$ | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{r}}$ or $\widehat{\boldsymbol{\beta}}$) | RMSE($\widehat{\boldsymbol{\pi}}$ or $\widehat{\boldsymbol{\gamma}}$) |
|---|---|---|---|---|---|
| NegBin | ExpDINA | 100 | 0.0248 | 1.26 | 0.132 |
| | | 500 | 0.0113 | 0.75 | 0.067 |
| | | 1000 | 0.0078 | 0.55 | 0.048 |
| | | 1500 | 0.0064 | 0.45 | 0.040 |
| | | 2000 | 0.0056 | 0.41 | 0.035 |
| | ExpACDM | 100 | 0.0352 | 0.463 | 0.103 |
| | | 500 | 0.0194 | 0.219 | 0.048 |
| | | 1000 | 0.0140 | 0.151 | 0.033 |
| | | 1500 | 0.0115 | 0.125 | 0.027 |
| | | 2000 | 0.0097 | 0.105 | 0.022 |

Table S.3: RMSE for the estimated negative binomial parameters, $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{r}}, \widehat{\boldsymbol{\pi}})$ under ExpDINA and $(\widehat{\boldsymbol{p}}, \widehat{\boldsymbol{\beta}})$ under ExpACDM.

| Link | Model | N | runtime (s) | # of iterations |
|---|---|---|---|---|
| Lognormal | ExpDINA | 100 | 0.08 | 4.60 |
| | | 500 | 0.35 | 4.13 |
| | | 1000 | 0.73 | 4.16 |
| | | 1500 | 1.27 | 4.15 |
| | | 2000 | 1.67 | 4.20 |
| | ExpACDM | 100 | 5.22 | 2.97 |
| | | 500 | 8.54 | 2.71 |
| | | 1000 | 11.83 | 2.61 |
| | | 1500 | 12.35 | 2.59 |
| | | 2000 | 15.12 | 2.60 |

Table S.4: The average number of iterations and runtime for CDMs with lognormal link

### S.5.2 Additional simulations under generic identifiability

Here, we conduct additional simulation studies under a generic identifiable model to empirically assess estimation consistency. Similar to Section 5.1, we still consider the transformed-Normal ACDM with $K = 5, J = 20$. But now we consider a $\mathbf{Q}$-matrix that only satisfies the generic identifiability conditions in Theorem 2 (see the entries in bold font below), but does not satisfy the strict identifiability conditions in Theorem 1. More specifically, we assume

that each row of the $\mathbf{Q}$-matrix has exactly two nonzero entries:

$$\mathbf{Q} := \begin{pmatrix} \mathbf{Q}^{\text{sub}} \\ \mathbf{Q}^{\text{sub}} \\ \mathbf{Q}^{\text{sub}} \\ \mathbf{Q}^{\text{sub}}_{1:2} \end{pmatrix}, \quad \text{where } \mathbf{Q}^{\text{sub}} = \begin{pmatrix} \mathbf{1} & 0 & 0 & 1 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 1 \\ 0 & 0 & 1 & \mathbf{1} & 0 \\ 1 & 0 & 0 & 0 & \mathbf{1} \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix},$$

where $\mathbf{Q}^{\text{sub}}_{1:2}$ denotes the first two rows of $\mathbf{Q}^{\text{sub}}$. All continuous parameters are specified in the same way as in Section 5.1 in the main paper. With the above $\mathbf{Q}$-matrix, the model does not satisfy the strict identifiability conditions in Theorem 1 since the $\mathbf{Q}$-matrix does not contain an identity matrix.

Table S.5 presents the estimation results. We observe that the RMSEs of the continuous parameters decrease as the sample size $N$ increases, actually similarly at a $\frac{1}{\sqrt{N}}$ rate as in the strictly identifiable case. This result empirically demonstrates that this generic identifiable transformed-Normal ACDM can still be consistently estimated by our EM algorithm 2. We also remark that the structure of the $\mathbf{Q}$-matrix we consider here is similar to the one used in the real data analysis in Section 6, by viewing the first three attributes as content skills, and the next two as cognitive skills. Hence, the results in Table S.5 also suggest consistent parameter estimation for our real data analysis.

| Link | Model | $N$ | RMSE($\widehat{\boldsymbol{p}}$) | RMSE($\widehat{\boldsymbol{\beta}}$) | RMSE($\widehat{\boldsymbol{\gamma}}$) |
|---|---|---|---|---|---|
| Transformed-Normal | ExpACDM | 100 | 0.0206 | 0.156 | 0.156 |
| | | 500 | 0.0100 | 0.066 | 0.070 |
| | | 1000 | 0.0067 | 0.047 | 0.048 |
| | | 1500 | 0.0054 | 0.038 | 0.040 |
| | | 2000 | 0.0049 | 0.033 | 0.034 |

Table S.5: RMSE for the estimated transformed-Normal ACDM parameters.

# S.6 Data preprocessing details and analysis of the binary response accuracy

## S.6.1 Data preprocessing details

Before analyzing the TIMSS 2019 dataset in Section 6 in the main manuscript, we preprocess the data to deal with outliers. The left panel in Figure S.1 shows that there are two small groups of outliers at the far left and far right of the histogram. We conjecture that the outliers with small response times result from randomly guessing or running out of time, while the outliers with large response times result from taking breaks during the exam. We truncate the log response times to be between 0 and 6; that is, for any log response time smaller than 0, we truncate it to be zero, for any log response time greater than 6, we truncate it to be 6. Moreover, among the 622 students in the dataset, we exclude two students who did not respond to any questions. The preprocessed dataset consists of $N = 620$ students. The mean and the median of the total response time are approximately 30 minutes, as shown in the right panel of Figure S.1. Thus, we believe that a majority of the students completed the exam well within the 45-minute limit.

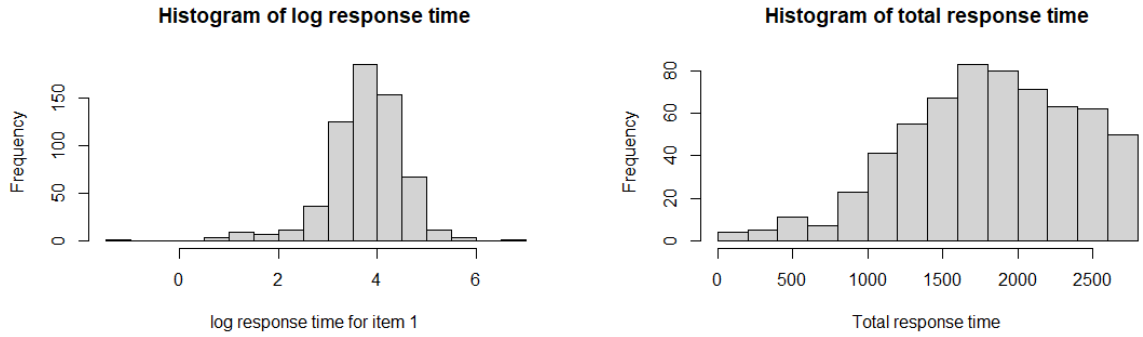


Figure S.1: Histograms of response time in seconds. The left panel shows the log response time for item number 1, and the right panel shows the total response time of the entire exam.

## S.6.2 Analysis of the binary response accuracy

We analyze the binary response accuracy in the TIMSS 2019 dataset and compare the estimated item parameters to those presented in Section 6 in the main paper. We consider

the response accuracy for the same students as in Section 6, and fit the ACDM for binary responses (de la Torre, 2011).

The estimated ACDM item parameters are displayed in Figure S.2. Interestingly, we observe that the larger entries in Figure 6 tend to correspond to smaller entries in Figure S.2. In particular, the intercepts $\{\beta_{j,0}\}_{j\in[J]}$ in the binary-response ACDM (the first column of Figure S.2) also show a discrepancy between the items belonging to the two different types: constructed-response items versus multiple-choice items. Note that intercept parameters in a binary-response ACDM can be interpreted as the guessing probability for those students lacking any required skills of an item. Therefore, it is very interpretable that the constructed-response items (such as items 2, 4, 7) have much smaller intercepts compared to those multiple choice items, which have intercepts fluctuating around the theoretical guessing probability of 1/4 (since there are four options in these multiple-choice items, random guessing will yield a 1/4 correct response rate).

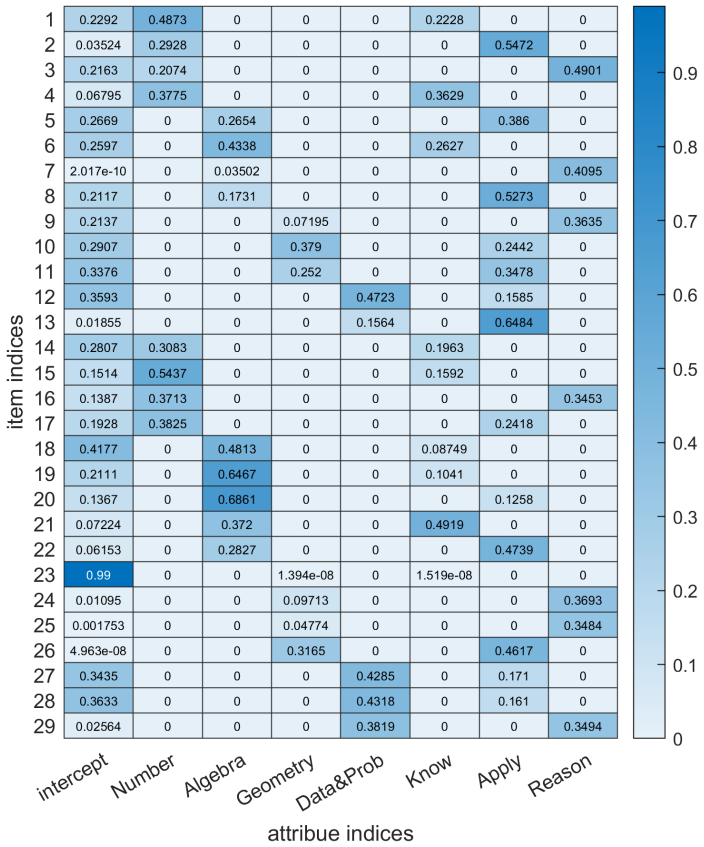


Figure S.2: Heatmap of the ACDM coefficients estimated from the binary response accuracy.

# 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(6A):3099–3132.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

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

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

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

Derksen, H. (2013). Kruskal's uniqueness inequality is sharp. *Linear Algebra and its Applications*, 438(2):708–712.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

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

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.

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, R., Heo, I., Liu, H., Shi, D., and Jiang, Z. (2023). Applying negative binomial distribution in diagnostic classification models for analyzing count data. *Applied Psychological Measurement*, 47(1):64–75.

Liu, R., Liu, H., Shi, D., and Jiang, Z. (2022). Poisson diagnostic classification models: A framework and an exploratory example. *Educational and Psychological Measurement*, 82(3):506–516.

Man, K. and Harring, J. R. (2019). Negative binomial models for visual fixation counts on test items. *Educational and Psychological Measurement*, 79(4):617–635.

Minchen, N. and de la Torre, J. (2018). A general cognitive diagnosis model for continuous-response data. *Measurement: Interdisciplinary Research and Perspectives*, 16(1):30–44.

Mityagin, B. S. (2020). The zero set of a real analytic function. *Mathematical Notes*, 107(3-4):529–530.

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