Exploratory General-response Cognitive Diagnostic Models with Higher-order Structures

Jia Liu*† Seunghyun Lee* Yuqi Gu*

*Department of Statistics, Columbia University
†School of Mathematics and Statistics, Northeast Normal University

Abstract

Cognitive Diagnostic Models (CDMs) are popular discrete latent variable models in educational and psychological measurement. While existing CDMs mainly focus on binary or categorical responses, there is a growing need to extend them to encompass a wider range of response types, including but not limited to continuous and count-valued responses. Meanwhile, incorporating higher-order latent structures has become crucial for gaining deeper insights into cognitive processes. We propose a general modeling framework for higher-order CDMs for rich types of responses. Our framework features a highly flexible data layer that is adaptive to various response types and measurement models for CDMs. Importantly, we address a challenging exploratory estimation scenario where the item-attribute relationship, specified by the Q-matrix, is unknown and needs to be estimated along with other parameters. In the higher-order layer, we employ a probit-link with continuous latent traits to model the binary latent attributes, highlighting its benefits in terms of identifiability and computational efficiency. Theoretically, we propose transparent identifiability conditions for the exploratory setting. Computationally, we develop an efficient Monte Carlo Expectation-Maximization algorithm, which incorporates an efficient direct sampling scheme and requires significantly reduced simulated samples. Extensive simulation studies and a real data example demonstrate the effectiveness of our methodology.

Keywords: Cognitive Diagnostic Models; Identifiability; Monte Carlo Expectation-Maximization (MCEM) Algorithm; Probit Model; Q-matrix.

1 Introduction

Cognitive Diagnostic Models (CDMs), or Diagnostic Classification Models (Templin et al., 2010), have emerged as a crucial tool for modeling educational assessment data with multidimensional

Correspondence to Yuqi Gu, Department of Statistics, Columbia University, New York, NY 10027, USA. Email: yuqi.gu@columbia.edu. This research is conducted when Jia Liu is a visiting PhD student in the Department of Statistics at Columbia University. This version: November 2024.

discrete (often binary) latent variables. Various diagnostic goals lead to CDMs with different measurement models; examples include the Deterministic Input Noisy Output “And” gate model (DINA; Junker and Sijtsma, 2001) with the conjunctive assumption, the Deterministic Input Noisy Output “Or” gate model (DINO; Templin and Henson, 2006) with the disjunctive assumption, the main-effect diagnostic models (DiBello et al., 2012; Maris, 1999; de la Torre, 2011) incorporating the additive effects of latent attributes, and the all-effect general diagnostic models (de la Torre, 2011; Von Davier, 2008; Henson et al., 2009) with a more saturated parameterization.

Currently, most existing CDMs are designed to model binary or polytomous response data. However, the diversification of examination modes and increased availability of educational and psychological data have enabled the collection of various response data types. Continuous response data arise in many scenarios, such as language proficiency tests scoring on a continuous scale and recording response time in computer-based assessments (Minchen et al., 2017). The modeling of response times has long been a topic of interest, see De Boeck and Jeon (2019) for a comprehensive overview. Another common response type is count responses, found in assessments recording the number of correct responses, the frequency of specific behaviors in classroom activities, the usage frequency of particular strategies in problem-solving tasks, and computer-based tests recording visit counts per item (Man and Harring, 2019; Liu et al., 2022). Rasch (1993) first proposed a Poisson-based item response theory (IRT) model for count data, and since then, many other models have been developed (Magnus and Thissen, 2017; Man and Harring, 2019, 2023).

A crucial element in a CDM is the relationship between observed item responses and latent attributes, specified by the Q-matrix (Tatsuoka, 1983). Recently, Lee and Gu (2024b) proposed a new cognitive diagnostic modeling framework for general response types with a prespecified Q-matrix. However, in many practical applications, the true Q-matrix may not be known a priori, necessitating an exploratory approach to infer the Q-matrix directly from the response data. In such challenging exploratory settings for flexible data types, estimating the Q-matrix reliably and efficiently is highly desirable but largely unknown. Beyond exploratory CDMs and general response types, integrating a higher-order layer into CDMs (de la Torre and Douglas, 2004; Templin et al., 2008) offers significant advantages. Such models uses one or more continuous latent traits to explain the binary attributes, providing a more nuanced understanding of the relationships between different skills, yielding a comprehensive and realistic representation of cognitive processes.

This paper makes the following key contributions. *First*, we develop a unified framework for modeling higher-order general-response cognitive diagnostic models (HO-GRCDMs). We formulate the bottom layer (data layer) of HO-GRCDMs using flexible exponential family distributions. This allows the model to directly adapt to different types of responses (binary, continuous, count, etc.) by altering the parametric family and various types of measurement assumptions (main-effect, all-effect, DINA, etc.) by modifying the latent covariate vector. In the higher-order layer, we employ a probit model to describe the relationship between the higher-order continuous latent traits and the binary latent attributes. The higher-order modeling approach was originally proposed by de la Torre and Douglas (2004) for binary response data, referred to as the higher-order CDM. We generalize it to general response types and employ a probit link instead of the logit link used in de la Torre and Douglas (2004). As will be discussed later, using a probit link for the higher-order layer provides significant theoretical and computational advantages.

Second, we establish identifiability for the proposed HO-GRCDMs. Model identifiability is a crucial prerequisite for valid statistical estimation, but it is a challenging issue for complex latent variable models such as HO-GRCDMs. While the identifiability of single-layer exploratory CDMs for categorical data has been extensively studied (e.g., Xu and Shang, 2018; Culpepper, 2019; Chen et al., 2020), much less is known about the identifiability of CDMs with higher-order structures. Lee and Gu (2024b) provides identifiability results for general response CDMs with a known Q-matrix, but that also does not guarantee the identifiability of an HO-GRCDM. Some existing studies established identifiability for CDMs with higher-order *discrete* latent structures, including the Bayesian pyramid model in Gu and Dunson (2023) and the DeepCDM in Gu (2024). However, the identifiability issue of CDMs with higher-order continuous latent traits is still underexplored, despite these models' popularity (de la Torre and Douglas, 2004; Templin et al., 2008; Wang et al., 2018). To our best knowledge, our identifiability results are the first to identify CDMs with multidimensional higher-order continuous latent traits.

Third, we propose an efficient Monte Carlo Expectation-Maximization (MCEM) algorithm for estimating HO-GRCDMs. The computational challenge of HO-GRCDM arise from three aspects: (a) the various types of response data, (b) the complex hierarchical structure that consists of both binary and continuous latent variables, (c) and the unknown Q-matrix. A typically approach to parameter estimation of such models is to regard the latent variables as missing data and employ Markov chain Monte Carlo (MCMC; Robert and Casella, 2004) or an EM-type algorithm, where the latter method is usually faster. However, the complex hierarchical structure in our model makes the maximization of the complete data log-likelihood intractable during the typical EM updates. In this paper, we propose an MCEM algorithm to maximize the regularized maximum likelihood to simultaneously estimate the Q-matrix and other parameters. Similar to Chen et al. (2015), we consider the Q-matrix estimation as a latent variable selection problem, and maximize log-likelihood with L_1 penalty. Our method provides the first non-MCMC method for parameter estimation in the category of CDMs with a multidimensional higher-order structure.

Our estimation framework incorporates an efficient direct sampling scheme and features significantly reduced simulated samples. The continuous latent traits θ in the higher-order layer are the only latent variables whose realization need to be sampled. After imputing the missing data θ, the M-step optimization becomes more tractable, and we solve this by the cyclical coordinate ascent (Friedman et al., 2010; Tay et al., 2023). Here, the simulation of θ has been a crucial issue in both IRT and higher-order CDM estimation. The MCMC method, commonly used for this purpose, often suffers from slow convergence and requires careful tuning of algorithm parameters. In this paper, we highlight that, benefiting from the use of a probit link for the higher-order layer, we can directly simulate θ from a unified skew-normal distribution according to the recent theoretical results of Li et al. (2023). Last but not least, initialization is an important issue for the efficiency of an algorithm. We employ an efficient Singular Value Decomposition (SVD)-based method for finding initial values, which is an extension of Chen et al. (2019) and Zhang et al. (2020) from the binary response case to the general response case. This non-iterative method is computationally fast and enjoys statistical consistency guarantees under certain conditions (Zhang et al., 2020).

The rest of this paper is organized as follows. Section 2 introduces the framework of HO-GRCDM. Section 3 presents the identifiability results for HO-GRCDM. Section 4 develops an efficient MCEM algorithm for parameter estimation. Section 5 conducts extensive simulation studies for HO-GRCDM under various measurement models, higher-order structures, and response types. Section 6 applies our methodology to a response time data set extracted from the TIMSS 2019 math assessment. Section 7 concludes the paper and discusses future research directions.

2 Model Setup

Assume there are N examinees responding to a test with J items. For each examinee, the observed response vector $\mathbf{R} = (R^1, \ldots, R^J)$ is a J-dimensional vector. Depending on the assessment design, the response R^j could be binary, polytomous, counted-valued, continuous, and so on. We represent an examinee's latent skill profile as a K-dimensional random vector, $\mathbf{A} = (A_1, \ldots, A_K)^\top \in \{0,1\}^K$, and let $\boldsymbol{\alpha} \in \{0,1\}^K$ be an arbitrary binary vector. Let $\mathbb{P}^j_\nu$ denote the prespecified parametric family for the jth response, with parameter $\nu = \nu_{j,\boldsymbol{\alpha}}$. Before introducing the specific notations, we first present the general form of HO-GRCDM as,

$$\begin{aligned} R^j \mid \mathbf{A} = \boldsymbol{\alpha}, \nu &\sim \mathbb{P}^j_{\nu_{j,\boldsymbol{\alpha}}}, \quad \text{where } \nu_{j,\alpha} = [\beta^j]^\top h(\alpha),\ j = 1, \ldots, J;\ \boldsymbol{\alpha} \in \{0,1\}^K; \quad (1) \\ P(A_k = 1|\theta; \lambda_{\mathbf{1}}{}^k, \lambda_0^k) &= f^{-1}(\theta^\top \lambda_{\mathbf{1}}{}^k + \lambda_0{}^k), \quad k = 1, \ldots, K; \quad (2) \\ \theta &\sim N(0, \Sigma_\theta). \quad (3) \end{aligned}$$

Equations (1)-(3) describe a CDM with a three-layer hierarchical structure, with the observed general-response data $\mathbf{R}$ at the bottom layer, the binary latent attributes $\mathbf{A}$ at the middle layer, and the higher-order Normal latent variables $\theta \in \mathbb{R}^D$ at the top layer.

A CDM typically consists of two main components: the measurement part and the latent part. The measurement part, defined in (1), describes how the observed responses $\mathbf{R}$ depend on the latent attributes $\boldsymbol{\alpha}$. We consider a very broad framework that allows flexible response types and various measurement models. The latent part, defined in (45) and (3), models the binary attributes. In the following sections, we separately define each part of the HO-GRCDM. For notational simplicity, for a positive integer M, let $[M]$ be the set of all positive integers $\{1, \ldots, M\}$.

2.1 Bottom Data Layer: CDMs with General Responses (GR-CDMs)

In the bottom layer (1), an examinee's observed responses depend on his/her statuses of K binary latent attributes. Here $A_k = 1$ indicates the k-th attribute is mastered; otherwise, $A_k = 0$. The $h(\alpha)$ is a latent covariate vector consisting of certain main effects and interaction effects of attributes, and β^j is a parameter vector that we will further specify. The parameter $\nu_{j,\alpha} = [\beta^j]^\top h(\alpha)$ is a linear combination of β^j and $h(\alpha)$. For convenience of presentation, we assume that for all items $j = 1, \dots, J$, the $\mathbb{P}^j_{\nu_{j,\alpha}}$ are the same parametric family, and omit the superscript.

We first elaborate on the choice of the parametric family $\mathbb{P}_\nu$ that can be used to model each response type. Below, we denote $g(\cdot;\ \nu)$ as the probability mass/density function (pmf/pdf) of the parametric family under consideration. The Bernoulli distribution with mean ν can be used to model binary responses (de la Torre, 2011; Maris, 1999). Alternatively, one can assume that ν is the logit transform of the Bernoulli mean, as for the case of the Logistic Linear Model (LLM, Maris, 1999). Equivalently, we model the Bernoulli mean as logistic$(\nu) := 1/(1+\exp(-\nu))$:

$$P(R^j = r|\mathbf{A} = \boldsymbol{\alpha}, \nu_{j,\boldsymbol{\alpha}}) = g(r;\ \text{logistic}(\nu_{j,\boldsymbol{\alpha}})) \quad = \quad \frac{\exp\left(-\nu_{j,\alpha}\right)^{1-r}}{1+\exp\left(-\nu_{j,\alpha}\right)}, \quad r = 0, 1. \tag{4}$$

In this work, we allow very diverse choices of non-categorical responses such as count and continuous data. For count-valued data, the Poisson distribution with mean ν can be used:

$$P(R^j = r|\mathbf{A} = \boldsymbol{\alpha}, \nu_{j,\boldsymbol{\alpha}}) = g(r;\ \nu_{j,\boldsymbol{\alpha}}) = \frac{(\nu_{j,\alpha})^r}{r!} \exp\left(-\nu_{j,\alpha}\right), \quad r = 0, 1, 2, \dots \tag{5}$$

For unbounded continuous responses, one can use a normal distribution, where $\nu_{j,\alpha}$ denotes the mean parameter. Together with an additional variance parameter σ_j^2, we have

$$g(r;\ \nu_{j,\boldsymbol{\alpha}}) \quad = \quad \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{\left(r-\nu_{j,\alpha}\right)^2}{2\sigma_j^2}\right\}, \tag{6}$$

For continuous responses with a constrained range, one can use transformed-normal distributions.

For example, the log-normal distribution can be used for modeling positive responses:

$$g(r;\ \nu_{j,\boldsymbol{\alpha}}) \;=\; \frac{1}{\sqrt{2\pi\sigma_j^2}r}\exp\left\{-\frac{\left(\log(r)-\nu_{j,\alpha}\right)^2}{2\sigma_j^2}\right\}, \tag{7}$$

and the logistic-normal distribution for responses within the range of (0,1):

$$g(r;\ \nu_{j,\boldsymbol{\alpha}}) \;=\; \frac{1}{\sqrt{2\pi\sigma_j^2}r(1-r)}\exp\left\{-\frac{\left(\log(r/(1-r))-\nu_{j,\alpha}\right)^2}{2\sigma_j^2}\right\}. \tag{8}$$

Alternatively, to model positive continuous responses, we can also use the Gamma distribution:

$$g(r;\ \nu_{j,\boldsymbol{\alpha}}) \;=\; \frac{\nu_{j,\alpha}^{s_j}}{\Gamma(s_j)}r^{s_j-1}\exp\left\{-\nu_{j,\alpha}\cdot r\right\}, \quad j=1,\cdots,J \tag{9}$$

Here, $\nu_{j,\alpha}>0$ is the rate parameter, and $s_j>0$ is the shape parameter.

Next, we specify the assumptions on the parameter $\nu_{j,\boldsymbol{\alpha}}$. In the literature of cognitive diagnostic modeling, a common assumption involves a pre-specified Q-matrix (Tatsuoka, 1983), $Q=[q_{jk}]_{J\times K}$, that describes which of the K attributes are measured by each of the J items. If $q_{jk}=1$, then the k-th attribute is measured by the j-th item; otherwise, $q_{jk}=0$. So, the value of $\nu_{j,\boldsymbol{\alpha}}$ must depend only on the attributes specified by the j-th row of the Q-matrix. While the Q-matrix is often assumed to be provided along the data, we consider a more challenging *exploratory estimation* scenario where the Q is unknown and need to be estimated along with other parameters.

Given the Q-matrix, one needs some structural assumptions on how $\nu_{j,\alpha}=[\beta^j]^\top h(\alpha)$ depends on the Q-matrix entries. Here, we present three popular measurement model assumptions on the parameters β^j and the function $h(\cdot)$. First, the *main-effect GR-CDM* assumes that $\beta^j=\left(\beta_0^j,\beta_1^j,\cdots,\beta_K^j\right)^\top$, and $h(\alpha)=(1,\alpha_1,\cdots,\alpha_K)^\top$. Then, we can write Eq. (1) as

$$P(R^j|\mathbf{A}=\alpha,\beta^j) \;=\; g\left(R^j;\ \beta_0^j+\sum_{k=1}^K\beta_k^j q_{j,k}\alpha_k\right). \tag{10}$$

For illustration, suppose that we are considering binary responses and $\mathbb{P}_\nu$ is the Bernoulli distribution, then (10) becomes the Additive Cognitive Diagnosis Model (ACDM; de la Torre).

Next, the *all-effect GR-CDM* assumes $\beta^j = \left(\beta_0^j, \beta_1^j, \cdots, \beta_K^j, \beta_{1,2}^j, \cdots, \beta_{K-1,K}^j, \cdots, \beta_{1,2,\dots,K}^j\right)^\top$, and $h(\alpha) = \left(1, \alpha_1, \cdots, \alpha_K, \alpha_1\alpha_2, \cdots, \alpha_{K-1}\alpha_K, \cdots, \prod_{k=1}^K \alpha_k\right)^\top$. The pmf/pdf becomes

$$\begin{aligned} P(R^j|\mathbf{A} = \alpha, \beta^j) &= g\Bigg(\mathbf{R}^j;\ \beta_0^j + \sum_{k=1}^K \beta_k^j q_{j,k}\alpha_k \\ &+ \sum_{1\leq k_1\leq k_2} \beta_{k_1k_2}^j \left\{q_{j,k_1}\alpha_{k_1}\right\}\left\{q_{j,k_2}\alpha_{k_2}\right\} + \cdots + \beta_{1,2,\dots,K}^j \prod_{k=1}^K \left\{q_{j,k}\alpha_k\right\}\Bigg). \quad (11) \end{aligned}$$

Note that for the case of binary responses, this becomes the GDINA model (de la Torre, 2011).

Finally, we introduce the GR-DINA model. We borrow the notations from the all-effect models, and take $\beta^j = \left(\beta_0^j,\ \beta_{\mathscr{K}_j}^j\right)^\top$ and $h(\alpha) = \left(1,\ \prod_{k\in\mathscr{K}_j}(q_{j,k}\alpha_k)\right)^\top$. Here, $\mathscr{K}_j = \left\{k \in [K]:\ q_{jk} = 1\right\}$ denotes the set of attributes that are measured by item j. Then the pmf/pdf is written as

$$P(R^j|\mathbf{A} = \alpha, \beta^j) = g\left(R^j;\ \beta_0^j + \beta_{\mathscr{K}_j}^j \prod_{k\in\mathscr{K}_j} (q_{j,k}\alpha_k)\right). \quad (12)$$

Note that the above formulation can be regarded as a special case of all-effect GR-CDM. Under binary responses, this model is the popular Deterministic Input, Noisy "And" gate model (DINA; Junker and Sijtsma, 2001) model with the conjunctive assumption. Under positive continuous responses, (12) becomes the continuous DINA model (c-DINA; Minchen et al., 2017).

For main-effect and all-effect GR-CDMs, the Q-matrix should constrain certain β-coefficients to be zero. For example, under the main-effect model, we must have $\beta_k^j = 0$ for k for which $q_{jk} = 0$. Under the all-effect model, we must have $\beta_S^j = 0$ when $S \not\subseteq \mathscr{K}_j$. Since we consider an unknown Q-matrix, the index of such zero coefficients is also unknown and needs to be estimated from data.

2.2 Latent Layers: Higher-Order Latent Trait Model for Binary Attributes

As shown in Equations (45) and (3), we consider a higher-order latent layer to model the attributes α. We introduce a D-dimensional *continuous* latent trait, $\theta = (\theta_1, \dots, \theta_D)^\top$. There are two common choices for the invertible link function f: the logit link function $f(x) = \log(x/1-x)$, and the probit link function $f(x) = \Phi^{-1}(x)$, where Φ is the cumulative distribution function of a standard normal random variable. In this paper, we employ the probit link function and let

$$P(A_k = 1|\theta, \boldsymbol{\lambda_1}^k, \lambda_0^k) \quad = \quad \Phi(\theta^\top \boldsymbol{\lambda_1}^k + \lambda_0^k), \qquad k = 1, 2, \ldots, K. \tag{13}$$

Here, $\boldsymbol{\lambda_1}^k = (\lambda_{1,1}^k, \ldots, \lambda_{1,D}^k)^\top$ and λ_0^k are the slopes and the intercept, respectively. Let $\boldsymbol{\lambda_1} = (\boldsymbol{\lambda_1}^1, \ldots, \boldsymbol{\lambda_1}^K)^\top$ be a matrix consisting of slope parameters of all K attributes, and $\boldsymbol{\lambda_0} = (\lambda_0^1, \ldots, \lambda_0^K)^\top$ be a vector consisting of intercept parameters of all K attributes. We assume a pre-specified binary matrix $Q^{(H)} = [q_{kd}^{(H)}]_{K \times D}$, which constrains the sparsity structure of $\boldsymbol{\lambda}_1$ to enhance the interpretability. The entry $q_{kd}^{(H)} = 1$ implies that the d-th continuous latent variable θ_d contributes to the mastering of the k-th binary latent attribute A_k. In Equation (3), θ is assumed to follow a D-variate normal distribution, $\theta \sim N(\mathbf{0}_D, \boldsymbol{\Sigma}_\theta)$, where the zero mean vector $\mathbf{0}_D$ is to fix the measurement origin, and the covariance matrix $\boldsymbol{\Sigma}_\theta = (\sigma_{dd'})$ has unit diagonal entries to fix the measurement units. That is, $\sigma_{dd} = 1$, for $d = 1, \ldots, D$. We also assume that the first item loading on each factor is positive to resolve the sign indeterminacy issue of the latent factors.

The higher-order latent layer is introduced to resemble an item response model (Lord, 1952; Birnbaum, 1968). This modeling approach was initially used in de la Torre and Douglas (2004), where the authors considered binary responses and assumed both the bottom layer's Q matrix and the latent layer's $Q^{(H)}$ matrix were known. In addition to the higher-order CDM, another approach introduced by Templin et al. (2008) employed a multivariate probit model with a *single* continuous latent factor. In that work, each binary attribute is derived by dichotomizing a Normal random variable at a specific threshold, with the K Normal variables modeled using a factor analysis structure.

Intuitively, the higher-order latent structure should be more parsimonious than the bottom layer, as it generally represents more abstract factors/traits in a higher level. For this aim, a *subscale structure* for $Q^{(H)}$ may be appropriate, where each attribute is associated with only one latent trait among all the D traits. In other words, an attribute A_k exclusively depends on a latent trait θ_d, and we have $q_{kd}^{(H)} = 1$. For all other groups indexed by $d' \neq d$, we assume $q_{kd'}^{(H)} = 0$ and $\lambda_{1,d'}^k = 0$ to not include their effect on α_k. However, in some cases, this structure may be too simple to capture the relationship between $\mathbf{A}$ and θ. To address this, the *bifactor structure* with an additional general factor could be a more flexible alternative yet still being parsimonious. In the bifactor structure, we assume that there are $D-1$ groups (indexed by $d=1,\dots,D-1$), and that each binary attribute is assigned to exactly one group. Each attribute α_k that belongs in group d is influenced by two latent factors: a general factor θ_1, and a group-specific factor θ_{d+1}. Consequently, we have $q_{k1} = q^{(H)}_{k(d+1)} = 1$. The effects of the other groups $d' \neq d$ are assumed to be zero, that is $q^{(H)}_{k(d'+1)} = 0$ and $\lambda^k_{1,d'+1} = 0$. In the remainder of the paper, we will assume that the higher-order model follows either the subscale or bifactor structure with a known $Q^{(H)}$ matrix/group information.

2.3 Benefits of the Probit Link for Modeling the Higher-order Layer

We elaborate on our rationale for choosing the probit link to model the higher-order layer, as opposed to the more common logit link. There are mainly two advantages in doing so. To see this, we first present the marginal distribution of the binary random vector $\mathbf{A}$ obtained using the probit link. For each $\boldsymbol{\alpha} \in \{0,1\}^K$, the marginal probability of $\mathbf{A} = \alpha$ under our model is

$$
\begin{aligned}
\pi_{\boldsymbol{\alpha}} := P(\mathbf{A}=\alpha) &= P\left(\varepsilon_1 \le (-1)^{(\alpha_1+1)}(\lambda_0^1 + \boldsymbol{\lambda}_{\mathbf{1}}{}^{1\top}\theta), \dots, \varepsilon_K \le (-1)^{(\alpha_K+1)}(\lambda_0^K + \boldsymbol{\lambda}_{\mathbf{1}}{}^{K\top}\theta)\right) \\
&= P\left((-1)^{(\alpha_1)}\lambda_0^1 + \sqrt{\boldsymbol{\lambda}_1^{1\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^1}\,\xi_1 \le 0, \dots, (-1)^{(\alpha_K)}\lambda_0^K + \sqrt{\boldsymbol{\lambda}_1^{K\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^K}\,\xi_K \le 0\right) \\
&= \Phi_K\left(\frac{(-1)^{(\alpha_1+1)}\lambda_0^1}{\sqrt{\boldsymbol{\lambda}_1^{1\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^1+1}}, \dots, \frac{(-1)^{(\alpha_K+1)}\lambda_0^K}{\sqrt{\boldsymbol{\lambda}_1^{K\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^K+1}},\ \mathbf{C}_\rho\right),
\end{aligned}
\tag{14}
$$

with a tetrachoric correlation matrix $\mathbf{C}_\rho = (C_\rho(k_1,k_2))_{K\times K}$:

$$
C_\rho(k_1,k_2) = \frac{(-1)^{(\alpha_{k_1}+\alpha_{k_2})}\boldsymbol{\lambda}_1^{k_1\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^{k_2}}{\sqrt{\boldsymbol{\lambda}_1^{k_1\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^{k_1}+1}\sqrt{\boldsymbol{\lambda}_1^{k_2\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^{k_2}+1}}
$$

for $k_1 \neq k_2 \in [K]$, and $C_\rho(k,k) = 1$ for $k \in [K]$. Here, ξ_k and ε_k are independent and identically distributed standard normal random variables, and Φ_K represents the cumulative distribution function (CDF) of a K-variate normal distribution; see more details in Fang et al. (2021).

The first virtue of using the probit link is for establishing model identifiability. Using (14), Fang et al. (2021) proved that the probit model's identifiability boils down to identifying the parameters $(\lambda_0^k, \boldsymbol{\lambda}_1^k)$ based on the threshold values $\lambda_0^1/(\boldsymbol{\lambda}_1^{1\top}\boldsymbol{\Sigma}_\theta\boldsymbol{\lambda}_1^1+1)^{1/2}$, and the pairwise tetrachoric correlations. This means that identifiability is reduced to a problem similar to the identifiability of linear factor models. In contrast, this property does not exist when using a logit link, which involves a complex convolution of Gaussian and logistic random variables.

The second advantage of using a probit link is on computation. First, the explicit form of the marginal distribution of $\mathbf{A}$ in (14) significantly reduces the computational complexity of parameter estimation. When using an EM-type algorithm, once the conditional expectation $E_\theta[l(\theta,\mathbf{A},\mathbf{R})|\mathbf{A},\mathbf{R}]$, where $l(\theta,\mathbf{A},\mathbf{R})$ is the log-likelihood, is computed, the whole conditional expectation $E_{\mathbf{A}}[E_\theta[l(\theta,\mathbf{A},\mathbf{R})|\mathbf{A},\mathbf{R}]|\mathbf{R}]$ in the E-step is available, as the out-layer expectation can be easily computed according to Equation (14). Second, utilizing the probit link enables the development of an efficient sampling scheme, as it allows direct sampling of θ from the target distribution $\theta|\mathbf{A},\mathbf{R},\lambda_{\mathbf{1}},\lambda_{\mathbf{0}}$. As detailed in Section 4.2.1, $\theta|\mathbf{A},\mathbf{R},\lambda_{\mathbf{1}},\lambda_{\mathbf{0}}$ follows a unified skew-normal distribution, whose samples can be obtained by a direct combination of samples from truncated normal and multivariate normal distributions.

3 Identifiability

We next propose conditions that guarantee the identifiability of HO-GRCDMs. For a statistical model $\{P_\nu\}$ indexed by a set of parameters ν, we say that the model is identifiable at the true parameter ν_0 when the equality between marginal distribution of the observed variables $P_\nu = P_{\nu_0}$ implies equal parameter values $\nu = \nu_0$. Identifiability is a fundamental prerequisite for consistent parameter estimation and valid model interpretation.

We first present separate identifiability conditions for (a) the bottom layer *exploratory* GR-CDM and (b) the higher-order continuous latent layer *as if the binary attributes were observed binary responses*. Then, we combine these results using a layer-wise proof argument similar to that in Gu (2024) and derive identifiability conditions for HO-GRCDMs. More specifically, we first marginalize out the top continuous layer and identify the GR-CDM parameters, including the proportion parameters $\boldsymbol{\pi} = (\pi_{\boldsymbol{\alpha}};\ \boldsymbol{\alpha} \in \{0,1\}^K)$ describing the marginal distribution of the binary attributes. Next, we use the estimated GR-CDM proportion parameters $\boldsymbol{\pi}$ and (14) to identify the parameters in the higher-order probit model. We define saturated GR-CDMs as follows.

Definition 1 (Saturated GR-CDM)**.** *A saturated GR-CDM with parameters* $(\boldsymbol{\pi}, \boldsymbol{\beta}, Q)$ *is a CDM without an higher-order structure, defined by* (1) *and with proportion parameters* $\pi_{\boldsymbol{\alpha}}$ *for each binary attribute pattern* $\boldsymbol{\alpha} \in \{0,1\}^K$*. Here,*$\boldsymbol{\pi} = (\pi_{\boldsymbol{\alpha}})$ *satisfy* $\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \pi_{\boldsymbol{\alpha}} = 1$*, and* $\pi_{\boldsymbol{\alpha}} > 0$.

We impose a monotonicity condition on the item parameters to avoid the sign-flipping for each latent attribute; that is, to distinguish between $A_k = 0$ and 1. Motivated by the popular monotonicity conditions for binary-response CDMs (Xu and Shang, 2018), we assume that

$$\nu_{j,\boldsymbol{\alpha}} > \nu_{j,\boldsymbol{\alpha}'} \quad \text{for all} \quad j \in [J],\ \boldsymbol{\alpha} \succeq \mathbf{q}_j,\ \boldsymbol{\alpha}' \not\succeq \mathbf{q}_j \tag{15}$$

holds for all saturated GR-CDMs, and also for all HO-GRCDMs. Here, $\mathbf{q}_j$ is the j-th row of Q. The assumption (15) means that the students possessing all required skills for the jth item have a larger ν-parameter in the distribution $R^j \mid \mathbf{A}$ than those who lack some required skills. This condition can be further simplified under specific GR-CDMs. For example, (15) is equivalent to $\beta^j_{\mathscr{K}_j} > 0$ under the DINA model, and to $\beta^j_k > 0$ for $q_{j,k} = 1$ under the ACDM. The following result is from Proposition 1 in Lee and Gu (2024a).

Proposition 1. *Under the saturated DINA/main-effect/all-effect GR-CDM that satisfies the monotonicity condition* (15)*, the model components* $(\boldsymbol{\pi}, \boldsymbol{\beta}, Q)$ *are identifiable up to a permutation of the K latent attributes when the following conditions hold.*

A. The true Q-matrix contains two submatrices $\mathbf{I}_K$ *after row swapping, i.e., Q can be written as*

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

B. Suppose that the Q-matrix is written as in A. For any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$*, there exists* $j > 2K$ *such that* $\nu_{j,\boldsymbol{\alpha}} \neq \nu_{j,\boldsymbol{\alpha}'}$.

In particular, condition B holds when the Q-matrix also contains another identity submatrix $\mathbf{I}_K$.

Conditions A and B resemble popular identifiability conditions for CDMs with categorical responses (Xu and Shang, 2018; Culpepper, 2019). Lee and Gu (2024b) showed that these conditions also suffice for identifying GR-CDMs, but under the confirmatory setting with a known Q-matrix.

Next, we present identifiability conditions for the higher-order probit layer in HO-GRCDMs. Here, we view the probit model as a parametric family with parameters $(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\Sigma}_\theta)$ and probability mass function in (14). Recall that we consider the subscale model and the bifactor model with a known higher-order loading structure $Q^{(H)}$ to model the K binary attributes.

We first present the necessary and sufficient conditions for the subscale model, which is proved by heavily utilizing the properties of the probit link mentioned after (14). This identifiability result may be of independent interest in the item response theory literature. We present the proofs of all theoretical results in the Supplementary Material.

Proposition 2. *Consider the subscale model with K attributes $(A_1, \ldots, A_K)$ and D-dimensional Gaussian latent factors θ_D. For $d \in [D]$, let $K_d = \sum_{k=1}^K q_{k,d}^{(H)}$ denote the number of attributes that belong to group d. Then, the model is identifiable if and only if one of the below conditions holds for all $d \in [D]$:*

C1. $K_d \geq 3$,

C2. $K_d = 2$, $\sigma_{dd'} \neq 0$ for some $d \neq d'$.

To summarize the above conditions C1 and C2, the subscale model with three or more attributes for each group is identifiable. Interestingly, to ensure identifiability, we require at least two attributes to belong to each group, which boils down to assuming condition A on the $Q^{(H)}$-matrix.

Next, we present identifiability conditions for the bifactor model, where the D-dimensional latent vector $\boldsymbol{\theta}$ consists of one general factor (θ_1) and $D-1$ group effects $(\theta_2, \ldots, \theta_D)$. Here, it is necessary to assume an additional orthogonal structure between the general and group-specific factors to resolve a trivial rotational ambiguity (see Section 4.1 in Fang et al., 2021). For each $d \in [D-1]$, let $L_d \subseteq [K]$ be the collection of attributes that belong to the dth group. We also let $\lambda_{1,d'}^{L_d}$ be a sub-vector of $\lambda_{1,d'} = (\lambda_{1,d'}^1, \ldots, \lambda_{1,d'}^K)^\top$ that consists of the indices $k \in L_d$.

Proposition 3 (Theorem 7 in Fang et al. (2021)). *Consider the bifactor model with K binary attributes and D-dimensional Gaussian latent variables with covariance $\Sigma_\theta = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \Sigma_\theta^\star \end{pmatrix}$, where $\Sigma_\theta^\star$ is a $(D-1)\times(D-1)$ symmetric positive definite matrix with $Diag(\Sigma_\theta^\star) = I_{D-1}$. Let $H := \{d \in [D-1] : \boldsymbol{\lambda}_{1,1}^{L_d}$ and $\boldsymbol{\lambda}_{1,d+1}^{L_d}$ are linear independent$\}$. Then, the model is identifiable if*

$$|\{k \in L_d : \lambda_{1,d+1}^k \neq 0\}| \geq 3 \ \textit{for all} \ \ d \in [D-1], \tag{16}$$

and one of the following conditions hold:

C3. $|H| \geq 3$,

C4. $|H| = 2$*, and there exists a group d such that L_d can be partitioned into $L_{d,1}$ and $L_{d,2}$ so that $\boldsymbol{\lambda}_{1,1}^{L_{d,a}}$ and $\boldsymbol{\lambda}_{1,d+1}^{L_{d,a}}$ are linearly independent for both partitions $a = 1,2$.*

Combining the above results, we can establish the desired identifiability result for HO-GRCDMs. The main argument is to sequentially identify the latent layers, similar to previous works for multilayer variants of CDMs (Gu and Dunson, 2023; Gu, 2024).

To resolve the latent attribute permutation issue in Proposition 1, we additionally assume that there is a pre-specified *anchor item* for each latent attribute. This means for each latent attribute A_k, we know that some item $j_k \in [J]$ measures A_k and does not measure any other attribute. Without loss of generality, combined with condition A, we can assume that the first K items are the anchor items for the K binary attributes.

Theorem 1. *The following two identifiability conclusions hold.*

(a) Consider an HO-GRCDM with a ***subscale*** *higher-order layer and known anchor items. The model is identifiable when the true bottom layer GRCDM parameters Q, β satisfy conditions A and B, and the true latent layer parameters $Q^{(H)}, \Sigma_\theta$ satisfy either condition C1 or C2.*

(b) Next, consider an HO-GRCDM with a ***bifactor*** *higher-order layer and known anchor items. The model parameters are identifiable when the true parameters satisfy conditions A, B, (16), and either C3 or C4.*

Theorem 1 provides our main result that guarantees the HO-GRCDMs are identifiable. Note that even with the anchor item assumption, we are still identifying all other $J-K$ rows of the Q-matrix. Note that by considering binary responses in the CDM, this result can be applied to establish identifiability of the celebrated HO-LTM with a probit link.

4 A New MCEM Algorithm

In this section, we develop an MCEM algorithm to estimate HO-GRCDMs. We first define some notations. Suppose there are N subjects response to a test consisting of J items. Let $i = 1,\dots,N$ and $j = 1,\dots,J$ denote the index of subjects and items, respectively. The test is designed to measure K binary latent attributes for each subject i, $\mathbf{A}_i = (A_{i1},\dots,A_{iK})^\top$, which is further determined by D higher-order continuous latent variable, $\theta_i = (\theta_{i1},\dots,\theta_{iD})^\top$. We slightly abuse notation and use $\mathbf{R} = [R_{ij}]_{N\times J} = (\mathbf{R}_1,\dots,\mathbf{R}_N)^\top$ to denote the observed response matrix. Let $\boldsymbol{\Theta} = (\theta_1,\dots,\theta_N)^\top$ be the $N \times D$ matrix that collects continuous latent variables for the N subjects, and define $\mathbf{A} = (\mathbf{A}_1,\dots,\mathbf{A}_N)^\top$ as the $N \times K$ matrix consisting of the binary attribute profiles for all N subjects.

Let β and $\boldsymbol{\lambda}$ denote the set of all the coefficient parameters in the first (1) and second layer (45), respectively. Additionally, we aim to estimate the Q-matrix in the CDM layer and the covariance matrix $\boldsymbol{\Sigma}_\theta$ for the continuous latent traits (3). Note that estimating Q by directly maximize the marginalized log-likelihood is computationally infeasible even for a moderate size of J and K. This is because one need to compute the profile likelihood based on each Q among $2^{J\times K}$ possible matrices, and find out the one that maximizes the profile likelihood. One solution to avoid such expensive computation is to consider the estimation of Q as a latent variable selection problem and solving it through a regularized maximum likelihood estimator (Chen et al., 2015). In particular, we maximize the regularized marginal log-likelihood $l(\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta|\mathbf{R})$ with an L_1 penalty $p_s(\beta)$:

$$(\widehat{\beta},\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\Sigma}}_\theta,\widehat{Q}) = \underset{\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta}{\arg\max}\,\left(l(\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta|\mathbf{R}) - N\cdot p_{\mathbf{s}}(\beta)\right), \tag{17}$$

where the log-likelihood function can be written as

$$l(\beta, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_\theta, Q|\mathbf{R}) \;=\; \sum_{i=1}^{N} \log \left\{ \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \prod_{j=1}^{J} P(R_{ij}|\boldsymbol{\alpha}; \beta^j) \int P(\boldsymbol{\alpha}|\theta; \boldsymbol{\lambda}) f(\theta|\boldsymbol{\Sigma}_\theta) \mathrm{d}\theta \right\}, \quad (18)$$

and the penalty function is defined as

$$p_{\mathbf{s}}(\beta) = \sum_{j=1}^{J} p_s(\beta^j) = \sum_{j=1}^{J} \left(s \sum_{\beta_k^j \in \beta^j} |\beta_k^j| \right), \quad (19)$$

where s is the regularization parameter. Here, the Q-matrix does not need to appear explicitly in the log-likelihood expression because its information is implicitly captured by the sparsity of the coefficients. After solving (17), Q can be estimated by identifying the non-zero pattern of $\widehat{\beta}$.

The mechanism for identifying the entries q_{jk} in the Q-matrix varies across different measurement models. For *main-effect* models, Q can be recovered using the rule $q_{jk} = \mathbf{1}(\beta_k^j \neq 0)$, where $\mathbf{1}$ is the indicator function. For *all-effect* models, theoretically, each row of Q should be identified by the highest-order non-zero coefficient. Specifically, if $\exists S \subseteq [K]$ such that $\beta_S^j \neq 0$ and $\beta_{S'}^j = 0$ for all $S' \subseteq [K], S \subset S'$, then $q_{jk} = 1$ for $k \in S$; otherwise, $q_{jk} = 0$. However, this strict rule may not be always applicable because some estimated β-coefficients may be close to zero but not exactly zero. In practice, a more effective approach is either to choose the largest non-zero interaction coefficient or to truncate the coefficients before identifying Q. For the latter approach, we recommend practitioners set the truncation thresholds based on the general magnitude of their estimated coefficients. For the *DINA model*, since there should be only one non-zero coefficient for each item j, the largest non-zero interaction coefficient can be selected, and the corresponding q_{jk} can be identified as equal to one.

The maximization problem presented in Equation (17) is quite complex due to the summation of integrals inside the log function and cannot be solved directly. The Expectation-Maximization (EM) algorithm is a popular method that iterates between the E-step and the M-step to seek the maximizer, and we first introduce the penalized variant of the EM algorithm in Section 4.1. This basic EM algorithm still suffers from the intractable integrals in the E-step, which motivates us to propose a more scalable novel Monte Carlo EM algorithm in Section 4.2.

4.1 Penalized EM algorithm

We first introduce the basic procedure of a penalized EM algorithm. A usual EM algorithm alternates between two steps: in the E-step, the expected complete data log-likelihood is computed, and in the M-step, the parameters are updated by maximizing this expected log-likelihood. The penalized EM algorithm follows the same procedure but includes a penalty term in the M-step to regularize the parameter estimates.

The complete data log-likelihood in a HO-GRCDM is

$$
\begin{aligned}
l(\beta, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_\theta | \mathbf{R}, \mathbf{A}, \boldsymbol{\Theta}) &= \log\left(\prod_{i=1}^{N}\prod_{j=1}^{J} P(R_{ij}|\mathbf{A}_i;\beta^j) P(\mathbf{A}_i|\theta_i;\boldsymbol{\lambda}) f(\theta_i|\boldsymbol{\Sigma}_\theta)\right) \\
&\triangleq l_1(\beta, \mathbf{R}, \mathbf{A}) + l_2(\boldsymbol{\lambda}, \theta_i, \mathbf{A}) + l_3(\theta_i, \boldsymbol{\Sigma}_\theta),
\end{aligned} \tag{20}
$$

with

$$
\begin{aligned}
l_1(\beta, \mathbf{R}, \mathbf{A}) &= \sum_{i=1}^{N} \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} \mathbf{1}(\mathbf{A}_i = \boldsymbol{\alpha}) \sum_{j=1}^{J} \log\left(P\left(R_{ij}|\mathbf{A}_i = \boldsymbol{\alpha}; \beta^j\right)\right), & (21)\\
l_2(\boldsymbol{\lambda}, \theta_i, \mathbf{A}) &= \sum_{i=1}^{N} \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} \mathbf{1}(\mathbf{A}_i = \alpha) \sum_{k=1}^{K} \left(\alpha_k \log\Phi\left(\theta_i^\top \boldsymbol{\lambda}_\mathbf{1}{}^k + \lambda_0^k\right)\right. \\
&\quad \left. + (1-\alpha_k)\log\left(1 - \Phi\left(\theta_i^\top \boldsymbol{\lambda}_\mathbf{1}{}^k + \lambda_0^k\right)\right)\right), & (22)\\
l_3(\theta_i, \boldsymbol{\Sigma}_\theta) &= \sum_{i=1}^{N}\left(-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\theta_i^\top \boldsymbol{\Sigma}_\theta \theta_i\right). & (23)
\end{aligned}
$$

Let $\eta = (\beta, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_\theta)$ generically denotes the collection of all parameters. For any parameter, a superscript "(t)" denotes the values obtained in the tth iteration. Each iteration t of the penalized EM algorithm contains the following two steps:

E-Step: Compute the Q-function as the expectation of the complete data log-likelihood:

$$
Q^{(t)}(\beta, \boldsymbol{\lambda}, \Sigma_\theta) = E_{(\mathbf{A},\boldsymbol{\Theta})}\left[l(\beta, \boldsymbol{\lambda}, \boldsymbol{\Sigma}_\theta|\mathbf{R}, \mathbf{A}, \boldsymbol{\Theta}) \mid \mathbf{R}; \widehat{\beta}^{(t-1)}, \widehat{\boldsymbol{\lambda}}^{(t-1)}, \widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}\right], \tag{24}
$$

where the conditional expectation is with respect to $P(\mathbf{A},\boldsymbol{\Theta}|\mathbf{R})$. We break down $Q^{(t)}(\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta)$ into three parts,

$$Q^{(t)}(\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta) = Q_1{}^{(t)}(\beta) + Q_2{}^{(t)}(\boldsymbol{\lambda}) + Q_3{}^{(t)}(\boldsymbol{\Sigma}_\theta),$$

with

$$Q_1{}^{(t)}(\beta) = E_{\mathbf{A}}[l_1|\mathbf{R};\widehat{\eta}^{(t-1)}] = \sum_{i=1}^{N}\sum_{\alpha\in\{0,1\}^K}\sum_{j=1}^{J}\log\left(P\left(R_{ij}|\alpha;\beta^j\right)\right)\psi_{i,\alpha}^{(t-1)}, \tag{25}$$

$$Q_2{}^{(t)}(\boldsymbol{\lambda}) = E_{(\mathbf{A},\boldsymbol{\Theta})}[l_2|\mathbf{R},\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}] = \sum_{i=1}^{N}\sum_{\alpha\in\{0,1\}^K} E_{\theta_i}[l_2{}^{(i,\alpha)}|\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_\theta^{(t-1)}]\psi_{i,\alpha}^{(t-1)}, \tag{26}$$

$$Q_3{}^{(t)}(\boldsymbol{\Sigma}_\theta) = E_{(\mathbf{A},\boldsymbol{\Theta})}[l_3|\mathbf{R},\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}] = \sum_{i=1}^{N}\sum_{\alpha\in\{0,1\}^K} E_{\theta_i}[l_3{}^{(i)}|\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_\theta^{(t-1)}]\psi_{i,\alpha}^{(t-1)}. \tag{27}$$

Here the common term $\psi_{i,\alpha}^{(t-1)}$ has the following expression,

$$\begin{aligned}\psi_{i,\alpha}^{(t-1)} =& P(\mathbf{A}_i=\alpha|\mathbf{R}_i;\widehat{\beta}^{(t-1)},\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}) \\ =& \frac{P(\mathbf{R}_i|\mathbf{A}_i=\alpha;\widehat{\beta}^{(t-1)})P(\mathbf{A}_i=\alpha;\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)})}{\sum_{\alpha\in\{0,1\}^K}P(\mathbf{R}_i|\mathbf{A}_i=\alpha;\widehat{\beta}^{(t-1)})P(\mathbf{A}_i=\alpha;\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)})},\end{aligned} \tag{28}$$

and $l_2{}^{(i,\alpha)}$ and $l_3{}^{(i)}$ denote the corresponding terms in the summation indexed by i and α presented in Equations (22) and (23), respectively.

M-Step: Update the parameters by maximizing the penalized Q-function:

$$\widehat{\eta}^{(t)} = (\widehat{\beta}^{(t)},\widehat{\boldsymbol{\lambda}}^{(t)},\widehat{\boldsymbol{\Sigma}}_\theta^{(t)}) = \underset{\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta}{\arg\max}\, Q^{(t)}(\beta,\boldsymbol{\lambda},\boldsymbol{\Sigma}_\theta) - N\cdot p_{\mathbf{s}}(\beta), \tag{29}$$

This is typically a convex optimization for exponential family distributed responses.

4.2 Monte Carlo EM Algorithm with an Efficient Sampling Scheme

4.2.1 Monte Carlo Integration for E-Step

As mentioned earlier, the probit link offers a significant advantage by enabling direct computation of the conditional expectation $E_{\mathbf{A}}[\cdot|\mathbf{R};\widehat{\eta}^{(t-1)}]$ as shown in (25)-(27), due to the explicit form of $P(\mathbf{A}=\alpha;\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)})$ presented in (14). This simplifies the computational challenge of the E-Step to calculating the inner expectations, $E_{\theta_i}[\cdot|\mathbf{A}_i=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_{\theta}^{(t-1)}]$, as involved in (26)-(27). Furthermore, conditional on $\mathbf{A}_i=\alpha$, $\widehat{\boldsymbol{\lambda}}^{(t-1)}$, and $\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}$, θ_i's are independent and identically distributed (i.i.d.). Using a general notation θ to represent θ_i in Equations (26)-(27), we have

$$\begin{aligned}
& E_{\theta_i}\left[{l_2}^{(i,\alpha)}|\mathbf{A}_i=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}\right] \qquad (30)\\
=\ & E_{\theta}\left[\sum_{k=1}^{K}\left(\alpha_k\log\Phi\left(\theta^{\top}{\boldsymbol{\lambda}_{\mathbf{1}}}^{k}+\lambda_0^k\right)+(1-\alpha_k)\log\left(1-\Phi\left(\theta^{\top}{\boldsymbol{\lambda}_{\mathbf{1}}}^{k}+\lambda_0^k\right)\right)\right)\middle|\mathbf{A}_i=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_{\theta}^{(t-1)}\right],
\end{aligned}$$

and

$$E_{\theta_i}\left[{l_3}^{(i)}|\mathbf{A}_i=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}\right]=E_{\theta}\left[-\frac{D}{2}\log(2\pi)-\frac{1}{2}\log|\boldsymbol{\Sigma}_{\theta}|-\frac{1}{2}\theta^{\top}\boldsymbol{\Sigma}_{\theta}^{-1}\theta\right]. \qquad (31)$$

This representation implies that, when computing ${Q_2}^{(t)}$ or ${Q_3}^{(t)}$, only 2^K expectations, $E_{\theta}\left[\cdot|\mathbf{A}_i=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_{\theta}^{(t-1)}\right]$, for $\alpha\in\{0,1\}^K$, need to be evaluated, regardless of the sample size N.

Despite the significantly reduced computational complexity, the above expectations involve multidimensional integrals and cannot be evaluated in closed form. We propose to use Monte Carlo integration and draw M_t samples $\theta^{(m,\alpha)}$, $m=1,\ldots,M_t$ from $P(\theta|\mathbf{A}=\alpha;\widehat{\boldsymbol{\lambda}}^{(t-1)},\boldsymbol{\Sigma}_{\theta}^{(t-1)})$, $\alpha\in\{0,1\}^K$, in the tth iteration. For any function $w(\theta,\alpha)$ of θ and α, we can approximate its expectation as

$$E_{\theta}[w(\theta,\alpha)|\mathbf{A}=\alpha,\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}] \approx \frac{\sum_{m=1}^{M_t}w(\theta^{(m,\alpha)},\alpha)}{M_t}. \qquad (32)$$

By replacing $w(\theta,\alpha)$ with the corresponding terms within the square brackets in Equations (30) and (31), we obtain Monte Carlo approximations for these expectations.

Sampling from $f(\theta|\mathbf{A}=\alpha;\widehat{\boldsymbol{\lambda}}^{(t-1)},\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)})$ has been a challenging issue in the literature of both multidimensional IRT models and higher-order CDMs. A commonly used method is the MCMC method, including the Metropolis–Hastings (MH; Cai, 2010) sampler and the Gibbs sampler (Béguin and Glas, 2001; Culpepper, 2016). However, such methods suffer from slow convergence to the target distribution, thereby slowing down the algorithm. Fortunately, with a probit link used in the latent layers, directly sampling is feasible. According to Theorem 4.2 in Li et al. (2023),

$$\theta|\mathbf{A} = \alpha, \mathbf{R}; \lambda_{\mathbf{1}}, \lambda_{\mathbf{0}} \sim SUN_{D,K}(\mathbf{0}_D, \mathbf{\Sigma}_\theta, \mathbf{\Delta}_c, \gamma_c, \mathbf{\Gamma}_c). \tag{33}$$

Here, SUN denotes an unified skew-normal distribution (Arellano-Valle and Azzalini, 2006), where $\mathbf{\Delta}_c = \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top \mathbf{S}^{-1}$, $\gamma_c = \mathbf{S}^{-1}\mathbf{U}_2$, $\mathbf{\Gamma}_c = \mathbf{S}^{-1}(\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)\mathbf{S}^{-1}$, and

$$\begin{aligned} \mathbf{U}_1 &= diag(2\alpha_1 - 1, \ldots, 2\alpha_K - 1)\lambda_{\mathbf{1}}, \\ \mathbf{U}_2 &= diag(2\alpha_1 - 1, \ldots, 2\alpha_K - 1)\lambda_{\mathbf{0}}, \\ \mathbf{S} &= diag(\mathbf{U}_1^1{}^\top \mathbf{\Sigma}_\theta \mathbf{U}_1^1, \ldots, \mathbf{U}_1^K{}^\top \mathbf{\Sigma}_\theta \mathbf{U}_1^K), \end{aligned} \tag{34}$$

with $\mathbf{U}_1^k{}^\top$ denoting the k-th row of $\mathbf{U}_1$. Furthermore, by Corollary 4.3 in Li et al. (2023), if (33) holds, then the conditional distribution of θ is equal to the following distribution (where "$\overset{\mathrm{d}}{=}$" means "equal in distribution")

$$\theta|\mathbf{A} = \alpha, \mathbf{R}; \lambda_{\mathbf{1}}, \lambda_{\mathbf{0}} \overset{\mathrm{d}}{=} \mathbf{V}_0 + \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top (\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)^{-1} \mathbf{S}\,\mathbf{V}_1, \tag{35}$$

where $\mathbf{V}_0$ is independent of $\mathbf{V}_1$ and

$$\mathbf{V}_0 \sim N(\mathbf{0}_D, \mathbf{\Sigma}_\theta - \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top (\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)^{-1} \mathbf{U}_1 \mathbf{\Sigma}_\theta), \tag{36}$$

$$\mathbf{V}_1 \sim TN(\mathbf{0}_D, \mathbf{S}^{-1}(\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)\mathbf{S}^{-1}, -\infty, -\mathbf{S}^{-1}\mathbf{U}_2). \tag{37}$$

Here, $TN(\mathbf{0}_D, \mathbf{S}^{-1}(\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)^{-1}, -\infty, -\mathbf{S}^{-1}\mathbf{U}_2)$ denotes a K-variate truncated normal distribution with zero mean and covariance matrix $\mathbf{S}^{-1}(\mathbf{U}_1 \mathbf{\Sigma}_\theta \mathbf{U}_1{}^\top + \mathbf{I}_K)^{-1}$ and truncation below the threshold $-\mathbf{S}^{-1}\mathbf{U}_2$. This means that we can generate samples of θ by first drawing samples of $\mathbf{V}_1$ and $\mathbf{V}_0$, and then performing a linear combination of these. The sampling scheme is summarized in Algorithm 1. This approach is efficient, as sampling from both the multivariate normal distribution and the truncated normal distribution is straightforward.

Algorithm 1 Direct Sampling Scheme for the E-Step

```
Input:  The number of subjects $N$; the number of attributes $K$; model parameters $\widehat{\boldsymbol{\lambda}}^{(t-1)}$ and $\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}$.
 1. For each $\alpha \in \{0,1\}^K$:
        Compute  $\mathbf{U}_1{}^{(\alpha)}$, $\mathbf{U}_2{}^{(\alpha)}$ and $\mathbf{S}$ shown in Equation (14) using $\widehat{\boldsymbol{\lambda}}^{(t-1)}$ and $\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}$.
 2. For each $m$ in $\{1,2,\ldots,M_t\}$, perform:
            a) Sample $\mathbf{V}_0^{(\alpha,m)}$ from:
               $N(\mathbf{0}_D, \boldsymbol{\Sigma}_\theta - \boldsymbol{\Sigma}_\theta \mathbf{U}_1{}^\top(\mathbf{U}_1\boldsymbol{\Sigma}_\theta\mathbf{U}_1{}^\top + \mathbf{I}_K)^{-1}\mathbf{U}_1\boldsymbol{\Sigma}_\theta)$.
            b) Sample $\mathbf{V}_1^{(\alpha,m)}$ from:
               $\mathbf{V}_1 \sim TN(\mathbf{0}_D, \mathbf{S}^{-1}(\mathbf{U}_1\boldsymbol{\Sigma}_\theta\mathbf{U}_1{}^\top + \mathbf{I}_K)\mathbf{S}^{-1}, -\infty, -\mathbf{S}^{-1}\mathbf{U}_2)$.
            c) Compute $\theta^{(\alpha,m)} =$
               $\mathbf{V}_0^{(\alpha,m)} + \widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}(\mathbf{U}_1{}^{(\alpha)})^\top(\mathbf{U}_1{}^{(\alpha)}\widehat{\boldsymbol{\Sigma}}_\theta^{(t-1)}(\mathbf{U}_1{}^{(\alpha)})^\top + \mathbf{I}_K)^{-1}\mathbf{S}\mathbf{V}_1^{(\alpha,m)}$.
Output: $\theta^{(\alpha,m)}$ for each $m = 1,\ldots,M_t$, $\alpha \in \{0,1\}^K$.
```

4.2.2 The M-Step Implementation

In M-Step, the maximization of $Q^{(t)}$ can be divided into three optimization problems.

$$\widehat{\beta}^{(t)} = \arg\max_{\beta} Q_1{}^{(t)}(\beta) - N \cdot p_{\mathbf{s}}(\beta), \tag{38}$$

$$\widehat{\boldsymbol{\lambda}}^{(t)} = \arg\max_{\boldsymbol{\Theta}} Q_2{}^{(t)}(\boldsymbol{\lambda}), \tag{39}$$

$$\widehat{\boldsymbol{\Sigma}}_\theta^{(t)} = \arg\max_{\boldsymbol{\Sigma}_\theta} Q_3{}^{(t)}(\boldsymbol{\Sigma}_\theta), \quad \boldsymbol{\Sigma}_\theta \succeq 0,\ \sigma_{dd'} = 1,\ d = 1,\ldots,D. \tag{40}$$

The optimization objective in (38) incorporates an L_1 penalty and utilizes various functions $g(\cdot)$ according to the response data type and measurement models. The optimization of (39) falls within the field of generalized linear models with a probit link. We use the coordinate ascent method, as described in Friedman et al. (2010) and Tay et al. (2023), to solve both problems. This method is known for its flexibility and power in solving such optimization problems.

Regarding the optimization of Equation (40), to estimate a covariance matrix with the constraint that diagonal elements must be ones, we employ a method inspired by the approach used in Stan (Carpenter et al., 2017) and detailed by Lewandowski et al. (2009). Below, we elaborate on this method as a two-step estimation procedure. First, we estimate $\boldsymbol{\Sigma}_\theta$ without considering the constraints, and denote the resulting estimator as $\widehat{\boldsymbol{\Sigma}}_\theta^* = (\widehat{\sigma}_{dd'}^*)$. In each iteration t, according to Equations (27) and (31), this solution can be directly derived as,

$$\widehat{\boldsymbol{\Sigma}}_\theta^{*(t)} = \frac{1}{M_t N} \sum_{m=1}^{M_t} \sum_{\alpha \in \{0,1\}^K} \left(\theta^{(m,\alpha)} \theta^{(m,\alpha)\top} \sum_{i=1}^{N} \psi_{i,\alpha}^{(t-1)} \right). \tag{41}$$

For simplicity, we neglect the iteration-specific notation (t) and introduce the computation based on a general $\widehat{\boldsymbol{\Sigma}}_\theta^*$.

Next, we compute the final estimator $\widehat{\boldsymbol{\Sigma}}_\theta$ that satisfies our diagonal constraints by reparameterizing the variance by its Cholesky decomposition $\boldsymbol{\Sigma}_\theta = \mathbf{G}^\top \mathbf{G}$. Let $\mathbf{G} = [g_{dd'}]$, and let $\mathbf{g}_d$ denote the d-th column of $\mathbf{G}$. To ensure the diagonal entries of $\widehat{\boldsymbol{\Sigma}}_\theta$ are ones, the upper triangular Cholesky factor $\mathbf{G}$ must satisfy $\|\mathbf{g}_d\| = 1$. Based on $\widehat{\boldsymbol{\Sigma}}_\theta^*$, we can use the following procedure to obtain $\widehat{\mathbf{G}}$:

$$\widehat{g}_{dd'} = \begin{cases} 0 & \text{if } d > d', \\ 1 & \text{if } d = d' = 1, \\ z_{dd'} & \text{if } d = 1 < d', \\ z_{dd'} g_{d-1,d'} (1 - z_{d-1,d'}^2)^{1/2} & \text{if } 1 < d \leq d'. \end{cases} \tag{42}$$

Here, $z_{dd'} = \tanh(\widehat{\sigma}_{dd'}^*)$. After obtaining $\widehat{\mathbf{G}}$, we compute $\widehat{\boldsymbol{\Sigma}}_\theta$ based on $\widehat{\boldsymbol{\Sigma}}_\theta = \widehat{\mathbf{G}}^\top \widehat{\mathbf{G}}$.

We summarize the above Monte Carlo EM Algorithm for the HO-GRCDM in Algorithm 2.

Remark 1. *Many HO-GRCDMs may have additional dispersion parameters. In these cases, additional steps are needed in the M step to update these parameters explicitly or with the help of existing optimizers. For example, in the log-normal CDM case, there are additional standard deviation parameters γ_j that need to be estimated. After obtaining estimates for all β^j, the γ_j can be solved explicitly by $\gamma_j = \sum_{i=1}^N \sum_{\alpha \in \{0,1\}^K} (R_{ij} - \nu_{j,\alpha})/(N \cdot 2^K)$ in each M step.*

Remark 2. *The Monte Carlo EM algorithm offers flexibility in handling complex models straightforwardly while achieving high accuracy. However, it is known for its relatively high computational cost, especially when the sample size N is large, due to the necessity for a sufficient number of Monte Carlo samples. We are pleased to report that this issue has a rather negligible effect*

Algorithm 2 Monte Carlo EM Algorithm for the HO-GRCDM

```
**Input:** Response matrix $\mathbf{R}$, number of binary latent attributes $K$, Monte Carlo sample size $M_t$ for $t = 1, 2, \ldots$, latent layer Q-matrix $Q^{(H)}$, initial values $\beta^{(\mathbf{0})}$, $\widehat{\boldsymbol{\lambda}}^{(0)}$, and $\widehat{\boldsymbol{\Sigma}}_{\theta}^{(0)}$.
*While not converged do:*
**E-Step:**
  *E1.* For $m = 1, 2, \ldots, M_t$ and $\alpha \in \{0,1\}^K$, sample $\theta^{(m,\alpha)}$ from $P(\theta | \mathbf{A} = \alpha; \widehat{\boldsymbol{\lambda}}^{(t-1)}, \boldsymbol{\Sigma}_{\theta}^{(t-1)})$ according to Algorithm 2.
  *E2.* Use the draws obtained in step (E1) to approximate the expectations $E_{\theta_i}\left[{l_2}^{(i,\alpha)} | \mathbf{A}_i = \alpha, \widehat{\boldsymbol{\lambda}}^{(t-1)}, \widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}\right]$ and $E_{\theta_i}\left[{l_3}^{(i)} | \mathbf{A}_i = \alpha, \widehat{\boldsymbol{\lambda}}^{(t-1)}, \widehat{\boldsymbol{\Sigma}}_{\theta}^{(t-1)}\right]$ according to Equations (30)-(31).
  *E3.* Substitute the expectations computed in step (E2) into the corresponding parts in Equations (25)-(27) to compute ${Q_1}^{(t)}(\beta)$, ${Q_2}^{(t)}(\boldsymbol{\lambda})$, and ${Q_3}^{(t)}(\boldsymbol{\Sigma}_{\theta})$.
**M-Step:**
  *M1.* Apply the coordinate ascent algorithm to solve the optimization problem in Equations (38) and (39) to obtain $\widehat{\beta}^{(t)}$ and $\widehat{\boldsymbol{\lambda}}^{(t)}$, and solve for the other model parameters if there are any.
  *M2.* Follow the procedure described in Section 4.2.2 to compute $\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t)}$.
  *M3.* Estimate the $Q$-matrix based on $\widehat{\beta}^{(t)}$, as discussed in the *Q-matrix estimation* section.
**Output:** Updated parameters $\widehat{\beta}^{(t)}$, $\widehat{\boldsymbol{\lambda}}^{(t)}$, and $\widehat{\boldsymbol{\Sigma}}_{\theta}^{(t)}$, along with an estimated $Q$-matrix.
```

on our proposed Algorithm 2. Instead of approximating the entire expectation of the E-step using Monte Carlo integration, only the inner expectation requires approximation. Moreover, only 2^K expectations need to be estimated, regardless of the sample size. For instance, even with a relatively large $K = 8$, only 256 integrals need to be estimated. This significantly differs from the traditional estimation approaches in IRT models, where the computational cost of Monte Carlo sampling would notably increase as the sample size N grows.

Remark 3. *Initialization is an important issue for the EM algorithm. An efficient approach is to use the SVD-based algorithm to obtain the initial values (Chen et al., 2019; Zhang et al., 2020). This method leverages the low-rank nature of the design matrix to capture the principal components of the data, thus providing stable and informative starting points for the iterative fitting process. The detailed procedures for initialization are presented in the Supplementary Material.*

5 Simulation Studies

In this section, we conduct extensive simulation studies to investigate the performance of the proposed MCEM algorithm for HO-GRCDMs. We consider three sample sizes: $N = 500$, 1000, and 2000, under the configuration $(J, K, D) = (30, 7, 3)$. We examine three types of models for the bottom data layer: the main-effect model, the all-effect model, and the DINA model. Within each bottom-layer model category, three types of response models are considered: (a) Bernoulli CDM for binary data, (b) Poisson CDM for count data, and (c) Lognormal CDM for continuous data. In the Bernoulli CDM case, the logistic distribution is used to model the probability of a correct response. Since both the Lognormal and Gamma distribution can be used to model positive continuous responses, we consider the Lognormal distribution as a representative case and postpone the simulation under the Gamma distribution to the Supplementary Material. The distributions for each model are detailed in Equations (4), (5), and (7). The first K items are serve as anchor items. The bottom layer Q-matrix, $Q_{30\times 7}$, is shown in Equation (43). For each bottom layer model, we explore two higher-order layer structures: the subscale structure and the bifactor structure. The true slope coefficients for the higher-order layer are also presented in Equation (43). The unconstrained parameters in Σ_θ are sampled from a uniform distribution $U(0.1, 0.3)$.

It is straightforward to verify that the above models are identifiable. Specifically, $Q_{30\times 7}$ satisfies the conditions in Proposition 1, and $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ satisfy the the conditions in Propositions 2 and 3, respectively. Therefore, the identifiability conditions in Theorem 1 are satisfied.

$$Q_{30\times 7} = \begin{pmatrix} & & & \mathrm{I}_7 & & & \\ & & & \mathrm{I}_7 & & & \\ & & & \mathrm{I}_7 & & & \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad \boldsymbol{\lambda}_1^{\text{subscale}} = \begin{pmatrix} 1.5 & 0.0 & 0.0 \\ 1.5 & 0.0 & 0.0 \\ 0.0 & 1.2 & 0.0 \\ 0.0 & 1.2 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{pmatrix}, \quad \boldsymbol{\lambda}_1^{\text{bifactor}} = \begin{pmatrix} 0.5 & 0.6 & 0.0 \\ 0.5 & 0.6 & 0.0 \\ 0.8 & 0.7 & 0.0 \\ 0.8 & 0.7 & 0.0 \\ 1.0 & 0.0 & 0.9 \\ 1.0 & 0.0 & 0.9 \\ 1.0 & 0.0 & 0.9 \end{pmatrix}. \tag{43}$$

The Monte Carlo sample size M_t for sampling θ is determined by the formula $M_t = s_0 + s_1 \cdot (t-1)$. In this study, we set $s_0 = s_1 = 5$, meaning five samples are simulated in the first iteration, with the number of samples increasing by five in each subsequent iteration. The convergence criterion is $\max \|\widehat{\eta}^{(t)} - \widehat{\eta}^{(t-1)}\| < 0.04$ for three successive iterations. Initialization is implemented using the SVD-based algorithm mentioned in Remark 3 and detailed in the Supplementary Material. The coordinate ascent part in our algorithm is conducted by using the R package *glmnet* (Hastie et al., 2021). For each model fitting, we apply the proposed MCEM algorithm multiple times across a sequence of regularization parameters s, which vary across distributions and are listed in the Supplementary Material. The regularization parameter that produces the smallest BIC value is then selected to finalize the model fitting.

5.1 Simulations for Main-Effect HO-GRCDMs

For the main-effect models, we set the coefficients β_k^j in the same way as Lee and Gu (2024b):

$$\beta_0^j = c_0, \quad \beta_k^j = \frac{c_1}{\sum_{k=1}^K q_{jk}}, \quad \forall j \in [J], k \in [K],$$

where (c_0, c_1) are two constants, set to (-1,3) for the Lognormal-CDM, (0.5,1) for the Poisson-CDM, and (-2,4) for the Bernoulli-CDM. For the Lognormal-CDM, the dispersion parameter σ_j is set to 1 for all J items. 100 independent replications are conducted in each setting.

We report the estimation accuracy for the continuous parameters in Table 1, by displaying the Root Mean Squared Errors (RMSE) and absolute biases (aBias). Note that the differences in the bottom layer parametric families may lead to results that are not directly comparable across these models. In particular, the Bernoulli model adopts a nonlinear parametrization for the correct response probability, so its probability parameters are on a different scale than those under other models. As shown in Table 1, the estimation accuracy for all the parameters improves as the sample size increases. Furthermore, to examine the recovery of the discrete Q-matrix, we report the proportion of correctly estimated rows and entries in Q in Table 12. It can be seen that the estimation accuracy of Q is reasonably high and improves as the sample size grows. The simulation results in Table 1 and Table 12 provide empirical verification of our identifiability results.

Model	Higher-Order Structure	N	RMSE				aBias			
			β	λ	Σ_θ	γ	β	λ	Σ_θ	γ
Lognormal	Subscale	500	0.170	0.107	0.064	0.030	0.153	0.085	0.053	0.024
		1000	0.131	0.076	0.048	0.022	0.120	0.058	0.039	0.018
		2000	0.111	0.058	0.040	0.015	0.105	0.052	0.034	0.013
	Bifactor	500	0.166	0.240	0.178	0.032	0.148	0.188	0.164	0.026
		1000	0.129	0.097	0.172	0.023	0.118	0.156	0.159	0.018
		2000	0.109	0.063	0.158	0.011	0.102	0.130	0.143	0.013
Poisson	Subscale	500	0.265	0.345	0.164	–	0.220	0.274	0.135	–
		1000	0.221	0.232	0.115	–	0.189	0.188	0.092	–
		2000	0.199	0.191	0.096	–	0.173	0.154	0.074	–
	Bifactor	500	0.238	0.451	0.162	–	0.201	0.348	0.146	–
		1000	0.209	0.328	0.130	–	0.182	0.263	0.119	–
		2000	0.190	0.278	0.122	–	0.172	0.227	0.113	–
Bernoulli	Subscale	500	0.399	0.176	0.072	–	0.351	0.135	0.056	–
		1000	0.348	0.086	0.051	–	0.316	0.067	0.042	–
		2000	0.313	0.083	0.041	–	0.294	0.058	0.032	–
	Bifactor	500	0.389	0.279	0.147	–	0.340	0.200	0.124	–
		1000	0.344	0.189	0.132	–	0.312	0.149	0.110	–
		2000	0.305	0.167	0.106	–	0.285	0.131	0.102	–

Table 1: RMSE and aBias for the Main-Effect HO-GRCDM

Higher-Order Structure		Lognormal			Poisson			Bernoulli		
	N	500	1000	2000	500	1000	2000	500	1000	2000
Subscale	P_R	0.807	0.897	0.967	0.805	0.914	0.957	0.644	0.854	0.951
	P_E	0.970	0.985	0.994	0.958	0.982	0.956	0.937	0.977	0.993
Bifactor	P_R	0.767	0.854	0.945	0.811	0.918	0.969	0.647	0.840	0.933
	P_E	0.960	0.977	0.992	0.961	0.983	0.992	0.937	0.974	0.990

Table 2: Proportion of Correctly Recovered Rows (P_R) and Entries (P_E) in Q-matrix for the Main-Effect HO-GRCDM.

5.2 Simulations for All-Effect HO-GRCDMs

For the all-effect models, we set the true coefficients β_k^j as $\beta_0^j = c_0$, and

$$\beta_S^j = \begin{cases} \frac{c_1}{2^{|\mathscr{K}_j|}} & \prod_{l \in S} q_{j,l} = 1 \\ 0 & \prod_{l \in S} q_{j,l} = 0 \end{cases}$$

where $\mathscr{K}_j = \{k \in [K]; q_{jk} = 1\}$, $S \subseteq [K] \backslash \emptyset$. Here, c_0 and c_1 are the same constants that we have defined in Section 5.1. Similar to the main-effect model scenario, we apply the proposed MCEM algorithm to fit each model and conduct 100 independent replications for each setting. RMSE and aBias are computed for evaluating estimation accuracy.

These simulation results are presented in Table 3. As shown in Table 3, the estimation accuracy for all parameters improves as the sample size increases. Compared to the main-effect case, the accuracy of the continuous parameters is lower, which is unsurprising given the significantly larger number of parameters in this case. To examine the recovery of Q, we report the proportion of correctly estimated rows and entries in Q in Table 4. As explained in the Q *estimation* section, following the strict rule tends to yield an onverly dense matrix. Instead, we identify the j-th row of Q based on the largest non-zero interaction coefficient for item j. The estimation accuracy of Q improves as the sample size grows. In fact, the recovery of Q is better than in the main-effect case. This is likely due to the larger number of parameters in the all-effect case, which may help to recover the Q-matrix by providing more information about the dependence structure between the items and the attributes.

Model	Higher-Order Structure	N	RMSE				aBias			
			β	λ	Σ_θ	γ	β	λ	Σ_θ	γ
Lognormal	Subscale	500	0.211	0.159	0.059	0.031	0.192	0.127	0.047	0.025
		1000	0.152	0.109	0.046	0.023	0.137	0.084	0.038	0.019
		2000	0.116	0.083	0.042	0.016	0.106	0.066	0.032	0.013
	Bifactor	500	0.191	0.209	0.143	0.031	0.170	0.167	0.131	0.025
		1000	0.140	0.150	0.141	0.021	0.125	0.120	0.129	0.017
		2000	0.114	0.123	0.126	0.017	0.104	0.095	0.109	0.013
Poisson	Subscale	500	0.310	0.258	0.072	–	0.292	0.199	0.059	–
		1000	0.280	0.174	0.054	–	0.272	0.139	0.044	–
		2000	0.270	0.134	0.045	–	0.264	0.113	0.037	–
	Bifactor	500	0.310	0.233	0.126	–	0.293	0.187	0.110	–
		1000	0.282	0.196	0.122	–	0.274	0.159	0.109	–
		2000	0.270	0.160	0.098	–	0.266	0.133	0.080	–
Bernoulli	Subscale	500	0.481	0.326	0.064	–	0.416	0.185	0.053	–
		1000	0.388	0.154	0.051	–	0.346	0.105	0.038	–
		2000	0.335	0.062	0.040	–	0.304	0.051	0.033	–
	Bifactor	500	0.479	0.355	0.171	–	0.414	0.279	0.149	–
		1000	0.393	0.191	0.106	–	0.349	0.152	0.103	–
		2000	0.323	0.178	0.102	–	0.296	0.140	0.097	–

Table 3: RMSE and aBias for the All-Effect HO-GRCDM

Higher-Order Structure		Lognormal			Poisson			Bernoulli		
	N	500	1000	2000	500	1000	2000	500	1000	2000
Subscale	P_R	0.942	0.958	0.982	0.883	0.923	0.972	0.827	0.953	0.992
	P_E	0.991	0.994	0.997	0.982	0.988	0.996	0.972	0.993	0.999
Bifactor	P_R	0.928	0.945	0.983	0.850	0.952	0.975	0.837	0.945	0.975
	P_E	0.989	0.992	0.998	0.983	0.994	0.996	0.974	0.992	0.996

Table 4: Proportion of Correctly Recovered Rows (P_R) and Entries (P_E) in Q-matrix for the All-Effect HO-GRCDM.

5.3 Simulations for DINA HO-GRCDMs

For the DINA models, we set the true coefficients β_k^j as $\beta_0^j = c_0$, and

$$\beta_S^j = \begin{cases} c_1 & \text{if } S = \mathscr{K}_j, \\ 0 & \text{otherwise.} \end{cases}$$

As mentioned in section 2.1, DINA model can be regarded as a special case of the all-effect model, meaning that we can use the same estimation procedure with all-effect model to estimate DINA model. Here, we set c_0 and c_1 same as Section 5.1, and implement the MCEM algorithm.

These simulation results are presented in Table 5. As demonstrated in Table 5, the estimation accuracy for all parameters improves as the sample size increases. Furthermore, to examine the recovery of Q, we report the proportion of correctly estimated rows and entries in Q in Table 6. Since there should be only one non-zero coefficient per item, we identify the j-th row of Q based on the largest non-zero interaction coefficient for item j. Again, the estimation accuracy of Q improves as the sample size grows. The results are similar to those in the all-effect case. In the confirmatory case, the DINA model has fewer parameters than the all-effect model, making it easier to estimate. However, in the exploratory case, it is fitted as an all-effect model, leading to similar computational cost. This may explain the comparable estimation accuracy in the two cases.

Model	Higher-Order Structure	N	RMSE				aBias			
			β	$\boldsymbol{\lambda}$	Σ_θ	γ	β	$\boldsymbol{\lambda}$	Σ_θ	γ
Lognormal	Subscale	500	0.205	0.185	0.064	0.031	0.189	0.121	0.050	0.025
		1000	0.146	0.145	0.048	0.023	0.135	0.106	0.037	0.019
		2000	0.113	0.087	0.037	0.016	0.106	0.080	0.027	0.013
	Bifactor	500	0.204	0.218	0.171	0.033	0.189	0.168	0.159	0.027
		1000	0.150	0.158	0.151	0.022	0.138	0.126	0.141	0.018
		2000	0.117	0.143	0.150	0.016	0.111	0.118	0.135	0.013
Poisson	Subscale	500	0.387	0.196	0.081	–	0.377	0.163	0.067	–
		1000	0.359	0.186	0.080	–	0.353	0.135	0.054	–
		2000	0.338	0.143	0.065	–	0.335	0.098	0.048	–
	Bifactor	500	0.391	0.206	0.131	–	0.380	0.166	0.120	–
		1000	0.357	0.152	0.130	–	0.352	0.124	0.117	–
		2000	0.337	0.129	0.127	–	0.334	0.102	0.114	–
Bernoulli	Subscale	500	0.458	0.322	0.073	–	0.403	0.180	0.057	–
		1000	0.348	0.129	0.055	–	0.316	0.101	0.042	–
		2000	0.320	0.109	0.036	–	0.300	0.051	0.028	–
	Bifactor	500	0.472	0.361	0.162	–	0.417	0.281	0.142	–
		1000	0.383	0.208	0.158	–	0.353	0.162	0.136	–
		2000	0.324	0.191	0.108	–	0.305	0.155	0.109	–

Table 5: RMSE and aBias for the DINA HO-GRCDM

Higher-Order Structure		Lognormal			Poission			Bernoulli		
	N	500	1000	2000	500	1000	2000	500	1000	2000
Subscale	P_R	0.898	0.915	0.947	0.928	0.955	0.975	0.790	0.890	0.975
	P_E	0.985	0.987	0.992	0.990	0.993	0.996	0.968	0.983	0.996
Bifactor	P_R	0.888	0.908	0.942	0.943	0.963	0.983	0.784	0.897	0.950
	P_E	0.984	0.986	0.991	0.991	0.995	0.998	0.966	0.985	0.993

Table 6: Proportion of Correctly Recovered Rows (P_R) and Entries (P_E) in Q-matrix for the DINA HO-GRCDM.

6 Real Data Analysis

In this section, we demonstrate the application of the HO-GRCDM to response time data from the 2019 TIMSS mathematics assessment (Fishbein et al., 2021). We analyze the response time data collected from United Arab Emirates students responding to booklet 1 in the eighth-grade math assessment[1]. This dataset includes the time spent on each item screen (in seconds) by 1599 students on 28 items within a total exam time of 45 minutes. Data points with log response times less than 0 or greater than 6 are regarded as outliers, potentially resulting from students' random guessing, running out of time, or taking breaks during the exam. These outliers are considered missing data (NA), and the corresponding observations are deleted, resulting in a total of 1163 observations. The data are then transformed from seconds to minutes. Table 6 summarizes some descriptive information about the items, including item type, item description, and the slope and location parameters obtained when fitting the two parameter item response model (2PL).

The TIMSS 2019 math assessment is designed to examine three cognitive skills (Knowing, Applying, and Reasoning) and four content skills (Number, Algebra, Geometry, and Data and Probability), where each item specified to measure one cognitive skill and one content skill. The assessment also provides a provisional Q-matrix and $Q^{(H)}$-matrix shown in Tables 8-9. We apply the developed MCEM algorithm to fit a main-effect HO-GRCDM with a subscale higher-order structure (using the provided $Q^{(H)}$-matrix) to the dataset. We elaborate more regarding the rationale for the choosing the structure of the HO-GRCDM in subsequent paragraphs. Here, note that we

[1] The data and the listed descriptive information are available from https://timssandpirls.bc.edu/timss2019/.

work under the exploratory framework with an unknown Q-matrix, and do not use any information in the provisional Q-matrix. This is because our purpose is to derive findings on the test structure and item characteristics that might provide valuable insights for test developers by assessing the validity of the test design Q-matrix.

Table 7: Descriptive item information. In the second column, MC denotes multiple choice items and CR denotes constructed response items.

Item	Item Type	Label	Slope	Location
1	MC	Octagon with equivalent shading	1.65	0.55
2	CR	Time when Pat finishes last lap	1.28	-0.26
3	MC	Multiples of 3	1.20	0.68
4	CR	Convert decimal to a fraction	1.24	0.39
5	MC	Expression for area of rectangle	1.32	0.67
6	MC	Expression with exponents of y	1.01	0.11
7	CR	Number of matches for figure 10	0.86	0.29
8	MC	Graph of y = 2x	1.24	1.62
9	MC	Rotation and reflection	1.13	1.59
10	MC	Surface area of the prism	1.50	0.99
11	MC	Value of angle x outside triangle	1.20	0.27
12	MC	Number of balls in a bag	1.19	-0.10
13	MC	Liv's smartphone use	1.91	0.76
14	CR	Statements for all values of integer a	0.74	1.04
15	MC	Arrow to show 5/12 on number line	1.49	0.74
16	CR	Value of fraction X in square	1.32	1.13
17	CR	Number of blue beads on bracelet	0.74	0.07
18	MC	Value of 2(6x - 3y)	1.29	0.09
19	MC	Expression equivalent to 2y + 6xy2	0.86	0.66
20	CR	Formula for stopping distance	1.16	0.66
21	CR	Value of x given perimeter of triangle ABC	1.60	0.92
22	MC	Additional point on a straight line	1.25	0.75
23	CR	Value of angle x in a quadrilateral	1.31	-0.15
24	CR	Methods of folding paper	0.50	0.33
25	CR	Coordinates to complete KLMN	1.23	0.72
26	CR	Mean temperature for 5 days	1.51	0.69
27	CR	Best graph for town information	1.57	0.14
28	CR	Bar graph of newspaper sales	1.05	1.47

Note: The Slope and Location refer to the item slope and location parameters obtained by fitting an item response model.

Item	Number	Algebra	Geometry	Data and Prob.	Knowing	Applying	Reasoning
1	1	0	0	0	1	0	0
2	1	0	0	0	0	1	0
3	1	0	0	0	0	0	1
4	1	0	0	0	1	0	0
5	0	1	0	0	0	1	0
6	0	1	0	0	1	0	0
7	0	1	0	0	0	0	1
8	0	1	0	0	0	1	0
9	0	0	1	0	0	0	1
10	0	0	1	0	0	0	1
11	0	0	1	0	0	1	0
12	0	0	0	1	0	1	0
13	0	0	0	1	0	1	0
14	1	0	0	0	1	0	0
15	1	0	0	0	0	1	0
16	1	0	0	0	0	0	1
17	1	0	0	0	1	0	0
18	0	1	0	0	1	0	0
19	0	1	0	0	1	0	0
20	0	1	0	0	0	1	0
21	0	1	0	0	0	1	0
22	0	1	0	0	0	1	0
23	0	0	1	0	0	1	0
24	0	0	1	0	0	0	1
25	0	0	1	0	0	0	1
26	0	0	0	1	1	0	0
27	0	0	0	1	0	1	0
28	0	0	0	1	0	0	1

Table 8: Q-matrix of the TIMSS 2019 data set

	Number	Algebra	Geometry	Data and Prob.	Knowing	Applying	Reasoning
Content	1	1	1	1	0	0	0
Cognitive	0	0	0	0	1	1	1

Table 9: $Q^{(H)\top}$ matrix of the TIMSS 2019 data set

Regarding the choice of parametric families to model the response time, the log-normal distribution and Gamma distribution are two commonly used distributions (De Boeck and Jeon, 2019; Maris, 1993; Klein Entink et al., 2009; Van der Linden, 2006). The log-normal model is often used when the logarithm of the response times follows a normal distribution, while the Gamma model is suitable for modeling positive continuous variables with skewed distributions. To choose between these two models, we first fit both models to the data and compute their BIC values. The obtained BIC values are 73165.46 for the Gamma model and 86664.8 for the log-normal model. Furthermore, we plot the probability histogram for the response time of each item and fit a density curve using the spline method. Using the estimated parameters obtained by fitting the log-normal and Gamma HO-GRCDM, their corresponding density curves are also plotted. We show plots for the last 10 items (items 19-28) in Figure 1 and present plots for all items in Supplementary Material E. By examining the histogram and comparing the density curves, we found that the response time variables for most items are right-skewed, and the Gamma model's curve (red line) overlaps more with the empirical density (blue line) than the log-normal model (green line), indicating that the

Gamma model has a better fit. Therefore, we use the Gamma distribution for the response time to fit a main-effect HO-GRCDM. We conduct an additional simulation study in Supplementary Material D to investigate the performance of the MCEM algorithm for estimating the main-effect HO-GRCDM with a Gamma distribution.

We also comment on the choice of the measurement model and the latent layer structure for our HO-GRCDM. It is common to use the main-effect models to understand response times (Sternberg, 1980; Maris, 1993; De Boeck and Jeon, 2019; Lee and Gu, 2024b). Lee and Gu (2024b) discusses the rationale for adopting the additive model assumption in more detail. As for the high-order layer, rather than focusing on assessing general cognitive ability, our primary goal is to capture fine-grained distinctions between subskills, for which the subscale model is more appropriate.

Figure 2(a) presents heatmap of the estimated bottom-layer parameters. The fitted model reveals a sparse structure of bottom-layer coefficients—only three columns (the 1st, 4th, and 6th columns) have non-zero coefficients. In addition, there is an estimated positive correlation of 0.35 between the cognitive skill and the content skill. Using the estimated parameters, we compute $P(\mathbf{A}_i = \alpha | \mathbf{R}_i; \widehat{\beta}, \widehat{\boldsymbol{\lambda}}, \widehat{\Sigma}_\theta)$ for each student, selecting the $\alpha \in \{0,1\}^K$ with the highest value as the estimate of their attribute profile. To explore the sparse structure and explain the three attributes with non-zero coefficients, we compute correlations among the seven attributes based on the estimated attributes of students and show the results in Figure 2(b). The first observation from the correlation plot is that the three cognitive attributes, attributes 5-7, are extremely highly correlated, with correlations up to 0.98. There are also high correlations among the four content attributes (attributes 1-4), with a correlation of 0.98 between attributes 1 and 3, and a correlation of 0.76 between attributes 2 and 4. This indicates that the content attributes may further divided into two groups: (1, 3) and (2, 4). Additionally, the correlations between one cognitive attribute and another content attribute are much smaller than that from content attributes or cognitive attributes.

The high correlations in attribute groups (1,3), (2,4), and (5,6,7) indicate that attributes within the same group are hard to distinguish. This is also reflected in the estimated bottom-layer coefficients β: only one attribute in each group has non-zero coefficients. The 6th column coefficients can be regarded as the coefficients for all of the cognitive skills: Knowing, Reasoning, and Applying. For attributes 1-4, we found that only items 1-15 have non-zero coefficients on the first attribute, with the majority measuring “Number” and “Algebra”. Items 24-28 have non-zero coefficients for attribute 4, and these items measure either “Geometry” or “Data and Probability”. These observations suggest that attribute groups (1,3), (2,4), and (5,6,7) correspond to (Number, Algebra), (Geometry, Data and Probability), and (Knowing, Reasoning, Applying), respectively.

The insights into the relationships among the attributes are both intuitive and interpretable. However, the estimated coefficients do not align with the test-design based Q-matrix in Table 8. A similar dataset was analyzed in Lee and Gu (2024b) using the designed Q-matrix, where their results revealed interesting observations about the intrinsic dependence among the attributes, which motivated our attempt to apply the HO-GRCDM to the TIMSS data. In this paper, our main focus is on developing a new modeling framework, providing identification results, and proposing an efficient algorithm for fitting this model, which together form a comprehensive toolkit for practitioners. We encourage applied researchers to design tests specifically tailored for HO-GRCDMs in the future, for which our methodology is well-suited and could perform optimally.

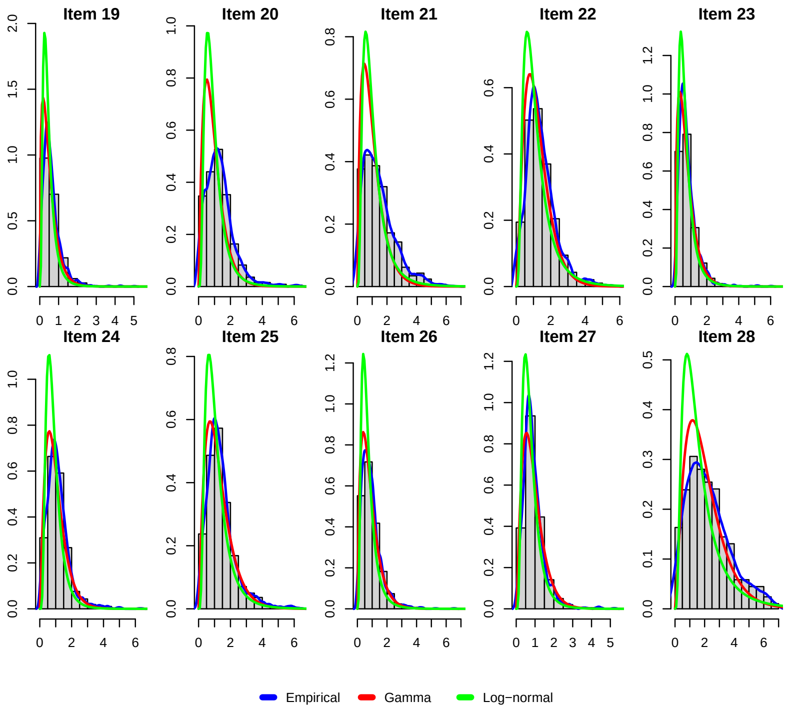
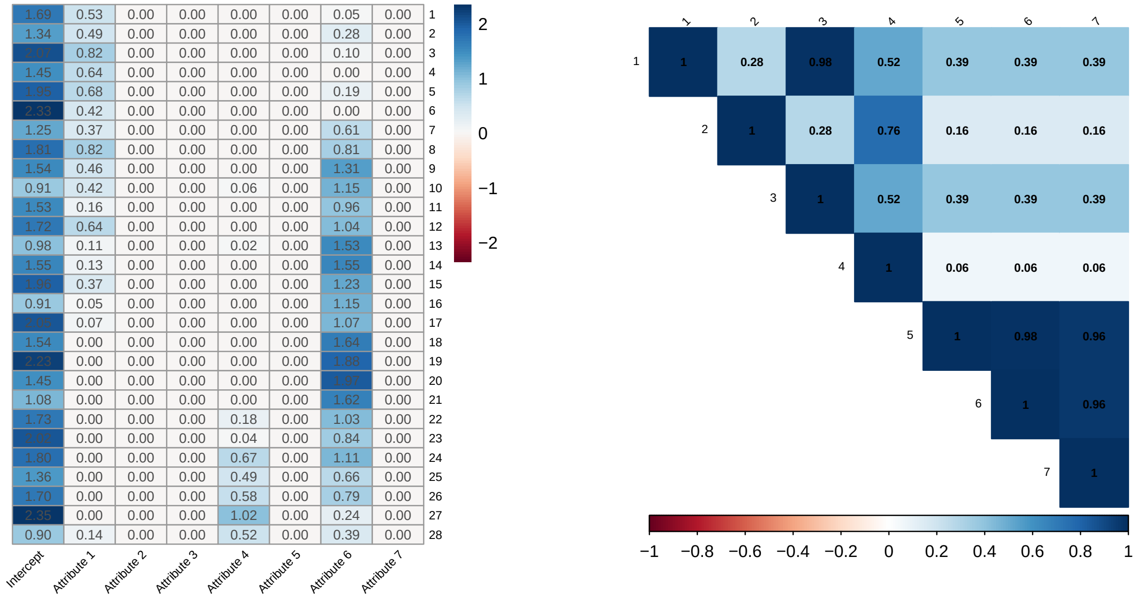


Figure 1: TIMSS data analysis. Probability Histogram and Fitted Density Curves (Empirical Density, Gamma Model, and Log-Normal Model) for Response Time Data (in minutes).



(a)

(b)

Figure 2: TIMSS data analysis. (a): Heatmap for Estimated Bottom-Layer Parameters. (b): Correlation Plot of the Estimated Latent Attributes.

7 Discussion

We have proposed a general modeling framework, HO-GRCDM, for modeling CDMs with general responses and a higher-order structure. This framework features high flexibility in (1) addressing various types of response data, (2) adapting to a variety of measurement models, and (3) considering an exploratory settings with an unknown Q-matrix. Furthermore, our models have a rich representational power in its hierarchical structure to hunt for higher-order cognitive information. We provide interpretable identifiability conditions in terms of the Q and $Q^{(H)}$ matrices that ensure the validity and accuracy of model fitting. The probit link used for the higher-order layer facilitates our identifiability theory as well as the development of an efficient MCEM algorithm for parameter estimation. Compared to existing MCMC and EM methods, our MCEM algorithm has lower computational complexity due to an explicit conditional expectation formula for α and an efficient sampler for θ. Extensive simulation studies under various response types and measurement models are conducted to examine the efficiency of the proposed algorithm.

There are several promising directions for future work that build upon the HO-GRCDM framework. *First*, incorporating more than two layers would help explore deeper and more nuanced diagnostic information. Gu (2024) proposed a new family of DeepCDMs featuring multiple, potentially deep, entirely discrete latent layers for cognitive diagnosis. However, that work focuses on binary response variables and binary latent variables. It would be interesting to develop a framework applicable to general responses and incorporate multiple discrete latent layers. *Second*, while it is typical to consider binary attributes in CDMs, extending them to polytomous attributes can provide a more nuanced representation of latent traits or skills (von Davier, 2005; Karelitz, 2004; Chen and de la Torre, 2013). *Third*, it is worth considering a fully exploratory setting with both the Q and $Q^{(H)}$ matrices being unknown. Investigating identifiability and estimation in this setting in the future could provide further insights into cognitive processes and test design.

Supplementary Material. The Supplementary Material contains the proofs of the identifiability results, the initialization strategy based on SVD, and additional details about the simulation studies and real data analysis.

Funding. Yuqi Gu's research is partially supported by the NSF grant DMS-2210796.

References

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574.

Béguin, A. A. and Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional irt models. *Psychometrika*, 66:541–561.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis–hastings robbins–monro algorithm. *Psychometrika*, 75:33–57.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Annals of Statistics*, pages 177–214.

Chen, J. and de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37(6):419–437.

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

Chen, Y., Li, X., and Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, 84:124–146.

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

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, 81(4):1142–1163.

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

De Boeck, P. and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10:422756.

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

de la Torre, J. and Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (2012). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively diagnostic assessment*, pages 361–389. Routledge.

Fang, G., Guo, J., Xu, X., Ying, Z., and Zhang, S. (2021). Identifiability of bifactor models. *Statistica Sinica*, 31:2309–2330.

Fishbein, B., Foy, P., and Yin, L. (2021). TIMSS 2019 user guide for the international database. *Hentet fra https://timssandpirls. bc. edu/timss2019/international-database*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CreateSpace.

Gu, Y. (2024). Going deep in diagnostic modeling: Deep cognitive diagnostic models (DeepCDMs). *Psychometrika*, 89(1):118–150.

Gu, Y. and Dunson, D. B. (2023). Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426.

Hastie, T., Qian, J., and Tay, K. (2021). An introduction to glmnet. *CRAN R Repositary*, 5:1–35.

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.

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

Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments*. University of Illinois at Urbana-Champaign.

Klein Entink, R. H., Fox, J.-P., and van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74:21–48.

Lee, S. and Gu, Y. (2024a). Deep discrete encoders: Identifiable deep generative models for rich data types with discrete latent layers. *Preprint*.

Lee, S. and Gu, Y. (2024b). New paradigm of identifiable general-response cognitive diagnostic models: beyond categorical data. *Psychometrika*.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Li, J., Gibbons, R., and Rockova, V. (2023). Sparse Bayesian multidimensional item response theory. *arXiv preprint arXiv:2310.17820*.

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.

Lord, F. (1952). A theory of test scores. *Psychometric monographs*.

Magnus, B. E. and Thissen, D. (2017). Item response modeling of multivariate count data with zero inflation, maximum inflation, and heaping. *Journal of Educational and Behavioral Statistics*, 42(5):531–558.

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.

Man, K. and Harring, J. R. (2023). Detecting preknowledge cheating via innovative measures: A mixture hierarchical model for jointly modeling item responses, response times, and visual fixation counts. *Educational and Psychological Measurement*, 83(5):1059–1080.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, 58:445–469.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64:187–212.

Minchen, N. D., de la Torre, J., and Liu, Y. (2017). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics*, 42(6):651–677.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests.* ERIC.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Science & Business Media.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General*, 109(2):119.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20:345–354.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106.

Templin, J., Henson, R. A., et al. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford press.

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

Templin, J. L., Henson, R. A., Templin, S. E., and Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, 32(7):559–574.

Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181–204.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2):i–35.

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

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1):57–87.

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

Zhang, H., Chen, Y., and Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika*, 85(2):358–372.

Supplementary Material

This Supplementary Material is organized as follows. Section A includes the proofs of the identifiability results. Section B presents the initialization strategy based on the singular value decomposition. Section C and Section D provides some additional details for the simulation studies and real data analysis, respectively.

A Proof of Identifiability Results

Proof of Proposition 2. We separate the arguments for sufficiency and necessity. Recall that for each d, we assume that there exists a *pivot row* k_d such that $\lambda_{1,d}^{k_d} > 0$.

Sufficiency. We prove that the model is identifiable under the given condition. For each $k \in [K]$, let Z_k denote the unique group that the kth item belongs to, and define $\delta_k := \frac{\lambda_{1,Z_k}^k}{\sqrt{1+(\lambda_{1,Z_k}^k)^2}}$. It is clear that $\lambda \mapsto \frac{\lambda}{\sqrt{1+\lambda^2}}$ is a one-to-one mapping. Hence, using Proposition 2 in Fang et al. (2021), it suffices to show that one can uniquely recover $\delta = (\delta_1, \ldots, \delta_K)$ and Σ, given the values

$$C_\rho(k,l) = \frac{\lambda_1^{k\top}\Sigma\lambda_1^l}{\sqrt{1+\lambda_1^{k\top}\Sigma\lambda_1^k}\sqrt{1+\lambda_1^{l\top}\Sigma\lambda_1^l}} = \sigma_{Z_kZ_l}\delta_k\delta_l$$

for $k \neq l$.

Fix any d. We first prove that δ_k for $k \in L_d := \{k \in [K] : Z_k = d\}$ can be identified. Let us first consider the first scenario with $K_d \geq 3$. For $k \neq l \in L_d$, we have $Z_k = Z_l = d$. Because $\sigma_{dd} = 1$, $C_\rho(k,l)$ simplifies into $\delta_k\delta_l$. Using the fact that $\delta_{k_d} > 0$ for the pivot row indexed by k_d, we can uniquely determine all δ_k for $k \in L_d$. To see this, one can simply take two additional indices $l \neq m \in L_d$ and consider the equations

$$C_\rho(k_d,l) = \delta_{k_d}\delta_l, \quad C_\rho(k_d,m) = \delta_{k_d}\delta_m, \quad C_\rho(l,m) = \delta_l\delta_m.$$

Next, we consider the case when $K_d = 2$ and $\sigma_{dd'} \neq 0$ for some $d' \neq d$. Since $\delta_{k_{d'}} \neq 0$ and $\sigma_{dd'} \neq 0$, one can recover the ratio of δ_ks for $k \in L_d = \{k_d, k_d+1\}$ by computing the ratio of $C_\rho(k,k_{d'}) = \sigma_{dd'}\delta_k\delta_{k_{d'}}$'s. Then, we can uniquely determine δ_k's for $k \in L_d$ using the value of

$C_\rho(k_d, k_d+1) = \delta_{k_d}\delta_{k_d+1}$.

Finally, it remains to recover the off-diagonal entries of Σ. But this directly follows by noting that $\sigma_{dd'} = \frac{C_\rho(k_d,k_{d'})}{\delta_{k_d}\delta_{k_{d'}}}$ for any $d \neq d'$. Here, note that $\delta_{k_d}\delta_{k_{d'}} \neq 0$ by the pivot row assumption, so the fraction is always well-defined.

Necessity. We show that the conditions in the theorem are necessary for identifiability using the proof by contradiction. Suppose an identifiable model has a d such that (1) $K_d = 1$ or (2) $K_d = 2$ and $\sigma_{dd'} = 0$ for all $d' \neq d$. Under the first case with $K_d = 1$, the d-th group is only measured by the k_d-th item. Then, the parameters $\delta_{k_d}, \{\sigma_{dd'}\}_{d' \neq d}$ are only reflected in $C_\rho(k_d, l) = \sigma_{dZ_l}\delta_{k_d}\delta_l$'s, for $l \neq k_d$. Even when assuming that the values of δ_l's are known, this is a system of $D-1$ equations with D unknowns, and does not exhibit a unique solution. Hence, we have a contradiction.

Next, we consider the case where $K_d = 2$ and $\sigma_{dd'} = 0$ for all $d' \neq d$. Then, $L_d = \{k_d, k_d+1\}$. Because $\sigma_{dd'} = 0$ for $d' \neq d$, $\rho_{kl} = 0$ for all $k \in L_d$. Hence, δ_{k_d} and δ_{k_d+1} needs to be determined only by the value of $C_\rho(k_d, k_d+1) = \delta_{k_d}\delta_{k_d+1}$. This is a single equation with two unknowns and not solvable, giving the contradiction. □

Proof of Theorem 1. The proof follows by sequentially applying the identifiability results for the CDM and the higher-order probit model in a layerwise manner. First, we consider the bottom layer CDM with parameters $(\boldsymbol{\pi}, \boldsymbol{\beta}, Q)$. Here, we are setting $\pi_{\boldsymbol{\alpha}} = P(\mathbf{A} = \boldsymbol{\alpha})$ by marginalizing out the probit latent layer using eq. (14), as mentioned in the main text. Then, under conditions A and B, Proposition 1 gives the identifiability of the CDM parameters $(\boldsymbol{\pi}, \boldsymbol{\beta}, Q)$, up to latent variable permutation. As we assume the knowlege of anchor items for each latent variable, we can identify these parameters without any trivial ambiguity.

Now, having identified $\boldsymbol{\pi}$, we fully know the marginal distribution of the latent variables. Now, we can apply Propositions 2 or 3 to identify the probit model parameters $(\Sigma_\theta, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)$ and the proof is complete. Note that the conditions for the probit parameters in Theorem 1 are exactly those in Propositions 2 or 3. □

B Initialization Procedure for Simulation Study

B.1 Initialization of Bottom Layers for Various Response Distributions

We use singular value decomposition (SVD) to find the starting values for the bottom layers. We first start at the Bernoulli model case and present Algorithm 1 for binary data as below.

Algorithm 3 Initialization for Bernoulli models

```
1. Input response data $\mathbf{R} = (r_{ij})_{N\times J}$, number of attributes $K$, link function g, and truncation parameter $\varepsilon_{N,J} > 0$.
2. Apply the singular value decomposition to $\mathbf{R} = \sum_{j=1}^{J} \tau_j \mathbf{u}_j \mathbf{v}_j^\top$, where $\tau_1 \geq \tau_2 \geq \dots \tau_J$ are the singular values, and $\mathbf{u}_j$s and $\mathbf{v}_j$s are left and right singular vectors, respectively.
3. Let $\mathbf{X} = (x_{ij})_{N\times J} = \sum_{k=1}^{\tilde{K}} \tau_k \mathbf{u}_k \mathbf{v}_k^\top$, where $\tilde{K} = \max\left\{K+1, \mathrm{arg\,max_k}\left\{\tau_{\mathrm{k}} \geq 1.01\sqrt{\mathrm{N}}\right\}\right\}$
4. Let $\widehat{\mathbf{X}} = (\widehat{x}_{ij})_{N\times J}$ be defined as
   $$\widehat{x}_{ij} = \begin{cases} \varepsilon_{N,J} & if \quad x_{ij} < \varepsilon_{N,J} \\ x_{ij} & if \quad \varepsilon_{N,J} \leq x_{ij} \leq 1-\varepsilon_{N,J} \\ 1-\varepsilon_{N,J} & if \quad x_{ij} \geq 1-\varepsilon_{N,J} \end{cases}$$
5. Let $\tilde{\mathbf{M}} = (\tilde{m}_{ij})_{N\times J}$, where $\tilde{m}_{ij} = g(\widehat{x}_{ij})$.
6. Let $\widehat{\beta}_0 = (\widehat{\beta}_0^1, \dots, \widehat{\beta}_0^J)$, where $\widehat{\beta}_0^j = (\sum_{i=1}^{N} \tilde{m}_{ij})/N$.
7. Apply singular value decomposition to $\widehat{\mathbf{M}} = (\tilde{m}_{ij} - \widehat{\beta}_0^j)_{N\times J}$ to have $\widehat{\mathbf{M}} = \sum_{j=1}^{J} \widehat{\tau}_j \widehat{\mathbf{u}}_j \widehat{\mathbf{v}}_j^\top$, where $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \dots \widehat{\tau}_J$ are the singular values, and $\widehat{\mathbf{u}}_j$s and $\widehat{\mathbf{v}}_j$s are left and right singular vectors, respectively.
8. Apply varimax to $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \dots, \widehat{\mathbf{v}}_J)$, and let $\tilde{\mathbf{V}}$ be the rotated version of $\widehat{\mathbf{V}}$.
9. Output $\widehat{\beta} = (\beta_k^j)_{J\times K} = \frac{1}{\sqrt{N}}(\widehat{\tau}_1 \tilde{\mathbf{v}}_1, \dots, \widehat{\tau}_K \tilde{\mathbf{v}}_K)$, $\widehat{\beta}_0$.
```

Algorithm 3 is based on the SVD-based estimator in Zhang et al. (2020). It utilizes SVD twice. The initial application of SVD, followed by the inverse transformation (Steps 2-5), serves to denoise and linearize the data. Subsequently, the second application of SVD (Steps 6-7) performs PCA on the linearized data. For further discussions on the details, please refer to Zhang et al. (2020) and Chatterjee (2015). The difference from Zhang et al. (2020) is that we apply the Varimax rotation to achieve a sparse and more interpretable factor loading structure in Step 8.

We utilize initialization for the HO-GRCDMs with other general responses based on a similar idea to that of Algorithm 3. Firstly, the procedure is simpler and more directly for Transformed-normal distribution. Let T_{ij} denote the transformed response variable. For example, $T_{ij} = \log(R_{ij})$ for log-normal distribution, $T_{ij} = \log(R_{ij}/(1-R_{ij}))$ for logistic-normal distribution, and so forth. After transforming R_{ij} to T_{ij}, there is no need to linearize data and truncate variable. The procedure of finding the starting points for Transformed-normal distributions is listed in Algorithm 4.

Algorithm 4 Initialization for Transformed-Normal models

```
1. Input transformed response data $\mathbf{T} = (t_{ij})_{N\times J}$, number of attributes $K$, link function g.
2. Apply the singular value decomposition to $\mathbf{T} = \sum_{j=1}^{J} \tau_j \mathbf{u}_j \mathbf{v}_j^\top$, where $\tau_1 \geq \tau_2 \geq \dots \tau_J$ are the
   singular values, and $\mathbf{u}_j$s and $\mathbf{v}_j$s are left and right singular vectors, respectively.
3. Let $\mathbf{X} = (x_{ij})_{N\times J} = \sum_{k=1}^{\tilde{K}} \tau_k \mathbf{u}_k \mathbf{v}_k^\top$, where $\tilde{K} = \max\left\{K+1, \arg\max_{\mathrm{k}}\left\{\tau_{\mathrm{k}} \geq 1.01\sqrt{\mathrm{N}}\right\}\right\}$,
   compute $\widehat{\beta}_0^j = (\sum_{i=1}^{N} x_{ij})/N$, $j = 1, \dots, J$.
4. Apply varimax to $\widehat{\mathbf{V}} = (\mathbf{v}_1, \dots, \mathbf{v}_J)$, and let $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_J)$ be the rotated version of $\widehat{\mathbf{V}}$.
5. Output $\widehat{\beta} = (\beta_k^j)_{J\times K} = \frac{1}{\sqrt{N}}(\tau_1 \tilde{\mathbf{v}}_1, \dots, \tau_K \tilde{\mathbf{v}}_K)$, $\widehat{\beta}_0 = (\widehat{\beta}_0^1, \dots, \widehat{\beta}_0^J)$.
```

The initialization of Poisson models presented in Algorithm 5. It is similar to the Bernoulli case. The difference is SVD is applied to transformed data $T_{ij} = \log(R_{ij}+1)$ instead of the original data R_{ij} to help stabilize the variance and reduce skewness, then the data is transformed back to the original scale in step 5.

Algorithm 5 Initialization for Poisson models

```
1. Input response data $\mathbf{R} = (r_{ij})_{N\times J}$, number of attributes $K$, link function g, and truncation parameter $\varepsilon_{N,J} = \log(1)$.
2. Transforming the data $\mathbf{R}$ to $\mathbf{T} = (t_{ij})_{N\times J}$, with $t_{ij} = \log(R_{ij}+1)$. Apply the singular value decomposition to $\mathbf{T} = \sum_{j=1}^{J} \tau_j \mathbf{u}_j \mathbf{v}_j^\top$, where $\tau_1 \geq \tau_2 \geq \dots \tau_J$ are the singular values, and $\mathbf{u}_j$s and $\mathbf{v}_j$s are left and right singular vectors, respectively.
3. Let $\mathbf{X} = (x_{ij})_{N\times J} = \sum_{k=1}^{\tilde{K}} \tau_k \mathbf{u}_k \mathbf{v}_k^\top$, where $\tilde{K} = \max\left\{K+1, \arg\max_{\mathrm{k}}\left\{\tau_{\mathrm{k}} \geq 1.01\sqrt{\mathrm{N}}\right\}\right\}$
4. Let $\widehat{\mathbf{X}} = (\widehat{x}_{ij})_{N\times J}$ be defined as
   $$\widehat{x}_{ij} = \begin{cases} \varepsilon_{N,J} & if \quad x_{ij} < \varepsilon_{N,J} \\ x_{ij} & if \quad x_{ij} \geq \varepsilon_{N,J} \end{cases}$$
5. Let $\tilde{\mathbf{M}} = (\tilde{m}_{ij})_{N\times J}$, where $\tilde{m}_{ij} = \exp(\widehat{x}_{ij}) - 1$.
6. Compute $\widehat{\beta}_0^j = (\sum_{i=1}^{N} \tilde{m}_{ij})/N$, $j = 1, 2, \dots, J$.
7. Apply singular value decomposition to $\widehat{\mathbf{M}} = (\tilde{m}_{ij} - \widehat{\beta}_0^j)_{N\times J}$ to have $\widehat{\mathbf{M}} = \sum_{j=1}^{J} \widehat{\tau}_j \widehat{\mathbf{u}}_j \widehat{\mathbf{v}}_j^\top$, where $\widehat{\tau}_1 \geq \widehat{\tau}_2 \geq \dots \widehat{\tau}_J$ are the singular values, and $\widehat{\mathbf{u}}_j$s and $\widehat{\mathbf{v}}_j$s are left and right singular vectors, respectively.
8. Apply varimax to $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \dots, \widehat{\mathbf{v}}_J)$, and let $\tilde{\mathbf{V}} = (\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_J)$ be the rotated version of $\widehat{\mathbf{V}}$.
9. Output $\widehat{\beta} = (\beta_k^j)_{J\times K} = \frac{1}{\sqrt{N}}(\widehat{\tau}_1 \tilde{\mathbf{v}}_1, \dots, \widehat{\tau}_K \tilde{\mathbf{v}}_K)$, $\widehat{\beta}_0 = (\widehat{\beta}_0^1, \dots, \widehat{\beta}_0^J)$.
```

B.2 Initialization of Higher-Order Layers

Once the initial values for the bottom layer are obtained, the conditional probability $P(\mathbf{R}_i^j|\alpha,\beta^j)$ for each model can be computed according to Equations (5)-(9). For each $\alpha \in \{0,1\}^K$, we can then compute the normalized likelihood of $\boldsymbol{\alpha}$,

$$P_\alpha = \frac{\prod_{i=1}^N P(\mathbf{R}_i^j|\alpha,\beta^j)}{\sum_{\alpha'}\prod_{i=1}^N P(\mathbf{R}_i^j|\alpha',\beta^j)}, \tag{44}$$

and use it as an initial approximation of $\pi_{\boldsymbol{\alpha}}$, the marginal proportion of the attribute pattern $\boldsymbol{\alpha}$. The approach of approximating marginals via normalized likelihood terms is well-recognized in Bayesian inference (Gelman et al., 2013) and in probabilistic modeling (Bishop and Nasrabadi, 2006).

Using P_α as an initial approximation of $\pi_{\boldsymbol{\alpha}}$, we generate a set of pseudo-attribute data according to the distribution $\{P_\alpha : \boldsymbol{\alpha} \in \{0,1\}^K\}$. This step aims to find appropriate starting points for $\lambda_{\mathbf{1}}{}^k$ and λ_0^k in the following model:

$$P(A_k = 1|\theta,\lambda_{\mathbf{1}}{}^k,\lambda_0^k) \;\; = \;\; f^{-1}(\theta^\top \lambda_{\mathbf{1}}{}^k + \lambda_0^k), \quad k = 1,\ldots,K. \tag{45}$$

Given the generated pseudo-attribute data $\mathbf{A}_{pseudo}$, this process is reduced to finding starting points for an item factor analysis model with a probit link $\Phi(\cdot)$, for which Algorithm 3 can be used to obtain the initial values. The complete initialization procedure is given in Algorithm 6. Note that, the sample size of $\mathbf{A}_{pseudo}$ is flexible and is not necessary to align with the response data size N. A larger number of $\mathbf{A}_{pseudo}$ will help ensure that the pseudo-attribute data distribution is sufficiently close to P_α, while a smaller number of $\mathbf{A}_{pseudo}$ can render larger randomness but enable faster computation if N is very large.

Algorithm 6 Complete Initialization Procedure

```
1. Response data $\mathbf{R} = (r_{ij})_{N \times J}$, number of attributes $K$, and link function $g$.
2. Based on the response type, apply one of Algorithms 3-5 accordingly to find initial values
   of $\widehat{\beta}$.
3. Compute $P_\alpha$ according to Equation (44).
4. Generate a set of pseudo-attribute data $\mathbf{A}_{pseudo}$ according to $P_\alpha$.
5. Impute $\mathbf{A}_{pseudo}$ into Equation (45) and apply Algorithm 3 to get the initial values of $\lambda$,
   using the probit link function $\Phi(\cdot)$.
6. Output the initial values $\widehat{\beta}$ and $\widehat{\lambda}$.
```

C Simulation Study Details

C.1 Simulation Details

Table 10 presents the sequences of the tuning parameters s for different sample sizes and models in the simulation study. The sequences are chosen to decrease as the sample size increases, following the theoretical suggestion for regularization parameter selection Chen et al. (2015). The magnitude of the sequences differs across models because of variations in the true parameter values.

Sample size	Model			
	Lognormal	Poisson	Bernoulli	Gamma
500	$(0.10, 0.12, 0.14)$	$(0.08, 0.09, 0.10)$	$(0.022, 0.023, 0.024)$	$(0.08, 0.09, 0.10)$
1000	$(0.08, 0.10, 0.12)$	$(0.07, 0.08, 0.09)$	$(0.020, 0.021, 0.022)$	$(0.08, 0.09, 0.10)$
2000	$(0.06, 0.08, 0.10)$	$(0.06, 0.07, 0.08)$	$(0.018, 0.019, 0.020)$	$(0.06, 0.07, 0.08)$

Table 10: Sequences of tuning parameters s for different sample sizes and models (Lognormal, Poisson, Bernoulli, and Gamma) used in the simulation study.

Model	Higher-Order Structure	N	RMSE				aBias			
			β	λ	Σ_θ	s	β	λ	Σ_θ	s
Gamma	Subscale	500	0.143	0.396	0.236	0.108	0.131	0.295	0.167	0.087
		1000	0.102	0.244	0.159	0.094	0.091	0.171	0.089	0.080
		2000	0.077	0.103	0.039	0.092	0.069	0.086	0.032	0.079
	Bifactor	500	0.140	0.459	0.241	0.110	0.127	0.347	0.228	0.088
		1000	0.094	0.326	0.196	0.102	0.084	0.250	0.186	0.085
		2000	0.073	0.245	0.177	0.093	0.066	0.196	0.161	0.072

Table 11: RMSE and aBias for the Gamma Model within the Main-Effect HO-GRCDM

C.2 Additional Simulation for the Gamma Model within a Main-Effect CDM

Like the other simulations conducted for main-effect models, we also set the coefficients β_k^j according to:

$$\beta_0^j = c_0, \quad \beta_k^j = \frac{c_1}{\sum_{k=1}^K q_{jk}}, \quad \forall j \in [J], k \in [K],$$

where (c_0, c_1) are two constants, set to (1, 2) for the Gamma-CDMs. These constants are chosen to match the scale of parameters obtained in the Empirical Data Analysis section. The estimation procedure is shown in Algorithm 1 in the paper. Here, the built-in R function *optim* is used to estimate the shape parameters. The obtained RMSE and aBias are presented in Table 11, and the proportion of correctly recovered rows and entries are shown in Table 12. It can be seen that the estimation accuracy of both model parameters and Q improves as the sample size grows.

	N	500	1000	2000
Subscale	P_R	0.727	0.787	0.854
	P_E	0.954	0.965	0.978
Bifactor	P_R	0.726	0.778	0.851
	P_E	0.955	0.965	0.977

Table 12: Proportion of Correctly Recovered Rows (P_R) and Entries (P_E) in Q-matrix for the Gamma Model within the Main-Effect HO-GRCDM

D Histograms and Fitted Density Curves for TIMSS Data

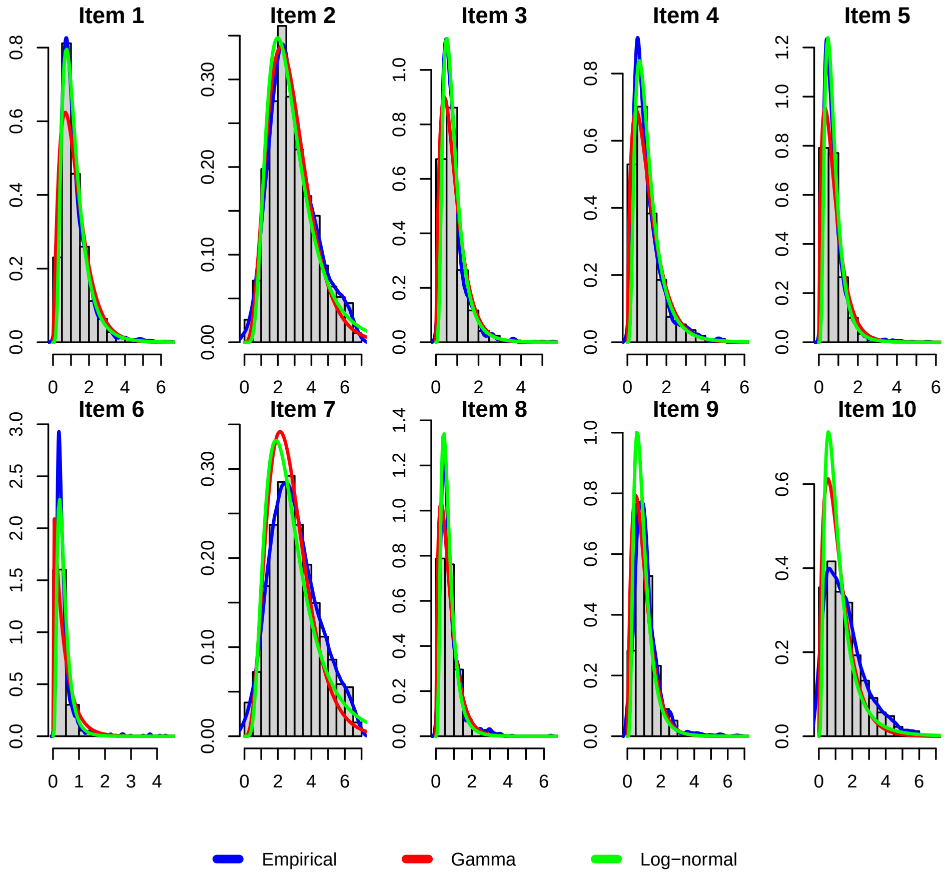


Figure 3: Probability Histogram and Fitted Density Curves (Empirical Density, Gamma Model, and Log-Normal Model) for Response Time Data (in Minutes) for Items 1-10.

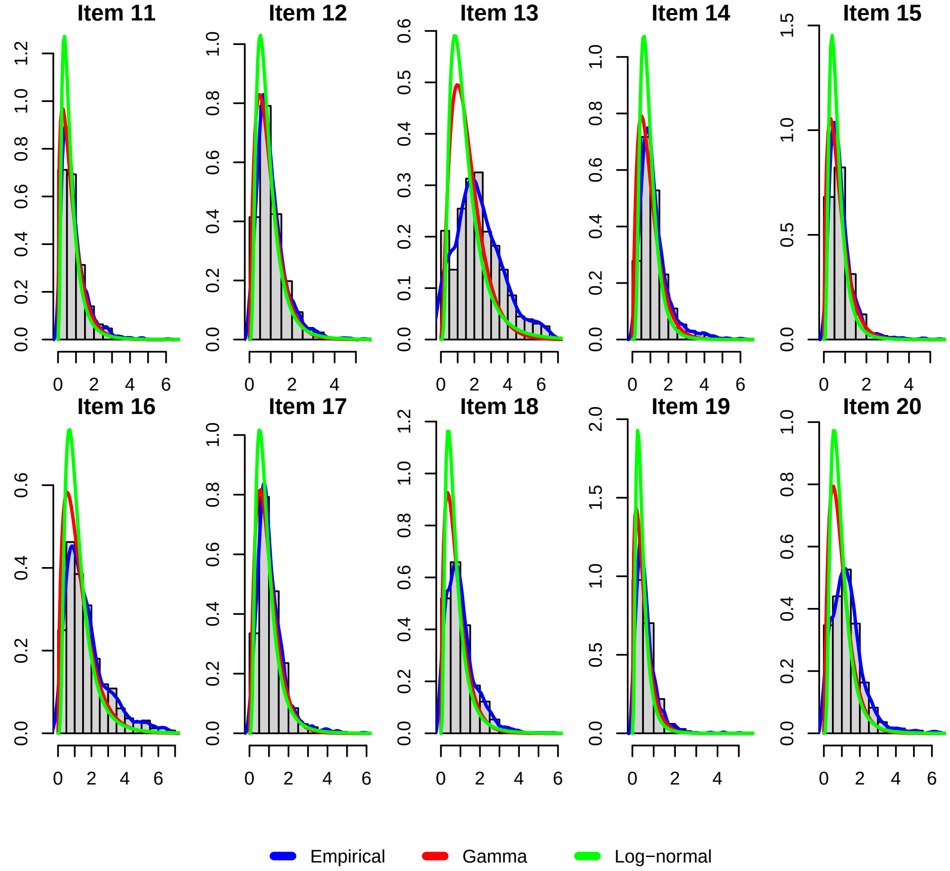


Figure 4: Probability Histogram and Fitted Density Curves (Empirical Density, Gamma Model, and Log-Normal Model) for Response Time Data (in Minutes) for Items 11-20.

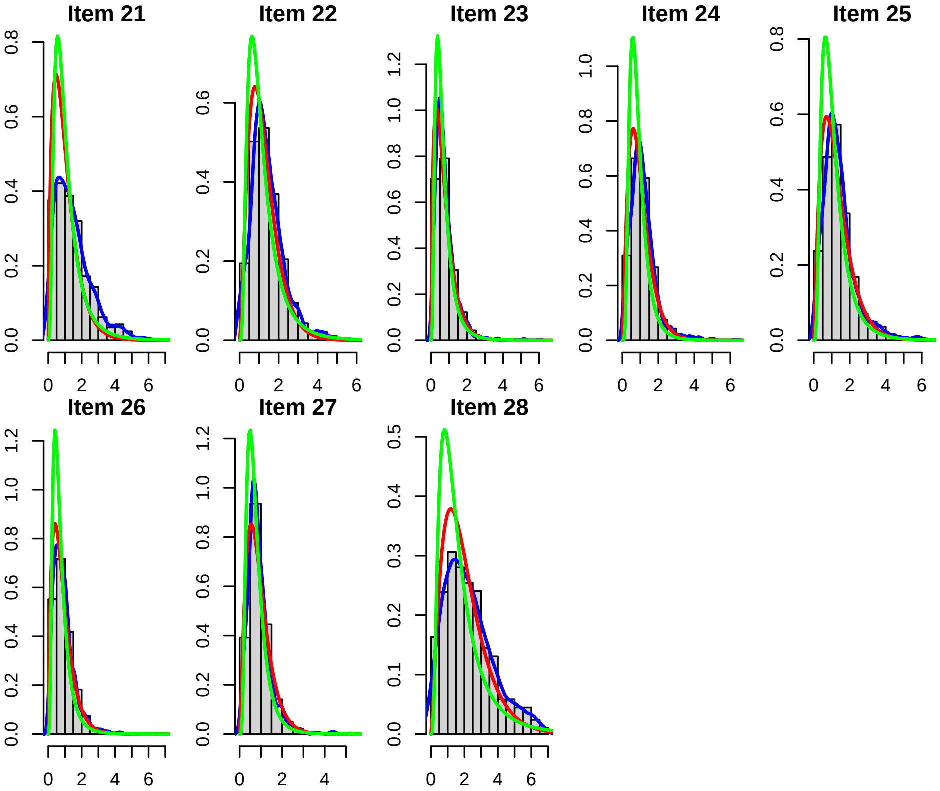


Figure 5: Probability Histogram and Fitted Density Curves (Empirical Density, Gamma Model, and Log-Normal Model) for Response Time Data (in Minutes) for Items 21-28.