Psychometrika (2026), **91**, 669–694
doi:10.1017/psy.2026.10093

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Uncovering Latent Structures: A Bayesian Approach to Estimating Q-Matrix and Attribute Hierarchies in Cognitive Diagnostic Models

Xue Wang[1], Yinghan Chen[2] and Shiyu Wang[3]

[1]Key Laboratory of Applied Statistics of MOE, Key Laboratory of Big Data Analysis of Jilin Province, School of Mathematics and Statistics, Northeast Normal University, Changchun, Jilin, China; [2]Department of Mathematics and Statistics, University of Nevada Reno, USA; [3]Department of Educational Psychology, University of Georgia, Athens, GA, USA

Corresponding author: Shiyu Wang; Email: swang44@uga.edu

(Received 15 May 2025; revised 27 January 2026; accepted 2 February 2026)

Abstract

Cognitive diagnostic models (CDMs) provide fine-grained diagnostic feedback by modeling the relationship between latent attributes and item responses. Two key components required for CDM implementation are the Q-matrix, which links items to attributes, and the attribute hierarchy, which defines prerequisite relationships among attributes. In many practical settings, both structures are specified by experts based on cognitive theory. In this article, we propose a novel Bayesian estimation method that simultaneously learns the Q-matrix, the attribute hierarchy, and the parameters of the deterministic inputs, noisy "and" gate (DINA) model. We develop a Metropolis–Hastings within Gibbs algorithm and integrate a mini-batch strategy to improve computational efficiency. We conducted a series of simulation studies to evaluate the performance of the proposed algorithm under varying conditions, including sample size, test length, hierarchy structure, mini-batch size, and threshold settings. The results demonstrate strong recovery rates for both latent structures and item parameters, confirming the accuracy and robustness of our method. A real data application further illustrates the utility of the proposed framework in uncovering interpretable diagnostic structures. Our findings offer practical guidance for researchers seeking to implement CDMs when both the Q-matrix and attribute hierarchy are unknown.

Keywords: Bayesian method; cognitive diagnosis models; hierarchical structure; Q-matrix

1. Introduction

Cognitive diagnostic models (CDMs) are a family of restricted latent class models that describe how a person's underlying discrete latent traits, in conjunction with item characteristics, influence their responses to test items. CDMs provide fine-grained diagnostic information about whether an individual has mastered specific skills or attributes. This detailed feedback makes CDMs particularly valuable in a wide range of applications across education, behavioral sciences, and social sciences (Chen et al., 2015; De La Torre & Douglas, 2004, 2008; Tatsuoka, 1984; Von Davier, 2008).

As a class of confirmatory psychometric models, CDMs require several key components to be specified prior to application to a dataset. Two essential latent structures must be defined. The first is the Q-matrix (Tatsuoka, 1983), which specifies the relationship between items and the attributes they measure. The second structure concerns the relationships among the latent attributes themselves. Although many earlier studies did not impose any assumptions about hierarchical structures among

attributes, accumulating evidence suggests that hierarchical dependencies are frequently present in practice. For example, in mathematics, multiplication is essentially a simplified form of addition, such as $2 \times 4 = 2 + 2 + 2 + 2 = 8$. Therefore, mastering addition is widely regarded as a fundamental prerequisite for mastering multiplication. Recognizing such dependencies, scholars have explored ways to incorporate attribute hierarchies into CDMs. As a result, a growing body of methodological research has aimed to formally integrate hierarchical structures—where mastery of one attribute is required before another—into CDMs (e.g., Leighton et al., 2004; Ma & Xu, 2022; Templin & Bradshaw, 2014).

In many CDM-related studies, the Q-matrix and attribute hierarchy are often fully or partially specified by test developers based on cognitive theory or expert judgment. More recently, data-driven approaches have emerged that estimate the Q-matrix, attribute hierarchy, or both directly from examinees' response data. The literature on Q-matrix estimation can be broadly divided into two categories: validation and refinement of Q-matrix with partial knowledge (Chiu, 2013; DeCarlo, 2012; de la Torre & Chiu, 2016; Templin & Henson, 2006; Wang et al., 2020), and completely data-driven estimation of unknown Q-matrix (Chen et al., 2015; Chen et al., 2018; Chung, 2019; Liu et al., 2012, 2013; Xu & Shang, 2018). The estimation methods include both frequentist (Chen et al., 2015; Liu et al., 2012, 2013; Xiang, 2013) and Bayesian approaches (Chen et al., 2018; Chung, 2019; Culpepper, 2019; Templin & Henson, 2006).

For the estimation of attribute hierarchy structures, Wang and Lu (2021) introduced two exploratory methods using regularization techniques to learn hierarchies directly from data. Liu et al. (2022) proposed a z-statistic approach to assess the significance of structural parameter estimates, and Zhang et al. (2024) refined this method for improved accuracy. Moreover, Lee and Gu (2024) proposed the latent conjunctive Bayesian networks (LCBNs) to unify the attribute hierarchy and the Bayesian network, and developed a two-step EM algorithm to estimate the hierarchy structure and item parameters. Assuming a known Q-matrix, Chen and Wang (2023) developed a Bayesian framework for estimating attribute hierarchies using a Metropolis-within-Gibbs sampler.

A few studies have sought to estimate both the Q-matrix and the attribute hierarchy jointly. For example, Ma et al. (2023) proposed a penalized likelihood approach to first select significant latent classes and item parameter structures, then recover the number of attributes, the attribute hierarchy, the Q-matrix, and item-level models. These developments reflect a growing interest in flexible, data-driven frameworks for cognitive diagnosis modeling.

The goal of this study is to address the problem of simultaneously learning the Q-matrix, attribute hierarchy, and CDM parameters from examinees' item response data. Building on the deterministic inputs, noisy "and" gate (DINA) model, we propose a Bayesian estimation method that jointly estimates the attribute hierarchy, Q-matrix, item parameters, and examinees' latent attribute profiles—without assuming prior knowledge of either the Q-matrix or the attribute hierarchy. Our approach extends the work of Chen and Wang (2023) on attribute hierarchy learning and Chung (2019) on Q-matrix estimation by developing a Metropolis–Hastings within Gibbs (MH-Gibbs) algorithm to estimate all parameters in a unified framework. To improve computational efficiency, we further incorporate a mini-batch strategy, enabling faster estimation while maintaining accuracy. Moreover, compared with Ma et al. (2023), which adopts a two-step procedure to estimate the attribute hierarchy and the Q-matrix, our method allows for the simultaneous estimation of all parameters in a unified framework. Moreover, unlike deterministic methods, our Bayesian approach summarizes the estimation of Q and G based on posterior samples. This enables a more flexible, theory-driven validation process, where experts can evaluate the plausibility of multiple high-probability structures informed by domain knowledge, rather than relying solely on a single point estimate.

In the remainder of this article, we first present the problem setup in Section 2. Section 3 introduces our proposed Bayesian estimation framework and details the development of an MH-Gibbs algorithm for jointly estimating the attribute hierarchy, Q-matrix, item parameters, and examinees' latent attribute profiles under the DINA model. In Section 4, we provide further discussions on several issues of implementing the algorithm to obtain final point estimation of model parameters, including the selection of mini-batch sample size, the potential label-switching issues, and the determination of cut-off

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

values for finalizing the latent structures. In Section 5, we conduct a comprehensive simulation study to evaluate the performance of the proposed algorithm and investigate the impact of various data conditions. Section 6 applies the proposed method to a real dataset to further demonstrate its practical utility. Finally, Section 7 concludes the article with a summary of key findings and outlines directions for future research.

2. Preliminaries: Model setup

2.1. DINA model

To introduce DINA model formulation, let i $(i = 1,\ldots,N)$ denote examinees, j $(j = 1,\ldots,J)$ denote items, and k $(k = 1,\ldots,K)$ denote attributes. Whether examinee i masters the attribute k is denoted as a binary variable α_{ik}, where $\alpha_{ik} = 1$ if examinee i has mastered the attribute k and $\alpha_{ik} = 0$ otherwise. We use a K-dimensional binary vector, $\boldsymbol{\alpha}_i = (\alpha_{i1},\ldots,\alpha_{ik},\ldots,\alpha_{iK})^{\mathrm{T}}$, to represent the latent attribute profile of examinee i. Therefore, there are $C = 2^K$ possible attribute profiles. Let class c denote the bijection index of the attribute profile, where $c = 0,\ldots,C-1$.

Furthermore, a $J \times K$-dimensional Q-matrix, $\mathrm{Q} = (\boldsymbol{q}_1,\ldots,\boldsymbol{q}_j,\ldots,\boldsymbol{q}_J)^{\mathrm{T}}$, is introduced to describe which attributes are measured by each item. Here, $\boldsymbol{q}_j^{\mathrm{T}} = (q_{j1},\ldots,q_{jk},\ldots,q_{jK})$ is the j-th row of Q-matrix, corresponding to measured attributes by item j with $q_{ik} = 1$ if the attribute k is required by item j, and $q_{ik} = 0$ otherwise. Let Y_{ij} denote the binary item response variable for the examinee i to the item j, where Y_{ij} equals 1 if the ith examinee correctly answered the item j, and 0 otherwise.

The DINA model assumes an examinee need to master all the required attributes for item j to have a high chance to answer that item correctly. To reflect this assumption, the ideal response variable $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is defined to formulate the item response model. Here, $\eta_{ij} = 1$ indicates that examinee i possesses all the required attributes for item j; otherwise, $\eta_{ij} = 0$. In addition, two item parameters are introduced for each item j: the slipping parameter s_j and the guessing parameter g_j which can be formally defined by

$$s_j = P(Y_{ij} = 0|\eta_{ij} = 1),\ g_j = P(Y_{ij} = 1|\eta_{ij} = 0). \tag{1}$$

Then, the probability of a correct response for the DINA can be expressed as

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, g_j, s_j) = g_j^{1-\eta_{cj}}(1-s_j)^{\eta_{cj}}, \tag{2}$$

where $\boldsymbol{\alpha}_c$ denotes one possible attribute profile value and η_{cj} is the corresponding ideal response variable.

2.2. Attribute hierarchy and Q-matrix

One type of latent structure with CDM is the attribute hierarchy, which may be commonly considered when defining attributes within many educational and psychological tests. Based on the nature of assessment, attributes may have different types of dependency, where one attribute may be a prerequisite for another attribute. Four types of popular hierarchical structure are discussed by Leighton et al. (2004), which are presented as four directed acyclic graphs (DAGs) in Figure 1. For a DAG, each node represents an attribute, and a directed edge between two nodes indicates a prerequisite relationship, where the attribute represented by the source node is required for mastering the attribute represented by the target node. For instance, with the linear hierarchy represented by Figure 1(a), the directed edge $1 \to 2$ indicates that the mastery of α_1 is a prerequisite for mastering α_2.

In this article, we adopt the same notation for describing the attribute hierarchy as Chen and Wang (2023). Let V and E denote the set of attributes and the directed edges, respectively, and $G(V;E)$ denote the attribute hierarchy structure. Mathematically, one $G(V;E)$ can be presented by its corresponding adjacency matrix, G. For example, let $G(V;E)$ represent the linear hierarchy in Figure 1(a). With this structure, $V = \{1,2,3,4\}$ and $E = \{1 \to 2, 2 \to 3, 3 \to 4\}$. This particular $G(V;E)$ can be denoted by its

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

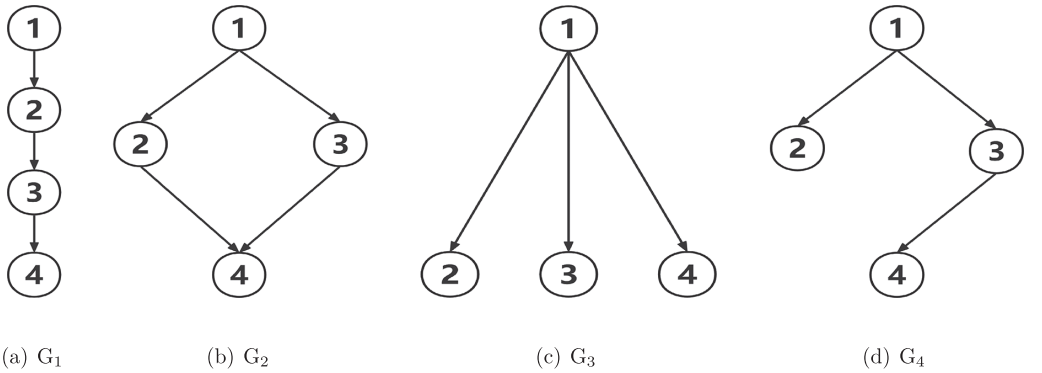

Figure 1. Four attribute hierarchy structures, arranged from left to right: (a) G_1 Linear; (b) G_2 Convergent; (c) G_3 Divergent; and (d) G_4 Unstructured.

adjacency matrix G as

$$\mathrm{G} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\mathrm{G}(k,k') = 1$ represents there is a directed edge $k \to k'$ and 0 otherwise. When there is no hierarchical structure between any attributes, all elements in G are 0, and we denote this structure as G_{null}. Because of this one-to-one correspondence between a DAG and its adjacency matrix G, we will just use the adjacency matrix G to denote its corresponding hierarchical structure $G(V,E)$.

In addition to the direct relationships discussed above, there are also indirect dependencies among attributes. To better capture both direct and indirect dependencies between attributes, the reachability matrix R is introduced, where $\mathrm{R}(k,k') = 1$ represents that there exists a path from k to k', and it can be obtained through Boolean addition and multiplication of the adjacency matrix (Tatsuoka, 2012). The reachability matrix R of Figure 1(a) can be expressed as

$$\mathrm{R} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that different hierarchical structures may lead to the same reachability matrix. For example, consider the structure G^*, corresponding to $V^* = \{1,2,3,4\}$ and $E^* = \{1 \to 2, 2 \to 3, 3 \to 4, 1 \to 4\}$. The reachability matrix of G^* will be the same as that of G for Figure 1(a), meaning that both exhibit the same dependencies between attributes. However, it is clear that the structure of G in Figure 1(a) is simpler. Therefore, our proposed Bayesian estimation method estimates the transitive reduction of G, which refers to a subgraph of G with the fewest edges that maintains the same reachability as G (Chen & Wang, 2023).

2.2.1. The impact of G to latent attribute profile distribution

Typically when assuming no attribute hierarchy among K attributes, that is, under G_{null}, there are $C = 2^K$ possible attribute profiles. With a specific attribute hierarchical structure, the number of possible attribute profiles is reduced from 2^K. For example, consider the linear structure G_1 in Figure 1(a), the attribute profile $(1,0,1,0)$ violates the relationship that α_2 is the prerequisite of α_3, thus cannot exist with G_1. Therefore, the attribute profile space depends on the specific hierarchical structure G. Denote C_G the

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

number of permissible attribute profiles under G. Here, we provide a definition of an equivalence class. Let $\boldsymbol{\alpha}_c \overset{\mathrm{G}}{=} \boldsymbol{\alpha}_{\tilde{c}}$ indicate that $\boldsymbol{\alpha}_c$ and $\boldsymbol{\alpha}_{\tilde{c}}$ are equivalent under structure G. Specifically, if a mastered attribute (i.e., an attribute with a value of 1) has a prerequisite that has not been mastered, it will be reassigned to 0. For example, as mentioned above, the attribute profile $(1,0,1,0)$ is not permissible under linear structure G_1 and is thus reassigned to (1,0,0,0), that is, $(1,0,1,0) \overset{\mathrm{G}_1}{=} (1,0,0,0)$.

2.2.2. Q-matrix under G

Another important latent structure under CDM is the Q-matrix. When attribute hierarchy exists, we assume that the hierarchical structure also affects the space of the Q-matrix. In other words, the $\boldsymbol{q}$-vector needs to reflect the underlying attribute hierarchy as well. For example, considering the linear structure G_1, we need to make the notations consistent in Figure 1(a), $\mathbf{q} = (0,1,0,0)$ equivalent to $\mathbf{q} = (1,1,0,0)$, indicating that an item measuring α_2 also needs to measure α_1 under this linear structure. Therefore, we introduce reduced Q-matrix Q_r(Tatsuoka, 2012), which including all possible $\boldsymbol{q}$-vectors under the hierarchical structure. In this case, the total $2^K - 1$ possible $\boldsymbol{q}$-vectors under G_{null} are reduced to four possible $\boldsymbol{q}$-vectors under the linear structure G_1. In this study, a structured Q-matrix constructed from the possible $\boldsymbol{q}$-vectors in $\mathrm{Q}_r(\mathrm{G})$ is used to reflect the attribute hierarchical structure. The above relationship is shown by Equation (3), where each column in a Q-matrix represents a possible $\boldsymbol{q}$-vector:

$$\mathrm{Q}_{all}^{\mathrm{T}} = \begin{bmatrix} 1&0&0&0&1&1&1&0&0&0&1&1&1&0&1 \\ 0&1&0&0&1&0&0&1&1&0&1&1&0&1&1 \\ 0&0&1&0&0&1&0&1&0&1&1&0&1&1&1 \\ 0&0&0&1&0&0&1&0&1&1&0&1&1&1&1 \end{bmatrix} \overset{reduced}{\longrightarrow} \begin{pmatrix} 1&1&1&1 \\ 0&1&1&1 \\ 0&0&1&1 \\ 0&0&0&1 \end{pmatrix} \overset{\triangle}{=} \mathrm{Q}_r(\mathrm{G}_1)^{\mathrm{T}}. \tag{3}$$

3. Bayesian estimation

3.1. Bayesian formulation

To start with, denote item response matrix for N examinees across J items as $\mathbf{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_N)^{\mathrm{T}}$, where $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{iJ})^{\mathrm{T}}$, $i = 1, \ldots, N$, is the response vector of examinee i to J items. In addition, let $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_N)^{\mathrm{T}}$ be the $N \times K$ matrix of attribute profiles of all examinees. Denote the parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c, \ldots, \pi_C)$ as the latent class probability vector, that is, $\pi_c = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)$, satisfying $\sum \pi_c = 1$. Let $\boldsymbol{s} = (s_1, \ldots, s_J)^T$ and $\boldsymbol{g} = (g_1, \ldots, g_J)^T$ denote the slipping and guessing parameters for all items.

Given the conditional independent assumption, the joint distribution of $\mathbf{Y}$ given parameters $\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}, \boldsymbol{A}, \mathrm{Q}, \mathrm{G}$ can be written as

$$p(\mathbf{Y}|\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}, \boldsymbol{A}, \mathrm{Q}, \mathrm{G}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \sum_{c=1}^{C} \boldsymbol{\pi}_c \, p(Y_{ij}|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, g_j, s_j, \boldsymbol{q}_j, \mathrm{G}), \tag{4}$$

where

$$\begin{aligned} p(Y_{ij}|\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, g_j, s_j, \boldsymbol{q}_j, \mathrm{G}) &= \left(g_j^{1-\eta_{cj}}(1-s_j)^{\eta_{cj}}\right)^{y_{ij}} \left((1-g_j)^{1-\eta_{cj}} s_j^{\eta_{cj}}\right)^{1-y_{ij}}, \\ \eta_{cj} &= \prod_{k=1}^{K} \alpha_{ck}^{q_{jk}}. \end{aligned} \tag{5}$$

To estimate the DINA model item parameters $\boldsymbol{s}$ and $\boldsymbol{g}$, the examinee parameters $\boldsymbol{A}$ and $\boldsymbol{\pi}$, and the two latent structures G and Q, we first outline the joint posterior distribution of all model parameters.

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Let $p(s_j,g_j)$, $p(\boldsymbol{\alpha}_i \mid \boldsymbol{\pi},\mathrm{G})$, $p(\boldsymbol{\pi} \mid \mathrm{G})$, $p(\mathrm{Q} \mid \mathrm{G})$, and $p(\mathrm{G})$ be the corresponding prior of (s_j,g_j), $\boldsymbol{\alpha}_i$, $\boldsymbol{\pi}$, Q, and G, respectively. The joint posterior distribution of all model parameters is

$$\begin{aligned} & p(\boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G} \mid \mathbf{Y}) \\ & \quad \propto p(\mathbf{Y} \mid \boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G}) p(\boldsymbol{s},\boldsymbol{g})\, p(\mathbf{A} \mid \boldsymbol{\pi},\mathrm{G})\, p(\boldsymbol{\pi} \mid \mathrm{G})\, p(\mathrm{Q} \mid \mathrm{G})\, p(\mathrm{G}) \\ & \quad \propto p(\mathbf{Y} \mid \boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G}) \left(\prod_{j=1}^{J} p(s_j,g_j) \right) \left(\prod_{i=1}^{N} p(\boldsymbol{\alpha}_i \mid \boldsymbol{\pi},\mathrm{G}) \right) p(\boldsymbol{\pi} \mid \mathrm{G})\, p(\mathrm{Q} \mid \mathrm{G})\, p(\mathrm{G}), \end{aligned} \tag{6}$$

where $p(\mathbf{Y} \mid \boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G})$ is the joint distribution function (4). In the following section, we specify the prior distributions of each parameter, followed by the inference of their posterior distributions.

3.2. Posterior inference

Firstly, the prior of s_j and g_j follows the Beta distribution with parameter a_{s0}, b_{s0},a_{g0}, and b_{g0}:

$$p(s_j,g_j) \propto s_j^{a_{s0}-1}(1-s_j)^{b_{s0}-1} g_j^{a_{g0}-1}(1-g_j)^{b_{g0}-1} \mathcal{I}(0 < s_j + g_j < 1). \tag{7}$$

Then we can derive that posterior distribution of s_j is a Beta distribution with parameter a_{sj} and b_{sj} as following:

$$s_j \mid \mathbf{Y},\boldsymbol{g},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G} \sim \mathrm{Beta}(a_{sj},b_{sj}) \mathcal{I}(0 < s_j < 1 - g_j), \tag{8}$$

where

$$a_{sj} = \sum_{i=1}^{N} \eta_{ij}(1-y_{ij}), \quad b_{sj} = \sum_{i=1}^{N} \eta_{ij} y_{ij}.$$

And the posterior distribution of g_j is a Beta distribution with parameter a_{gj} and b_{gj} as following:

$$g_j \mid \mathbf{Y},\boldsymbol{s},\boldsymbol{\pi},\boldsymbol{A},\mathrm{Q},\mathrm{G} \sim \mathrm{Beta}(a_{gj},b_{gj}) \mathcal{I}(0 < g_j < 1 - s_j), \tag{9}$$

where

$$a_{gj} = \sum_{i=1}^{N} (1-\eta_{ij}) y_{ij}, \quad b_{gj} = \sum_{i=1}^{N} (1-\eta_{ij})(1-y_{ij}).$$

Following Culpepper (2015), we assign Beta(1,1) priors to both s_j and g_j in this article, that is, $a_{s0} = b_{s0} = a_{g0} = b_{g0} = 1$.

Secondly, the prior of $\boldsymbol{\pi}$ under structure G_{null} is a Dirichlet distribution with parameter $\boldsymbol{\delta}_0$:

$$p(\boldsymbol{\pi} \mid \boldsymbol{\delta}_0) \propto \prod_{c=0}^{C-1} \pi_c^{\delta_{0c}}, 0 \le \pi_c \le 1, \sum \pi_c = 1. \tag{10}$$

However, as discussed in Section 2.2, the existence of a hierarchical structure among attributes imposes constraints on the attribute profiles. Specifically, some attribute profiles become impermissible under the hierarchy structure G. This restriction reduces the set of permissible attribute profiles, thereby influencing the distribution of $\boldsymbol{\pi}$. Moreover, since G is allowed to change during the estimation process, the distributions of $\boldsymbol{\alpha}_i$ also change accordingly. Therefore, when incorporating a hierarchical structure G, it is necessary to adjust the corresponding priors to reflect the structural constraints imposed by G.

Let L_G denote the set of permissible attribute profiles under a specific hierarchical structure G and $C_G = |L_G|$ denote the number of permissible attribute profiles. Define $\boldsymbol{\pi}_G$ as the current latent class probability vector. Thus, the prior of $\boldsymbol{\pi}_G$ now follows an aggregated Dirichlet distribution with parameters $\boldsymbol{\delta}_0^G$:

$$p(\boldsymbol{\pi}_G \mid \boldsymbol{\delta}_0^G) \propto \prod_{\boldsymbol{\alpha}_c \in L_G} \pi_c^{\delta_{0c}^G}, \tag{11}$$

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

where

$$\pi_{\tilde{c}} = \sum_{\boldsymbol{\alpha}_c \overset{G}{=} \boldsymbol{\alpha}_{\tilde{c}}; c \in C} \pi_c, \quad \delta_{0\tilde{c}}^G = \sum_{\boldsymbol{\alpha}_c \overset{G}{=} \boldsymbol{\alpha}_{\tilde{c}}; c \in C} \delta_c.$$

Based on the prior in Eq. (11) and the joint posterior distribution in Eq. (6), we can derive that posterior distribution of $\boldsymbol{\pi}_G$ is a Dirichlet distribution with parameter $\boldsymbol{\delta}_G$ as follows:

$$\boldsymbol{\pi}_G \mid \mathbf{Y}, \boldsymbol{s}, \boldsymbol{g}, \mathbf{A}, \mathrm{Q}, \mathrm{G} \sim \mathrm{Dirichlet}(\boldsymbol{\delta}_G), \tag{12}$$

where

$$\delta_{\tilde{c}}^G = \sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{\tilde{c}}) + \delta_{0\tilde{c}}^G.$$

We set $\boldsymbol{\delta}_0 = (1, \ldots, 1)_L$ in our study, following the setting in Culpepper (2015).

Thirdly, the prior of $\boldsymbol{\alpha}_i$ follows a categorical distribution with parameter $\boldsymbol{\pi}$ under structure G_{null}:

$$p(\boldsymbol{\alpha}_i | \boldsymbol{\pi}) = \prod \pi_c^{\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)}, c = 0, \ldots, C-1. \tag{13}$$

Similar to parameter $\boldsymbol{\pi}$, the prior of $\boldsymbol{\alpha}_i$ also needs to be adjusted according to the hierarchy structure G. Accordingly, the prior of $\boldsymbol{\alpha}_i$ follows a collapsed categorical distribution with parameter $\boldsymbol{\pi}_G$ under structure G:

$$p(\boldsymbol{\alpha}_i | \boldsymbol{\pi}_G) = \prod_{\tilde{c} \in L_G} \pi_{\tilde{c}}^{\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{\tilde{c}})}. \tag{14}$$

Based on the prior in Eq. (14) and the joint posterior distribution in Eq. (6), we can derive that posterior distribution of $\boldsymbol{\alpha}_i$ is a categorical distribution with parameter $\boldsymbol{\rho}_i^G = (\rho_{i0}^G, \ldots, \rho_{iC_G-1}^G)$ as follows:

$$\boldsymbol{\alpha}_i \mid \mathbf{Y}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{\pi}_G, \mathrm{Q}, \mathrm{G} \sim \mathrm{Categorical}(\boldsymbol{\rho}_i^G), \tag{15}$$

where

$$\rho_{i\tilde{c}}^G \propto \pi_{\tilde{c}}^G \prod_{j=1}^{J} \left(g_j^{1-\eta_{\tilde{c}j}} (1-s_j)^{\eta_{\tilde{c}j}} \right)^{y_{ij}} \left((1-g_j)^{1-\eta_{\tilde{c}j}} s_j^{\eta_{\tilde{c}j}} \right)^{1-y_{ij}}.$$

Fourthly, as for the prior of the Q-matrix, we adopted the approach in Chung (2019) by introducing additional parameters for each Q-matrix entry and made corresponding adjustments. With K attributes, there are $H = 2^K - 1$ possible $\boldsymbol{q}$-vector patterns for each item under structure G_{null}, which is denoted as $\boldsymbol{\varepsilon}_{H \times K} = (\varepsilon_{hk})_{H \times K}$. Suppose $\boldsymbol{q}_j$ following a categorical distribution with parameter $\boldsymbol{\theta}_{0j} = (\theta_{0j1}, \ldots, \theta_{0jH})$:

$$p(\boldsymbol{q}_j \mid \boldsymbol{\theta}_{0j}) \propto \prod_{h=1}^{H} \theta_{0jh}^{\mathcal{I}(\boldsymbol{q}_j = \boldsymbol{\varepsilon}_h)}, \tag{16}$$

$$\boldsymbol{\theta}_{0j} = g(\boldsymbol{\Phi}_j), \quad \boldsymbol{\Phi}_j = (\boldsymbol{\phi}_{j1}, \ldots, \boldsymbol{\phi}_{jH})^{\mathrm{T}},$$

where $\boldsymbol{\phi}_{jh} = (\phi_{jh1}, \ldots, \phi_{jhK}), h \in (1, \ldots, H)$. We define $\phi_{jhk} = P(q_{jk} = 1)$ and the prior for ϕ_{jhk} is

$$\phi_{jhk} \sim \mathrm{Beta}(1, 1).$$

Therefore, the conditional posterior for each element ϕ_{jhk} is distributed as

$$\phi_{jhk} \mid q_{jk} \sim \mathrm{Beta}\left(1 + q_{jk}, 2 - q_{jk}\right). \tag{17}$$

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Moreover, the prior for $\boldsymbol{\theta}_{0j}$ is then modeled as

$$\begin{aligned}\boldsymbol{\theta}_{0j} = g(\boldsymbol{\Phi}_j) = (g(\boldsymbol{\phi}_{j1}),\ldots,g(\boldsymbol{\phi}_{jH}) \propto \Bigg(&\prod_{k=1}^{K}\phi_{j1k}^{\varepsilon_{1k}}\left(1-\phi_{j1k}\right)^{1-\varepsilon_{1k}},\ldots,\prod_{k=1}^{K}\phi_{jhk}^{\varepsilon_{hk}}\\ &\times\left(1-\phi_{jhk}\right)^{1-\varepsilon_{hk}},\ldots,\prod_{k=1}^{K}\phi_{jHk}^{\varepsilon_{Hk}}\left(1-\phi_{jHk}\right)^{1-\varepsilon_{Hk}}\Bigg).\end{aligned} \tag{18}$$

Considering that in the DINA model, restricting $\boldsymbol{\alpha}$ is equivalent to restricting the Q-matrix, we do not impose any additional structural constraints on the Q-matrix during its update process. Therefore, based on the prior in Eq. (18) and the joint posterior distribution in Eq. (6), we can derive that posterior distribution of $\boldsymbol{q}_j$ is a categorical distribution with parameter $\boldsymbol{\theta}_j = (\theta_{j1},\ldots,\theta_{jH})$ as follows:

$$\boldsymbol{q}_j \mid \mathbf{Y},\boldsymbol{s},\boldsymbol{g},\boldsymbol{\pi},\mathrm{G} \sim \mathrm{Categorical}(\boldsymbol{\theta}_j), \tag{19}$$

where

$$\theta_{jh} = \frac{\gamma_{jh}}{\sum_{h=1}^{H}\gamma_{jh}}, \tag{20}$$

$$\gamma_{jh} = p(\boldsymbol{\phi}_{jh})\prod_{i=1}^{N}\left(g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}\right)^{y_{ij}}\left((1-g_j)^{1-\eta_{ij}}s_j^{\eta_{ij}}\right)^{1-y_{ij}}. \tag{21}$$

Finally, for the hierarchical structure G, we adopt a uniform prior, which means for any two possible transitive reduction structure G and G^*, we have $p(\mathrm{G}) = p(\mathrm{G}^*)$. As we are not able to get the closed form of the posterior distribution of G, similar to Chen and Wang (2023), the MH sampling method is utilized to update G. First, we need to sample a new structure G^*. Note that the new G^* can only be obtained by either removing an existing edge or adding an edge to the current structure G. Denote p as the probability of adding an edge, indicating $1-p$ as the probability of removing an edge. We first use the probability p to determine whether the new structure G^* involves adding or removing an edge, and then randomly generate a possible G^*. It is worth to note that this sampling procedure only exists between transitive reduction Gs, meaning that each time a new G^* is sampled, it must have a different reachability matrix from G. According to this rule, each time an edge is added, it must be selected from the edges between two attributes that do not have any direct or indirect relationship. Such a pair of attributes can be found through 0s in the off-diagonal entries in the reachability matrix. Let $T\left(\mathrm{G},\mathrm{G}^*\right)$ denote the transition probability from G to the proposed new G^*. Define

$$T\left(\mathrm{G},\mathrm{G}^*\right) = \frac{1}{2\times \text{ number of pairs of non-connected nodes}}, \tag{22}$$

when adding an edge, and

$$T\left(\mathrm{G},\mathrm{G}^*\right) = \frac{1}{\text{number of existing edges}}, \tag{23}$$

when removing an edge. Then the accept ratio of G^* is

$$\begin{aligned} r &= \min\left\{1,\frac{p\left(\mathrm{G}^*\mid\mathbf{Y},\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathbf{Q}\right)T\left(\mathrm{G}^*,\mathrm{G}\right)}{p\left(\mathrm{G}\mid\mathbf{Y},\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathbf{Q}\right)T\left(\mathrm{G},\mathrm{G}^*\right)}\right\}\\ &= \min\left\{1,\frac{p\left(\mathbf{Y}\mid\mathrm{G}^*,\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathrm{Q}\right)p\left(\mathrm{G}^*\right)T\left(\mathrm{G}^*,\mathrm{G}\right)}{p(\mathbf{Y}\mid\mathrm{G},\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathrm{Q})p(G)T\left(\mathrm{G},\mathrm{G}^*\right)}\right\}\\ &= \min\left\{1,\frac{p\left(\mathbf{Y}\mid\mathrm{G}^*,\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathbf{Q}\right)T\left(\mathrm{G}^*,\mathrm{G}\right)}{p(\mathbf{Y}\mid\mathrm{G},\boldsymbol{s},\boldsymbol{g},\mathbf{A},\mathrm{Q})T\left(\mathrm{G},\mathrm{G}^*\right)}\right\}. \end{aligned} \tag{24}$$

Then, a random number is drawn from a uniform distribution and compared to the acceptance ratio r to determine whether to accept the proposed G^*.

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 1. MH-Gibbs sampling procedure

Algorithm
Input: $a_{s0}, b_{s0}, a_{g0}, b_{g0}, \boldsymbol{\delta}_0, T$
Initialization: $\mathbf{A}^{(0)}, \mathrm{Q}^{(0)}, \mathrm{G}^{(0)}$
for t in 1:T do
(1) Update $\boldsymbol{s}^{(t)}$ and $\boldsymbol{g}^{(t)}$: Sample $s_j^{(t)}$ and $g_j^{(t)}$ using Eq. (8) and Eq. (9) respectively
(2) Update $\boldsymbol{\pi}^{(t)}$: Sample $\boldsymbol{\pi}^{(t)}$ using Eq. (12)
(3) Update $\mathbf{A}^{(t)}$: Sample $\boldsymbol{\alpha}_i^{(t)}$ using Eq. (15)
(4) Update $\mathbf{Q}^{(t)}$: Sample $\mathrm{Q}^{(t)}$ using Eq. (19)
(5) Update $\mathbf{G}^{(t)}$: Sample $u \sim U(0,1)$; with pre-specified probability p:
if $u < p$ **do** add an edge to sample G^* **else do** remove an edge to sample G^*
Compute the accept ratio r using Eq. (24) and sample $w \sim U(0,1)$
if $w < r$ **do** accept G^* **else do** stay with G
end for

Our proposed MH within Gibbs sampling procedure is summarized in Table 1, and T represent the length of the Markov chain. The prespecified probability p can be adjusted to improve the acceptance ratio of the newly proposed structure G^*. In this article, we set $p = 0.5$.

4. Mini-batch and final estimation for model parameters

In this section, we summarize the implementation of the proposed algorithm, including the use of mini-batch method in each iteration, addressing the potential label switch issue, and the way to obtain the final estimation for two latent structures and model parameters after the sampling procedure. The overview of the complete procedure from sampling procedure to final parameter estimation is provided in Figure 2. We provided details in the rest of this section.

4.1. Mini-batch method in sampling procedure

We propose to use the mini-batch method, where, during each iteration of the sampling algorithm (Table 1), we randomly select a subset of samples to update $\boldsymbol{s}$, $\boldsymbol{g}$, $\boldsymbol{Q}$, and G. Note that $\boldsymbol{\alpha}$ and $\boldsymbol{\pi}$ cannot be updated using sub-samples, as they are parameters specific to each individual sample. The mini-batch method is a widely used technique in machine learning, commonly applied in stochastic gradient descent. The method works by updating model parameters using only a subset of the samples during each iteration. In recent years, the mini-batch method has also found applications in the Bayesian field, such as in MH (Wu et al., 2022), Gibbs sampling (De Sa et al., 2018), Hamiltonian Monte Carlo (HMC) (Zou & Gu, 2021), and variational Bayes (Hoffman et al., 2013). This approach offers two main advantages. First, it improves the computational efficiency of the algorithm. Second, the mini-batch method introduces some noise, thereby increasing the diversity of the sampling and helping to prevent rapid convergence to local optima.

When implementing mini-batch with our proposed algorithm, one needs to determine the mini-batch size. Common practice in machine learning involves selecting batch sizes that are powers of two, such as 32, 64, 128, or 256, due to their compatibility with CPU and GPU memory architectures (Keskar et al., 2016). Keskar et al. (2016) have discussed that large-batch sample size tend to converge to sharp minimizers, while small-batch sample size consistently converge to flatter minimizers. This suggests that using a larger batch size does not necessarily lead to better estimation performance. In contrast,

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

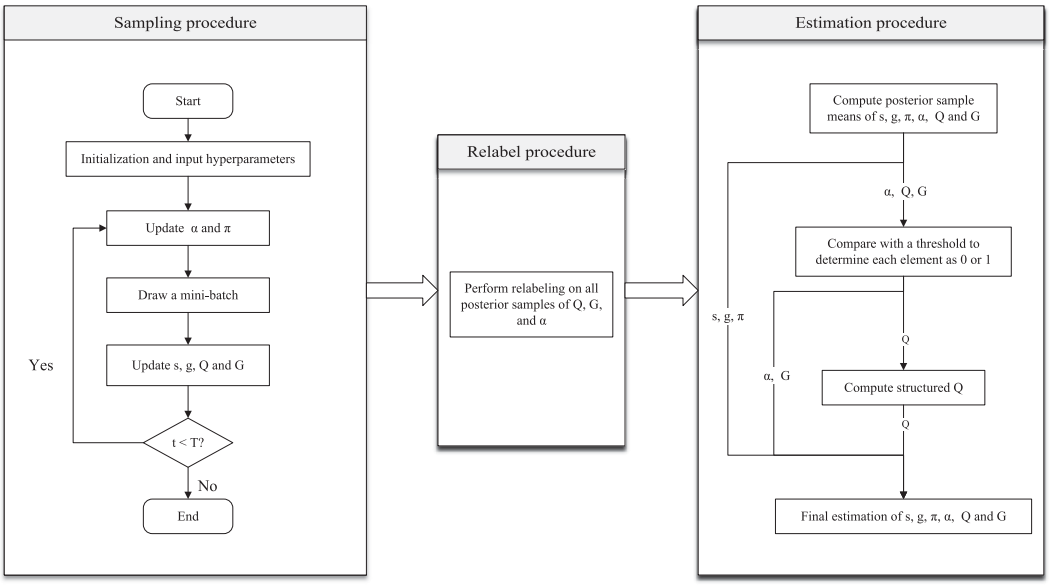

Figure 2. The overview of the complete Bayesian estimation procedure, including the sampling procedure, relabeling procedure, and final parameter estimation.

a relatively smaller batch size may introduce more stochasticity, which can help the algorithm escape from local optima. Given these considerations, we will specifically explore the impact of different mini-batch size to the model estimation results in the simulation section.

4.2. Label switch

Before computing the final estimations from the sampling procedure, it is necessary to address the issue of label switching. In the context of Bayesian methods, label switching is a common concern that may arise both within a single Markov chain and across multiple chains. In this article, the unknown Q-matrix may lead to label switching issues among several attributes, which also impacts $\boldsymbol{\alpha}$ and G. Therefore, it is crucial to impose a relabeling process on the iterative data. Denote the $J \times K \times T$-dimension $\mathcal{Q}$ as the iteration samples of Q. Following the relabeling method in Erosheva and Curtis (2017), we define $\overline{\mathcal{Q}}$ as the average of the sampled $\mathcal{Q}$, which serves as the first reference. For each $\mathrm{Q}^{(t)} \in \mathcal{Q}$, we leverage the Hungarian algorithm to address the relabeling issue. Specifically, we construct a $K \times K$ cost matrix $C^{(t)} = \{c_{kk'}^{(t)}\}$, where

$$c_{kk'}^{(t)} = \left\| \mathbf{q}_{k'}^{(t)} - \overline{\mathbf{q}}_k \right\|_1, \quad k, k' = 1, \ldots, K,\ t = 1, \ldots, T.$$

Let $\mathcal{P}$ denote the set of all permutations of $\{1, \ldots, K\}$. Then, the relabeled $\mathrm{Q}^{(t)}$, denoted as $\mathrm{Q}_R^{(t)}$, is obtained by solving the following assignment problem:

$$p^* = \arg\min_{p \in \mathcal{P}} \sum_{k=1}^{K} c_{k,p(k)}^{(t)}, \tag{25}$$

$$\mathrm{Q}_R^{(t)} = \mathrm{Q}_{p^*}^{(t)}, \quad t = 1, \ldots, T. \tag{26}$$

Note that the above steps need to be repeated. In other words, after obtaining the new relabeled $\mathcal{Q}_R$, we need to calculate the new average matrix $\overline{\mathcal{Q}}_R$ and repeat the relabel procedure until $\mathcal{Q}_R$ converge. In addition, the same permutation must be applied to both $\boldsymbol{\alpha}$ and G simultaneously, as they are also subject to the same relabel issues involved in Q matrix.

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

4.3. Final estimation summary

4.3.1. The estimation of G

The estimation of attribute hierarchy is reflected through the estimation of the adjacency matrix G. We can use the posterior sample mean as a posterior summary statistic for G, denoted as $\hat{\mathrm{G}}$. As a result, each element $\hat{\mathrm{G}}_{kk'}$ takes a value between 0 and 1, which summarizes the probability for a potential edge from attribute k to k'. If an explicit structure is needed for the interpretation of probable hierarchy relationships, a carefully chosen threshold can decide on the final structure. Here, we define a threshold c, referred to as the cut-off value, to make the final estimation of G. Specifically, for each element of $\hat{\mathrm{G}}$, if $\hat{\mathrm{G}}_{kk'} < c$, we set $\hat{\mathrm{G}}_{kk'} = 0$; otherwise, $\hat{\mathrm{G}}_{kk'} = 1$. Chen and Wang (2023) found that for edges that do not exist and have no indirect relationships between attributes, the estimated values are close to 0. For non-existing edges with indirect relationships between attributes, the estimated values range between 0.2 and 0.4. The specific value of the cut-off value will be explored in the simulation studies.

4.3.2. The estimation of Q matrix

After estimating the hierarchical structure, we obtain the estimated Q-matrix through the following two steps. First, similar to the estimation process of $\hat{\mathrm{G}}$, we compute the posterior sample mean of the Q-matrix, denoted as $\hat{\mathrm{Q}}$, and determine each element as 0 or 1 based on a threshold of 0.5. Second, given the estimated hierarchical structure $\hat{\mathrm{G}}$, we construct a structured $\hat{\mathrm{Q}}$ according to the reduced Q-matrix $\mathrm{Q}_r(\hat{\mathrm{G}})$, which serves as the final estimation of the Q-matrix.

4.3.3. The other model parameters

For the parameters $\boldsymbol{s}$, $\boldsymbol{g}$, and $\boldsymbol{\pi}$, we can use their posterior sample means as point estimates. And for the attribute profiles $\boldsymbol{\alpha}$, we compute the posterior means and then dichotomize each element using a threshold of 0.5.

5. Simulation studies

The purpose of this set of simulation studies is to evaluate the performance of the proposed MH-Gibbs algorithm in recovering DINA model parameters across the following factors: 1) sample size and test length, 2) mini-batch size, 3) cutoff values for G, 4) attribute profile distribution, and 5) different attribute hierarchy structures.

5.1. Design

We consider a set up that a total of $K = 4$ attributes were assessed, and we fix the slipping and guessing parameters in a practical level as $s_j = g_j = 0.2$. The summary of the five factors considered and their corresponding level is presented in Table 2. These factors were fully crossed, yielding 40 conditions. We conducted four chains in each replication, with the chain length of 20,000 and the burn in value of 10,000, and each simulation repeated 50 times.

The two test lengths represent relatively short and medium-length assessments. Recognizing that estimating a high-dimensional, unknown Q-matrix necessitates a larger sample size for accuracy, we increased the sample size paired with the increase of test length in our simulation design. This adjustment reflects both practical considerations and the need for a sufficiently large sample size. The four hierarchical structures of the attributes, as well as studied in previous related work (Chen & Wang, 2023), were therefore also considered in our study.

As for the mini-batch size, as noted in Section 3.3, common practice in machine learning involves selecting batch sizes that are powers of two. To balance convergence speed and computational efficiency, we avoided batch sizes that are excessively small or large. Consequently, we chose batch sizes ranging

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 2. Simulation factors and levels

Factor	Levels
Sample size and test length	(N = 500, J = 20)
	(N = 1,000, J = 40)
Minibatch size	N_1 = 128 and 256 for N = 500
	N_1 = 128,256, and 512 for N = 1,000.
Cutoff value	0.2, 0.3, and 0.4
Attribute profile distribution	Balanced and unbalanced
Attribute hierarchy	G_1, G_2, G_3, and G_4 in Figure 1.

from one-quarter to one-half of the total sample size. Specifically, for N = 500 with J = 20, we selected batch sizes of 128 and 256; for N = 1,000 with J = 40, we chose batch sizes of 128, 256, and 512.

Given the findings in Chen and Wang (2023), for edges that do not exist and have no indirect relationships between attributes, the estimated values are close to 0. For non-existing edges with indirect relationships between attributes, the estimated values range between 0.2 and 0.4. Therefore, when selecting the cut-off value, we choose 0.2, 0.3, and 0.4.

Finally, following the design in Chen and Wang (2023), we also considered a balanced and unbalanced attribute profile distribution. For the balanced distribution, the attribute profiles were randomly generated from one of its permissible patterns. For the unbalanced distribution, we first generated $\boldsymbol{\alpha}_i^* = (\alpha_{i1}^*, \ldots, \alpha_{ik}^*, \ldots, \alpha_{iK}^*)^{\mathrm{T}}$ from a multivariate normal distribution $\boldsymbol{\alpha}_i^* \sim \mathrm{N}(0_K, \boldsymbol{\Sigma}_{K\times K})$, where $0_K = (0, \ldots, 0)_{K\times 1}^{\mathrm{T}}$ and

$$\boldsymbol{\Sigma}_{\boldsymbol{K}\times\boldsymbol{K}} = \begin{bmatrix} 1 & \ldots & \sigma \\ \vdots & \ddots & \vdots \\ \sigma & \ldots & 1 \end{bmatrix}_{K\times K},$$

the off-diagonal elements of $\Sigma_{K\times K}$ are σ = 0.5. Then the relationships between the attribute profile $\boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_i^*$ can be expressed as $\alpha_{ik} = 1$ if $\alpha_{ik}^* > 0$ and $\alpha_{ik} = 0$ otherwise.

5.2. Q-matrix design

The establishment of parameter identifiability is a prerequisite for the accurate estimation and reliable interpretation of CDMs. To this end, it is essential to impose a set of constraints on the Q-matrix. Previous studies have discussed the identifiability conditions of parameters in CDMs (Chen et al., 2015; Chen et al., 2018; Gu & Xu, 2023; Ma et al., 2023). In particular, Gu and Xu (2023) explicitly addressed the identifiability conditions when a hierarchical structure exists among attributes and the Q-matrix is unknown. Therefore, in this study, we followed the identifiability conditions proposed by Gu and Xu (2023) when generating the Q-matrix. Specifically, they defined the "densified" version $\mathcal{D}^{\mathrm{G}}(\mathrm{Q})$ and the "sparsified" version $\mathcal{S}^{\mathrm{G}}(\mathrm{Q})$ of the Q-matrix. The "densified" version $\mathcal{D}^{\mathrm{G}}(\mathrm{Q})$ corresponds to what we referred to earlier as the structured Q-matrix in Section 2.2, which is defined as: for any $q_{j,k'} = 1$ and $k \to k'$, set $q_{j,k} = 1$. Conversely, the "sparsified" version $\mathcal{S}^{\mathrm{G}}(\mathrm{Q})$ is defined as: for any $q_{j,k'} = 1$ and $k \to k'$, set $q_{j,k} = 0$. For example, under the structure G_1, given a q-vector $\boldsymbol{q} = (0,1,1,0)$, the "densified" version $\mathcal{D}^{\mathrm{G}}(\boldsymbol{q})$ becomes $(1,1,1,0)$, whereas the "sparsified" version $\mathcal{S}^{\mathrm{G}}(\boldsymbol{q})$ becomes $(0,0,1,0)$. Moreover, under G_1, these three versions are equivalent, that is,

$$(0,1,1,0) \overset{G_1}{\sim} (1,1,1,0) \overset{G_1}{\sim} (0,0,1,0).$$

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Denote the Q-matrix contains the following two components:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}^0_{K\times K} \\ \mathbf{Q}^\star_{(J-K)\times K} \end{pmatrix}.$$

Specifically, we generate the Q-matrix satisfying the following three identifiability conditions: (A) the sub-matrix Q^0 is equivalent to the identity matrix I_K, that is, $\mathrm{Q}^0 \overset{\mathrm{G}}{\sim} I_K$; (B) the $\mathcal{S}^{\mathrm{G}}(\mathrm{Q})$ has at least three ones in each column; and (C) the $\mathcal{D}^{\mathrm{G}}\left(\mathrm{Q}^\star\right)$ contains K distinct column vectors.

Moreover, once we generated Q-matrices that satisfied the identifiability conditions, we uniformly adopted their corresponding "densified" versions, that is, the structured Q-matrices. It is worth noting that during the 50 replications under each simulation condition, we generated 50 distinct Q-matrices. This design may reduce comparability across different conditions. However, it allows us to better evaluate the performance and robustness of our algorithm across different Q-matrices under each condition.

5.3. Evaluation criteria

We evaluate the proposed algorithm through the following perspectives. First, the potential scale reduction factor (PSRF), denoted as $\hat{R}$ (Gelman & Rubin, 1992), was used to evaluate the convergence of the Gibbs algorithm across different simulation conditions. When the $\hat{R}$ is less than 1.1, the Markov chain is considered to have reached convergence. For each replication, four Markov chains were run. The chain that achieved the lowest deviance information criterion (DIC; Spiegelhalter et al., 2002), a widely used model selection criterion in the Bayesian framework, was considered the best and was selected for subsequent analysis.

Second, we used the average bias and root mean squared error (RMSE) to evaluate the accuracy of the estimation for the item parameters $\boldsymbol{s}, \boldsymbol{g}$, and attribute profile membership parameter $\boldsymbol{\pi}$. These evaluation indexes are defined as

$$\mathrm{ABias}(\boldsymbol{s}) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{R}\sum_{r=1}^{R}\left(\hat{s}_j^{(r)} - s_j\right), \quad \mathrm{ARMSE}(\boldsymbol{s}) = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{s}_j^{(r)} - s_j\right)^2},$$

$$\mathrm{ABias}(\boldsymbol{g}) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{R}\sum_{r=1}^{R}\left(\hat{g}_j^{(r)} - g_j\right), \quad \mathrm{ARMSE}(\boldsymbol{g}) = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{g}_j^{(r)} - g_j\right)^2},$$

$$\mathrm{ABias}(\boldsymbol{\pi}) = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\pi}_c^{(r)} - \pi_c\right), \quad \mathrm{ARMSE}(\boldsymbol{\pi}) = \frac{1}{C}\sum_{c=1}^{C}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\pi}_c^{(r)} - \pi_c\right)^2},$$

where R represent the replication times, and $\hat{\cdot}^{(r)}$ represent the estimated value of a parameter in the rth replication. In addition, we provide the 95% highest probability density interval (HPDI) of $\boldsymbol{s}, \boldsymbol{g}$, and $\boldsymbol{\pi}$, which is the narrowest interval that contains 95% of the posterior probability, to access the uncertainty related to these parameter estimation.

To evaluate the recovery of attribute profile estimation, we adopted the pattern-wise agreement rate (PAR), $\mathrm{PAR} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i)$, and the attribute-wise agreement rate (AAR), defined as $\mathrm{AAR}(k) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}(\hat{\alpha}_{ik} = \alpha_{ik})$. Here, $\hat{\boldsymbol{\alpha}}_i$ is the estimated value of $\boldsymbol{\alpha}_i$, and $\hat{\alpha}_{ik}$ is the estimated value of α_{ik}. Higher values of PAR and AAR indicate better recovery of attribute profile.

Third, to assess the recovery of Q-matrix, we calculate the Q-matrix recovery rate for each dimension (QRR) and the average Q-matrix recovery rate (AQRR) across the Q-matrix as follows:

$$\mathrm{QRR(k)} = \frac{1}{R}\sum_{r=1}^{R}\left(1 - \frac{\sum_{j=1}^{J}\left|\hat{q}_{jk}^{(r)} - q_{jk}\right|}{J}\right), \mathrm{AQRR} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{QRR(K)},$$

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

where $r = 1, 2, \ldots, R$ and $\hat{q}_{jk}^{(r)}$ is the estimated value of q_{jk} in the rth replication. The higher value of AQRR and ORR indicates better recovery of Q-matrix. Additionally, we also report the number of times the overall Q-matrix was correctly estimated across 50 replication.

Finally, when evaluating the recovery of the hierarchical structure, we considered four criteria: the true positive rate (TPR) and the true negative rate (TNR), based on the adjacency matrix (G-matrix), as well as the reachability TPR (RTPR) and the reachability TNR (RTNR), based on the reachability matrix (R-matrix):

$$\text{TPR} = \frac{\#\,\text{True positive edge}}{\#\,\text{True positive edge} + \text{False negative edge}},$$

$$\text{TNR} = \frac{\#\,\text{True negative edge}}{\#\,\text{True negative edge} + \text{False positive edge}},$$

$$\text{RTPR} = \frac{\#\,\text{True reachability positive edge}}{\#\,\text{True reachability positive edge} + \text{False reachability negative edge}},$$

$$\text{RTNR} = \frac{\#\,\text{True reachability negative edge}}{\#\,\text{True reachability negative edge} + \text{False reachability positive edge}}.$$

TPR is the proportion of positive edges that are correctly estimated, TNR denotes the proportion of negative edges that are correctly estimated, RTPR measures the proportion of reachability positive edges that are correctly estimated, and RTNR refers to the proportion of reachability negative edges that are correctly estimated. The higher and closer to 1 value of TPR, TNR, RTPR, and RTNR indicate a better recovery of the hierarchical structure G. The number of times the overall G was correctly estimated across 50 replications was also reported.

5.4. Simulation results

5.4.1. Convergence analysis

The results across all simulation conditions consistently indicated that the algorithm converged within 10,000 iterations. We report the simulation condition of $N = 500$, $J = 20$, and $N_1 = 256$ as an example in Appendix A of the Supplementary Material. Therefore, the following results are based the chain that has the smallest DIC among the four of the last 10,000 steps.

5.4.2. Computational time

We report the average computation time (in seconds) across all replications under each simulation condition in Table 3. Algorithm was operated in R programming language (R Core Team, 2017) on a desktop computer equipped with Intel(R) Core(TM) i5-10400 CPU @ 2.90 GHz, 16 GB RAM. The two R packages "Rcpp" (Eddelbuettel & François, 2011) and "RcppArmadillo" (Eddelbuettel & Sanderson, 2014) were used to enhance the computational efficiency. From the result in Table 3, when $N = 500$, it takes within 6 minutes across all conditions, and it takes within 19 minutes across all conditions when $N = 1{,}000$.

5.4.3. The impact of mini-batch sample size and cut-off value of G

We first explore how different levels of mini-batch size and cut off values of G impact the performance of the proposed Gibbs algorithm. The recovery results of the Q-matrix and the hierarchical structure G under different mini-batch sample sizes and cut-off values are documented in Appendix B of the Supplementary Material. Here, we summarize the optimal cut-off value and mini-batch size under various simulation conditions in Table 4. In this table, each cell lists the mini-batch size and cutoff

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 3. Average computation time of the proposed algorithm under each condition in seconds

	N	J	K	N_1	G_1	G_2	G_3	G_4
Balanced	500	20	4	128	165.0952	195.3910	223.8264	208.2667
	500	20	4	256	264.6138	296.7557	339.5586	261.4543
	1,000	40	4	128	586.5280	588.8976	656.1260	592.0493
	1,000	40	4	256	718.5412	687.6035	791.3040	724.0883
	1,000	40	4	512	985.2270	972.3729	1,078.9910	1,020.1100
Unbalanced	500	20	4	128	167.1727	173.9886	200.8386	196.8976
	500	20	4	256	251.8998	273.0060	339.1605	306.8017
	1,000	40	4	128	515.7451	536.0071	678.1804	576.9780
	1,000	40	4	256	626.4555	651.6926	807.8230	706.1459
	1,000	40	4	512	857.5112	898.6855	1,095.6220	967.4907

Table 4. Summary of the optimal mini-bath size and cut-off value

According to the overall recovery of Q				
	$N = 500$		$N = 1{,}000$	
	Balanced	Unbalanced	Balanced	Unbalanced
G_1	$N_1 = 256$ $c = 0.2/0.3$	$N_1 = 256$ $c = 0.3$	$N_1 = 512$ $c = 0.2/0.3$	$N_1 = 512$ $c = 0.2/0.3$
G_2	$N_1 = 256$ $c = 0.2$	$N_1 = 256$ $c = 0.2/0.3$	$N_1 = 256$ $c = 0.2/0.3/0.4$	$N_1 = 512$ $c = 0.2/0.3$
G_3	$N_1 = 128$ $c = 0.2$	$N_1 = 256$ $c = 0.2$	$N_1 = 256$ $c = 0.2/0.3/0.4$	$N_1 = 256/512$ $c = 0.2/0.3/0.4$
G_4	$N_1 = 256$ $c = 0.2/0.3$	$N_1 = 256$ $c = 0.2$	$N_1 = 512$ $c = 0.2/0.3$	$N_1 = 256/512$ $c = 0.2/0.3$
Summary	$N_1 = 128$(1 time)/256(7 times) $c = 0.2$(7 times)/0.3(4 times)		$N_1 = 256$(4 times)/512(6 times) $c = 0.2$(8 times)/0.3(8 times)/0.4(3 times)	
According to the overall recovery of G				
	$N = 500$		$N = 1{,}000$	
	Balanced	Unbalanced	Balanced	Unbalanced
G_1	$N_1 = 256$ $c = 0.3$	$N_1 = 256$ $c = 0.3$	$N_1 = 512$ $c = 0.3$	$N_1 = 512$ $c = 0.2/0.3$
G_2	$N_1 = 256$ $c = 0.2$	$N_1 = 256$ $c = 0.2$	$N_1 = 256$ $c = 0.2/0.3/0.4$	$N_1 = 512$ $c = 0.2/0.3$
G_3	$N_1 = 128/256$ $c = 0.2$	$N_1 = 256$ $c = 0.2$	$N_1 = 256$ $c = 0.2/0.3$	$N_1 = 256/512$ $c = 0.2/0.3$
G_4	$N_1 = 256$ $c = 0.2/0.3$	$N_1 = 256$ $c = 0.2$	$N_1 = 512$ $c = 0.2$	$N_1 = 256/512$ $c = 0.2/0.3$
Summary	$N_1 = 128$(1 time)/256(8 times) $c = 0.2$(6 times)/0.3(3 times)		$N_1 = 256$(4 times)/512(6 times) $c = 0.2$(7 times)/0.3(7 times)/0.4(1 time)	

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 5. Evaluate the recovery accuracy of the adjacency matrix G

		Balanced				
		TPR	TFR	RTPR	RTFR	$\hat{G} = G$
G_1	$N = 500, J = 20$	0.900	0.972	0.986	0.967	44
	$N = 1{,}000, J = 40$	0.927	0.983	0.976	0.960	44
G_2	$N = 500, J = 20$	0.985	0.997	0.996	0.997	49
	$N = 1{,}000, J = 40$	0.910	0.978	0.976	0.969	45
G_3	$N = 500, J = 20$	0.973	0.995	0.989	0.991	48
	$N = 1{,}000, J = 40$	0.947	0.988	0.977	0.982	46
G_4	$N = 500, J = 20$	1.000	1.000	1.000	1.000	50
	$N = 1{,}000, J = 40$	1.000	1.000	1.000	1.000	50
		Unbalanced				
		TPR	TFR	RTPR	RTFR	$\hat{G} = G$
G_1	$N = 500, J = 20$	0.853	0.963	0.978	0.957	40
	$N = 1{,}000, J = 40$	0.987	0.997	0.996	0.993	49
G_2	$N = 500, J = 20$	0.920	0.980	0.982	0.977	46
	$N = 1{,}000, J = 40$	0.980	0.995	0.996	0.994	49
G_3	$N = 500, J = 20$	1.000	1.000	1.000	1.000	50
	$N = 1{,}000, J = 40$	1.000	1.000	1.000	1.000	50
G_4	$N = 500, J = 20$	0.960	0.994	0.990	0.993	46
	$N = 1{,}000, J = 40$	1.000	1.000	1.000	1.000	50

Note: TPR and TFR are the true positive rate and true negative rate based on G-matrix; RTPR and RTFR are the true positive rate and true negative rate based on reachability matrix (R-matrix). The last column reports the number of times the entire G-matrix was recovered correctly.

value that resulted in the best recovery of Q and G in a specific simulation condition. For example, for the attribute hierarchy G_1 and sample size is 500 (test length is 20), the mini-batch size of 256, and the cut off value of 0.2 or 0.3 result in the highest recovery of Q and the mini-batch size of 256 and the cut off value of 0.3 resulted in the best recovery of G_1. From this table, we can see that for the cut-off value, under the condition $N = 500$, the highest recovery accuracy for both Q and G occurs when the cut-off value is 0.2, while for $N = 1{,}000$, both 0.2 and 0.3 yield high accuracy. For the mini-batch sample size, when $N = 500$, the highest recovery accuracy for both Q and G occurs when $N_1 = 256$. Among all eight conditions under $N = 500$, the optimal mini-batch sample size of $N_1 = 128$ appears only once, indicating that $N_1 = 256$ is clearly superior. When $N = 1{,}000$, $N_1 = 512$ outperforms $N_1 = 256$ only two times, indicating that the performance of $N_1 = 256$ and $N_1 = 512$ is similar. We speculate that this is because, with larger sample sizes, even if the mini-batch sample size is not very large, the continual sampling process allows the mini-batch samples to cover all the sample information since the ergodicity, leading to more accurate estimates. Based on these findings, we use mini-batch size $N_1 = 256$ when $N = 500$ and $N_1 = 512$ for $N = 1{,}000$, and cut off value of 0.2 in the subsequent analysis.

5.4.4. Recovery of G

Table 5 displays the recovery of the hierarchical structure G. Based on the results, we find that when $N = 500$, the proportion of accurately estimating G exceeds 80% across all replications and conditions,

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 6. Estimated adjacency matrices $\hat{G}$ of the proposed algorithm for $N = 1{,}000$, $J = 20$ condition

	Balanced				Unbalanced			
G_1	0.000	**0.469**	0.376	0.304	0.000	**0.471**	0.377	0.298
	0.000	0.000	**0.473**	0.343	0.000	0.000	**0.473**	0.372
	0.000	0.000	0.000	**0.436**	0.000	0.000	0.000	**0.476**
	0.031	0.040	0.040	0.000	0.007	0.008	0.007	0.000
G_2	0.000	**0.462**	**0.465**	0.300	0.000	**0.454**	**0.454**	0.299
	0.000	0.000	0.000	**0.431**	0.000	0.000	0.000	**0.472**
	0.000	0.000	0.000	**0.420**	0.000	0.000	0.000	**0.469**
	0.030	0.042	0.040	0.000	0.008	0.008	0.007	0.000
G_3	0.000	**0.466**	**0.432**	**0.446**	0.000	**0.470**	**0.473**	**0.479**
	0.001	0.000	0.001	0.011	0.000	0.000	0.000	0.000
	0.000	0.026	0.000	0.010	0.000	0.000	0.000	0.000
	0.000	0.023	0.000	0.000	0.000	0.000	0.000	0.000
G_4	0.000	**0.465**	**0.470**	0.372	0.000	**0.465**	**0.467**	0.372
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	**0.470**	0.000	0.000	0.000	**0.473**
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: The bolded entries in the matrix represent the real edges.

and this proportion increases to over 88% when $N = 1{,}000$. The TPR and TFR results indicate that, across all simulation conditions, the proposed algorithm achieves over 90% accuracy, except under G_1 in unbalanced condition, in estimating the hierarchical structure edges between attributes, and the RTPR and RTFR results show that the accuracy in estimating direct or indirect dependencies between attributes is even higher, reaching over 95%. Moreover, Table 6 presents the average posterior sample mean of G ($\hat{G}$) across 50 replications under the condition $N = 1{,}000$, $J = 20$, where the bold numbers indicate the true edges. The results show that our algorithm yields estimated values above 0.4 for the true edges between attributes. For non-existing edges but with indirect relationships between attributes, the estimated values range between 0.2 and 0.4. For edges that do not exist and have no indirect relationships between attributes, the estimated values are close to 0. This finding indicates that our method performs remarkably well in identifying the reachability relationships among attributes.

In cases where the hierarchical structure was incorrectly estimated, we found that the primary sources of error were the misestimation of directional relationships among attributes and, in some instances, the failure to identify certain edges. Notably, the recovery of structure G_1 was less ideal compared to the other structures when the sample size was 500. Upon further examination, we observed that most of the estimation failures were caused by incorrect directionality within the attribute hierarchy. For instance, in one replication, the true structure $1 \to 2 \to 3 \to 4$ was incorrectly estimated as $1 \to 2 \to 4 \to 3$, in which the direction between attributes 3 and 4 was reversed. Similar issues were found under structure G_3 in the balanced condition, where the true edges $1 \to 2$, $1 \to 3$, and $1 \to 4$ were sometimes incorrectly estimated as $2/3/4 \to 1$. Moreover, under structure G_3, many incorrect estimations involved missing edges connected to attribute 1. Regardless of which edge was missing, this consistently led to the lowest QRR value for the first attribute.

In addition, the overall recovery accuracy was the lowest under the linear structure G_1. To further investigate this, we conducted an additional analysis of the simulation data and found two main

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

contributing factors: (1) a small proportion of items measuring certain attributes in the Q-matrix and (2) an unbalanced distribution of examinees across the permissible attribute mastery patterns. Specifically, under both balanced and unbalanced conditions, estimation failures often occurred when fewer than 15% of items measured the fourth attribute. Moreover, in the unbalanced condition with $N = 500$, $J = 20$, the examinees with mastery patterns $(1,0,0,0)$ and $(1,1,0,0)$ accounted for less than 10%, sometimes even below 5%, leading to insufficient information for accurate estimation. When the sample size increased to $N = 1{,}000$, this issue was largely mitigated. These findings indicate that both limited item coverage for certain attributes and an unbalanced distribution of attribute mastery patterns can substantially reduce estimation accuracy.

Finally, an interesting finding is that, although the recovery of G shows similar overall accuracy under both balanced and unbalanced attribute profile conditions, there are subtle differences. For example, under a sample size of 500 and test length of 20, the recovery rate under the balanced condition is higher for all structures except G_3. However, when the sample size increases to 1,000 and test length increases to 40, the recovery rate under the unbalanced condition becomes higher. Additionally, under the balanced attribute profile condition, increasing the sample size and test length does not necessarily improve the recovery of G. In contrast, under the unbalanced attribute profile condition, the increase in sample size and test length lead to better recovery of G.

We conjecture that under the balanced condition, the sample of each attribute profile is already relatively well-balanced even at smaller sample sizes, leading to relatively good recovery rates. Therefore, increasing the sample size from 500 to 1,000 does not result in a significant improvement. In the unbalanced condition, however, the imbalance in the sample of each attribute profile results in relatively poor estimation performance at smaller sample sizes. As the sample size increases, the sample of each attribute profile increases, leading to a noticeable improvement in recovery rates.

5.4.5. Recovery of Q matrix

Table 7 presents the recovery accuracy of the Q-matrix. First, across all conditions, the percentage of replications that can fully recover the whole true Q matrix, ranging from 60%(30/50) to 100%(50/50), and with mean of 92%. Second, overall, the proposed algorithm generally resulted in the same or higher recovery of Q, calculated as the number of times the estimated Q-matrix matches the true Q-matrix. This improvement was observed with increasing test length and sample size across almost all conditions, except in the balanced condition with the hierarchical structure G_2. In this case, the number of correct Q estimations decreased from 48 to 45 when moving from $N = 500$, $J = 20$ to $N = 1{,}000$, $J = 40$.

From the element-wise perspective, based on the AQRR value, the element-wise recovery rate was above 97% across all simulation conditions. In the balanced conditions, under structures G_1, G_2, and G_3, the AQRR value showed a slight decrease from $N = 500$, $J = 20$ to $N = 1{,}000$, $J = 40$, while in other conditions, it showed a slight increase. The AQRR values are similar for sample sizes of 500 and 1,000. Similar to the overall recovery of Q, the element-wise recovery rate does not always improve with an increasing sample size. We speculate that this is because, although the sample size increased from 500 to 1,000, the test length also increased from 20 to 40, meaning the number of elements in Q that required estimation doubled thus increase the difficulty in estimation as well.

In addition, the recovery of Q, whether considering element-wise accuracy or overall accuracy, generally shows similar results and trends under both balanced and unbalanced conditions, with only slight differences. When $N = 1{,}000$, the overall number of correct recoveries of Q in the unbalanced condition is slightly higher than in the balanced condition.

Finally, given a sample size, the recovery of Q is similar across the four considered attribute hierarchical structures. However, in terms of the total number of correct recoveries for Q, the unbalanced condition with a sample size of 500 shows only 30 correct recoveries for G_1 (linear structure), which is noticeably lower than in the other conditions. Nevertheless, the corresponding QARR values are similar to those for the other structures. This is mainly closely related to the recovery of G under that condition. Specifically, by comparing Tables 4 and 5, we found that the total number of correct estimations for Q

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 7. Evaluate the recovery accuracy of Q-matrix

		Balanced					
		QRR1	QRR2	QRR3	QRR4	AQRR	$\hat{Q} = Q$
G_1	$N = 500, J = 20$	1.000	1.000	1.000	0.930	0.983	44
	$N = 1,000, J = 40$	1.000	1.000	1.000	0.903	0.976	44
G_2	$N = 500, J = 20$	1.000	1.000	1.000	0.984	0.996	48
	$N = 1,000, J = 40$	1.000	1.000	1.000	0.918	0.980	45
G_3	$N = 500, J = 20$	0.984	0.993	1.000	0.991	0.992	46
	$N = 1,000, J = 40$	0.969	0.979	0.998	0.992	0.985	46
G_4	$N = 500, J = 20$	1.000	1.000	1.000	0.991	0.998	49
	$N = 1,000, J = 40$	1.000	1.000	1.000	1.000	1.000	50
		Unbalanced					
		QRR1	QRR2	QRR3	QRR4	AQRR	$\hat{Q} = Q$
G_1	$N = 500, J = 20$	0.996	0.951	1.000	0.953	0.975	30
	$N = 1,000, J = 40$	1.000	1.000	1.000	0.984	0.996	49
G_2	$N = 500, J = 20$	0.991	0.997	0.996	0.950	0.984	44
	$N = 1,000, J = 40$	1.000	1.000	1.000	0.984	0.996	49
G_3	$N = 500, J = 20$	1.000	1.000	1.000	1.000	1.000	50
	$N = 1,000, J = 40$	1.000	1.000	1.000	1.000	1.000	50
G_4	$N = 500, J = 20$	0.989	0.999	0.982	0.993	0.991	45
	$N = 1,000, J = 40$	1.000	1.000	1.000	1.000	1.000	50

Note: QRR1–QRR4 represent the correct recovery rates for the four attributes, AQRR denotes the element-wise recovery rate of the Q-matrix, and the last column reports the number of times the entire Q-matrix was recovered correctly.

generally aligns closely with that of the hierarchical structure G, with the correct estimations for Q being less than or equal to those for G. This suggests that if G is estimated correctly, there is a high likelihood that Q will also be accurately estimated. Conversely, if G is estimated incorrectly, Q is unlikely to be estimated correctly, since G directly influences the permissible patterns for Q. For unbalanced condition with a sample size of 500 under the hierarchical structure G_1, only 40 out of 50 replications generate correct estimations of the overall G, which is lower than those from other conditions, thus also leads to the undesired recovery of Q.

5.4.6. Recovery of item parameters and attribute profile

Table 8 presents the ARMSE and ABias of the item parameters $\boldsymbol{s}$ and $\boldsymbol{g}$, as well as the class membership parameter $\boldsymbol{\pi}$. Since the true values of $\boldsymbol{s}$ and $\boldsymbol{g}$ are identical across all items, we further report the average HPDI (AHPDI) for $\boldsymbol{s}$ and $\boldsymbol{g}$, defined as the mean of the lower and upper bounds of the HPDIs across all items and over 50 replications. And we also report the AHPDI for $\boldsymbol{\pi}$ over 50 replications in the figures in Appendix C of the Supplementary Material. The results indicate that the parameters $\boldsymbol{s}$, $\boldsymbol{g}$, and $\boldsymbol{\pi}$ exhibit good performance in terms of bias, with bias values close to zero under all simulation conditions. Across all hierarchical structures, as the sample size and test length increase from $N = 500, J = 20$ to $N = 1,000, J = 40$, the average RMSE of $\boldsymbol{s}$ and $\boldsymbol{g}$ shows a declining trend, indicating that the accuracy of item parameter estimation improves with larger sample sizes. Simultaneously, the 95% AHPDI for parameters $\boldsymbol{s}$, $\boldsymbol{g}$, and $\boldsymbol{\pi}$ becomes narrower, further demonstrating that parameter estimation accuracy

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 8. Evaluate the estimation accuracy for the item parameters $\boldsymbol{s}, \boldsymbol{g}$, and the class membership probability parameter $\boldsymbol{\pi}$

		Balanced				
		$ARMSE_s(ABias_s)$	$95\%AHPDI_s$	$ARMSE_g(ABias_g)$	$95\%AHPDI_g$	$ARMSE_\pi(ABias_\pi)$
G_1	$N = 500, J = 20$	0.038(0.009)	[0.116, 0.305]	0.033(0.007)	[0.110, 0.308]	0.023(0.000)
	$N = 1{,}000, J = 40$	0.032(0.008)	[0.143, 0.274]	0.020(0.003)	[0.137, 0.270]	0.026(0.000)
G_2	$N = 500, J = 20$	0.037(0.012)	[0.112, 0.316]	0.031(0.007)	[0.109, 0.309]	0.018(0.000)
	$N = 1{,}000, J = 40$	0.033(0.011)	[0.140, 0.287]	0.021(0.004)	[0.138, 0.274]	0.028(0.000)
G_3	$N = 500, J = 20$	0.036(0.008)	[0.099, 0.323]	0.058(0.024)	[0.116, 0.337]	0.022(0.000)
	$N = 1{,}000, J = 40$	0.029(0.011)	[0.135, 0.288]	0.040(0.013)	[0.137, 0.274]	0.020(0.000)
G_4	$N = 500, J = 20$	0.037(0.010)	[0.109, 0.315]	0.037(0.010)	[0.109, 0.315]	0.014(0.000)
	$N = 1{,}000, J = 40$	0.022(0.006)	[0.133, 0.303]	0.022(0.004)	[0.143, 0.289]	0.006(0.000)
		Unbalanced				
		$ARMSE_s(ABias_s)$	$95\%AHPDI_s$	$ARMSE_g(ABias_g)$	$95\%AHPDI_g$	$ARMSE_\pi(ABias_\pi)$
G_1	$N = 500, J = 20$	0.025(0.003)	[0.130, 0.278]	0.038(0.010)	[0.095, 0.330]	0.013(0.000)
	$N = 1{,}000, J = 40$	0.016(0.002)	[0.151, 0.255]	0.024(0.003)	[0.127, 0.281]	0.007(0.000)
G_2	$N = 500, J = 20$	0.023(0.004)	[0.132, 0.279]	0.065(0.025)	[0.086, 0.375]	0.017(0.000)
	$N = 1{,}000, J = 40$	0.017(0.002)	[0.150, 0.255]	0.022(0.002)	[0.130, 0.277]	0.007(0.000)
G_3	$N = 500, J = 20$	0.030(0.007)	[0.116, 0.302]	0.030(0.005)	[0.112, 0.302]	0.011(0.000)
	$N = 1{,}000, J = 40$	0.019(0.003)	[0.141, 0.266]	0.019(0.001)	[0.139, 0.265]	0.004(0.000)
G_4	$N = 500, J = 20$	0.028(0.006)	[0.122, 0.294]	0.078(0.031)	[0.103, 0.367]	0.015(0.000)
	$N = 1{,}000, J = 40$	0.018(0.003)	[0.146, 0.261]	0.020(0.002)	[0.134, 0.271]	0.003(0.000)

Note: ARMSE, ABias, and AHPDI represent the average RMSE, average bias, and average high density posterior interval across all items or all latent classes.

benefits from larger samples. Furthermore, Table 9 reports the recovery accuracy of the attribute profiles $\boldsymbol{\alpha}$. The values of AAR and PAR also increase as the sample size grows, suggesting that the recovery accuracy of $\boldsymbol{\alpha}$ improves with larger sample sizes.

6. Real data application

In this section, we demonstrated the proposed algorithm through a data set collected from the Examination for the Certificate of Proficiency in English (ECPE). ECPE is a test developed and scored by the English Language Institute of the University of Michigan. This data set is based on 2,922 examinees's responses to 28 multiple-choice questions. Three skills are designed to be measured in ECPE: (1) knowledge of morphosyntactic rules, (2) cohesive rules, and (3) lexical rules (Buck & Tatsuoka, 1998; Henson & Templin, 2007).

6.1. Analysis procedures

Given the total sample size, we chose three mini-batch sample sizes N_1 = 512, 1,024, and 1,500 to implement the proposed algorithm. A total of four Markov chains were conducted for each selected batch size, and the chain length was set as 20,000 and burn in value is 10,000. In addition, the prior for each parameter is consistent with the settings used in the simulation study. Our purpose is to investigate whether the proposed algorithm can successfully recover the linear hierarchical structure $3 \rightarrow 2 \rightarrow 1$ among the three skills in the ECPE test, as identified in previous research (Templin & Bradshaw, 2014; Wang & Lu, 2021).

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 9. Evaluate the estimation accuracy for the attribute profile $\boldsymbol{\alpha}$

		Balanced					Unbalanced				
		AAR1	AAR2	AAR3	AAR4	PAR	AAR1	AAR2	AAR3	AAR4	PAR
G_1	$N = 500, J = 20$	0.968	0.959	0.964	0.929	0.839	0.976	0.981	0.978	0.949	0.897
	$N = 1{,}000, J = 40$	0.990	0.989	0.989	0.927	0.899	0.995	0.996	0.995	0.983	0.970
G_2	$N = 500, J = 20$	0.961	0.948	0.947	0.962	0.837	0.978	0.965	0.968	0.932	0.863
	$N = 1{,}000, J = 40$	0.985	0.984	0.983	0.925	0.8820	.994	0.993	0.994	0.983	0.966
G_3	$N = 500, J = 20$	0.933	0.930	0.920	0.917	0.749	0.962	0.939	0.946	0.935	0.824
	$N = 1{,}000, J = 40$	0.952	0.955	0.971	0.966	0.858	0.988	0.981	0.984	0.978	0.938
G_4	$N = 500, J = 20$	0.963	0.934	0.937	0.946	0.811	0.967	0.939	0.950	0.937	0.826
	$N = 1{,}000, J = 40$	0.988	0.978	0.981	0.983	0.934	0.993	0.988	0.990	0.988	0.963

Note: AAR1 represents the correct classification rate for the first attribute, AAR2 for the second attribute, AAR3 for the third attribute, and AAR4 for the fourth attribute. PAR stands for the pattern-wise agreement rate.

Table 10. Estimated adjacency matrix $\hat{\mathrm{G}}$ and the DIC value under different mini-batch conditions

	$N_1 = 512$			$N_1 = 1{,}024$			$N_1 = 1{,}500$		
$\hat{\mathrm{G}}$	0.000	0.053	0.052	0.000	0.029	0.006	0.000	0.065	0.000
	0.408	0.000	0.170	**0.212**	0.000	0.017	**0.426**	0.000	0.019
	0.374	**0.364**	0.000	**0.340**	**0.362**	0.000	**0.407**	**0.489**	0.000
DIC	83,271.820			82,280.660			82,837.410		

Note: The bold values represent the edges that are considered to exist with strong statistical evidence.

6.2. Results

First of all, the $\hat{R}$ values of $\boldsymbol{s}$ and $\boldsymbol{g}$ under $N_1 = 512$ condition are 1.000 and 1.000, respectively, under $N_1 = 1{,}024$ condition are 1.020 and 1.040, respectively, and under $N_1 = 1{,}500$ condition are 1.010 and 1.020, respectively. All $\hat{R}$ values are all below 1.1, indicating that the Markov chains have converged. Table 10 shows the $\hat{G}$, and the DIC values under different mini-batch conditions.

6.2.1. The estimated G

In terms of $\hat{G}$ values, the estimated values for the true edges in real data are somehow lower than those in the simulations. For example, the estimated value for edge $2 \to 1$ is only 0.212 (i.e., $\hat{G}(2,1) = 0.212$) under the $N_1 = 1{,}024$ condition. Yet results also show that the estimates for non-existent edges are close to zero, that is, all edges other than $2 \to 1$, $3 \to 1$, and $3 \to 2$ have estimates close to zero, indicating their absence. In addition, the results under both $N_1 = 512$ and $N_1 = 1{,}500$ consistently reveal the same hierarchical structure. Accordingly, our estimated hierarchical structure is $3 \to 2 \to 1$ which is consistent with previous studies mentioned above. Moreover, based on the DIC values, we selected the mini-batch size of 1,024, which provided the minimum DIC value, for further analysis.

6.2.2. The estimated Q

Finally, based on the result under $N_1 = 1{,}024$ condition, Table 11 presents both the expert labeled structured Q-matrix and our estimated structured Q-matrix, with highlighted elements indicating discrepancies between them. The differences occurred in items 9, 17, and 21, and mostly on the first two attributes. This is because attribute 3 is a prerequisite for attributes 1 and 2, and thus all its values are 1. The element-wise agreement of expert labeled Q and our estimated Q is 0.952, indicating that the element-wise consistency rate with the expert-labeled Q-matrix reaches 95.2%, further demonstrating the effectiveness of our algorithm.

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Table 11. The expert labeled Q-matrix and the estimated structured Q-matrix of from ECPE data

Item	Expert labeled structured **Q**			Estimated structured **Q**		
1	1	1	1	1	1	1
2	0	1	1	0	1	1
3	1	1	1	1	1	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	1	0	0	1
7	1	1	1	1	1	1
8	0	1	1	0	1	1
9	0	0	1	1	1	1
10	1	1	1	1	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	0	0	1	0	0	1
16	1	1	1	1	1	1
17	0	1	1	1	1	1
18	0	0	1	0	0	1
19	0	0	1	0	0	1
20	1	1	1	1	1	1
21	1	1	1	0	1	1
22	0	0	1	0	0	1
23	0	1	1	0	1	1
24	0	1	1	0	1	1
25	1	1	1	1	1	1
26	0	0	1	0	0	1
27	1	1	1	1	1	1
28	0	0	1	0	0	1

Note: The highlighted elements represent the differences between the true structured Q-matrix and the estimated structured Q-matrix.

7. Discussion

In this study, we proposed a Bayesian estimation method for the simultaneous estimation of the Q-matrix, the attribute hierarchy, and the parameters of the DINA model when both the Q-matrix and the attribute hierarchy are unknown. To implement this, we developed an MH-Gibbs sampling algorithm. To enhance computational efficiency, we incorporated a mini-batch approach and investigated the impact of mini-batch sample size on estimation accuracy. A series of simulation studies were conducted to evaluate the performance of the proposed algorithm across various conditions, including mini-batch sample size, cutoff values for G, different attribute hierarchy structures, sample size, and

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

test length. Overall, the simulation results support the accuracy and computational efficiency of our algorithm. Given a specified cut-off value and mini-batch size, the proposed algorithm demonstrates strong performance across most conditions. Except for the G_1 structure with $N = 500$ under the unbalanced condition, the overall recovery rates for both the Q-matrix and the adjacency matrix G exceed 88%, with several conditions achieving a perfect 100% recovery. At the element-wise level, the algorithm achieves recovery rates above 97.5% for the Q-matrix and over 95% for the adjacency matrix G. These results underscore the robustness and reliability of the proposed estimation method in jointly recovering both the Q-matrix and the attribute hierarchy structure, even when both are unknown. A real data application further demonstrates the validity of the estimated latent structures, G and Q, obtained from the proposed method.

We next provide further discussion on the major findings from our simulation studies and real data application, which may offer practical guidelines for applying the proposed algorithm to datasets with varying structures. First, regarding the selection of mini-batch size, our results suggest that using a mini-batch size between one-quarter and one-half of the total sample size generally yields accurate parameter estimates. However, we recommend that users test multiple mini-batch sizes when analyzing real data and select the optimal setting based on model fit indices such as the DIC. This recommendation stems from the fact that our simulations and empirical analyses were conducted on datasets with sample sizes ranging from approximately 500 to 3,000, and performance may vary in other contexts.

Next, regarding the selection of the cut-off value for G, our simulation and real data analyses suggest that a threshold between 0.2 and 0.3 generally yields accurate estimation of the attribute hierarchy under the DINA model. We observed that with larger sample sizes, the estimated values for existing edges tend to be higher, indicating increased precision as more information becomes available. Importantly, even when the estimated values for connected (i.e., direct and indirect) edges are relatively small (around 0.2), those for non-connected edges are typically close to zero. This distinction allows for reliable differentiation between edges with and without dependencies. While we used a cut-off of 0.2 in this study, we caution that this is not a universal threshold; users are encouraged to determine an appropriate cut-off based on the characteristics of their own data.

Finally, an interesting finding is that although G and the corresponding Q in some simulation conditions exhibit the least desirable recovery when the sample size is 500, the item parameters and attribute profiles under this condition still perform similarly to those of the other attribute hierarchical structure. This is consistent with the finding from Wang (2018) that the classification results from CDM can maintain at relatively high accuracy for a missspecified Q-matrix that satisfy certain conditions.

We conclude by discussing several limitations of the current study and potential directions for future research. First, to determine the final structure of the Q-matrix, we used a fixed threshold of 0.5. However, in the real-world ECPE dataset, we observed that some estimated values were very close to this cutoff (e.g., between 0.49 and 0.51), raising concerns about the appropriateness of using a strict threshold. Identifying a more flexible or data-driven approach for determining this threshold could improve Q-matrix recovery accuracy and represents an important area for future investigation.

Second, in this article, we set cut-off value as 0.2 on the ECPE data set since the posterior estimates of the adjacency matrix G clearly distinguish between existing and non-existing edges. However, this may not hold for other data sets. A straightforward strategy we currently consider is to try several different cut-off values, fit the model under each, and then use model selection criteria such as DIC to determine the most appropriate attribute hierarchy structure. However, constructing a theoretically grounded, data-driven procedure for determining the cut-off value would be an important direction for our future research, which could further improve and refine our proposed method.

Third, although many CDMs have been developed—such as the DINA model (Junker & Sijtsma, 2001), the deterministic inputs, noisy "or" gate (DINO) model (Templin & Henson, 2006), the generalized DINA (GDINA) model (De La Torre, 2011), the log-linear cognitive diagnosis model (LCDM) (Henson et al., 2009), and the general diagnostic model (GDM) (Von Davier, 2008)—this study focuses on developing an estimation method based on the DINA model. The DINA model is one of

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

the most widely used CDMs and has gained substantial popularity in educational and psychological measurement due to its interpretability and simplicity. We plan to extend our algorithm to accommodate more general CDMs, such as the LCDM, GDM, and GDINA models. However, such an extension may involve the following two main challenges. First, when an attribute hierarchy is present, under the DINA model's response mechanism, it is sufficient to impose constraints on either the attribute profiles or the Q-matrix. However, this property may not hold for other CDMs. Second, in the DINA model, the item parameters have explicit posterior forms, which allows the algorithm to be implemented within a Gibbs sampling framework. In contrast, this property may not be satisfied in other CDMs, posing another challenge for extending our method.

Additionally, the current study focused only on a fixed number of attributes K. Therefore, extending our method to scenarios where K is unknown would be an important direction for future research. Finally, although our use of the mini-batch approach significantly improves computational efficiency, the average runtime remains between 14 and 18 minutes for a sample size of $N = 1{,}000$. Thus, we aim to further optimize our implementation to improve scalability and computational efficiency, for example, by exploring potential enhancements in the Q-matrix updating procedure (Gu & Dunson, 2023).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10093.

Data availability statement. Publicly available dataset was analyzed in this study. The dataset can be found as follows: https://cran.r-project.org/web/packages/CDM/index.html.

Funding statement. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The authors declare that they have no competing interests.

Ethical standards. The authors declared no potential conflict of interests with respect to the research, authorship, and/or publication of this article. This research does not involve human participants and/or animals.

Disclosure of use of AI statement. The authors utilized ChatGPT5.2 to help with manuscript writing process. Generative artificial intelligence was used to rephrase authors' original written context, which is mainly for language polishing purposes. The authors reviewed the created content produced by this generative AI Tool and assert the content within this document is factually accurate and free of plagiarism. The authors take full responsibility for the submitted document.

References

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157.

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

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

Chen, Y., & Wang, S. (2023). Bayesian estimation of attribute hierarchy for cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *48*(6), 810–841.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

Chung, M. (2019). A Gibbs sampling algorithm that estimates the Q-matrix for the DINA model. *Journal of Mathematical Psychology*, *93*, 102275.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476.

Culpepper, S. A. (2019). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, *84*(2), 333–357.

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273.

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

De La Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595–624.

De Sa, C., Chen, V., & Wong, W. (2018). Minibatch Gibbs sampling on large graphical models. In *International conference on machine learning*. In J. Dy & A. Krause (Eds.), Proceedings of the 35th International Conferenceon Machine Learning, volume 80 of Proceedings of Machine Learning Research (pp. 1165–1173).: PMLR.

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

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*, 1–18.

Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, *71*, 1054–1063.

Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in Bayesian factor analysis. *Psychometrika*, *82*(2), 295–307.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gu, Y., & 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.

Gu, Y., & Xu, G. (2023). Identifiability of hierarchical latent attribute models. *Statistica Sinica*, *33*(4), 2561–2591.

Henson, R., & Templin, J. (2007). Large-scale language assessment using cognitive diagnosis models. Paper presented at the annual meeting of the National Council for Measurement in Education in Chicago, Illinois.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, *14*(1), 1303–1347.

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

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.

Lee, S., & Gu, Y. (2024). Latent conjunctive Bayesian network: Unify attribute hierarchy and Bayesian network for cognitive diagnosis. *The Annals of Applied Statistics*, *18*(3), 1988–2011.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, *19*(5A), 1790.

Liu, Y., Xin, T., & Jiang, Y. (2022). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*, *57*(5), 784–803.

Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, *88*(1), 175–207.

Ma, C., & Xu, G. (2022). Hypothesis testing for hierarchical structures in cognitive diagnosis models. *Journal of Data Science*, *20*(3), 279–302.

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Tatsuoka, K. K. (1984). Analysis of errors in fraction addition and subtraction problems. Computer-based Education Research Laboratory, University of Illinois.

Tatsuoka, K. K. (2012). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman, & R.L. Brennan, (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Routledge.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339.

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

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, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics*, *46*(1), 58–84.

Wang, D., Cai, Y., & Tu, D. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research*, 1–13.

https://doi.org/10.1017/psy.2026.10093 Published online by Cambridge University Press

Wang, S. (2018). Two-stage maximum likelihood estimation in the misspecified restricted latent class model. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 300–333.

Wu, T.-Y., Rachel Wang, Y., & Wong, W. H. (2022). Mini-batch Metropolis–Hastings with reversible SGLD proposal. *Journal of the American Statistical Association*, *117*(537), 386–394.

Xiang, R. (2013). *Nonlinear penalized estimation of true Q-matrix in cognitive diagnostic models* [Unpublished doctoral dissertation]. Columbia University.

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

Zhang, X., Jiang, Y., Xin, T., & Liu, Y. (2024). Iterative attribute hierarchy exploration methods for cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *50*(4), 682–713.

Zou, D., & Gu, Q. (2021). On the convergence of Hamiltonian Monte Carlo with stochastic gradients. In M. Meila & T. Zhang (Eds.), Proceedings of the 38th InternationalConference on Machine Learning, volume 139 of Proceedings of Machine LearningResearch (pp. 13012–13022).: PMLR.

Cite this article: Wang, X., Chen, Y. and Wang, S. (2026). Uncovering Latent Structures: A Bayesian Approach to Estimating Q-Matrix and Attribute Hierarchies in Cognitive Diagnostic Models. *Psychometrika*, 669–694. https://doi.org/10.1017/psy.2026.10093