Psychometrika (2025), **0**, 1–38
doi:10.1017/psy.2024.20

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Generalized Bayesian method for diagnostic classification models

Kazuhiro Yamaguchi[1], Yanlong Liu[2] and Gongjun Xu[2]

[1]University of Tsukuba, Tsukuba, Japan; [2]University of Michigan, Ann Arbor, MI, USA

Corresponding author: Kazuhiro Yamaguchi; Email: yamaguchi.kazuhir.ft@u.tsukuba.ac.jp

(Received 31 October 2024; revised 10 November 2024; accepted 11 November 2024)

Abstract

This study extends the loss function-based parameter estimation method for diagnostic classification models proposed by Ma, de la Torre, et al. (2023, Psychometrika) to consider prior knowledge and uncertainty of sampling. To this end, we integrate the loss function-based estimation method with the generalized Bayesian method. We establish the consistency of attribute mastery patterns of the proposed generalized Bayesian method. The proposed generalized Bayesian method is compared in a simulation study and found to be superior to the previous nonparametric diagnostic classification method—a special case of the loss function-based method. Moreover, the proposed method is applied to real data and compared with previous parametric and nonparametric estimation methods. Finally, practical guidelines for the proposed method and future research directions are discussed.

Keywords: diagnostic classification models; generalized Bayesian method; loss function-based method; parameter estimation

1. Introduction

Learning is an important aspect of human life. The current status of individual knowledge or depth of understanding must be evaluated to ensure efficient learning. Test analysis models called diagnostic classification models (DCMs; Rupp et al., 2010; von Davier & Lee, 2019) have been popularly employed to capture an individual's learning status. Notably, DCMs provide useful statistical tools to reveal individuals' current learning status based on the test's item responses. Latent knowledge or cognitive elements are called attributes and are expressed as latent categorical variables in DCMs. Moreover, DCMs are known as restricted latent class models (e.g., Rupp & Templin, 2008; Xu, 2017), wherein each possible set of attributes represents a latent class. In other words, attribute mastery patterns indicate the attributes that are either mastered or not mastered. Therefore, one of the DCMs' final outputs is the estimate of the attribute mastery patterns of individuals or attribute mastery probabilities.

Various parameter estimation methods for the DCMs have been actively developed. Parametric and nonparametric estimation methods are commonly used in DCMs. Parametric estimation methods assume parametric item response functions and structural models. Therefore, parametric estimation methods employ a likelihood function under the assumed model and include (penalized or regularized) maximum likelihood estimation (e.g., Chen et al., 2015; de la Torre, 2009; Ma, Ouyang, & Xu, 2023) and Bayesian estimation methods (e.g., Culpepper, 2015; Yamaguchi & Okada, 2020; Yamaguchi & Templin,

2022b), incorporating prior distributions for model parameters. Numerous parametric estimation methods have been developed and their properties have been studied (e.g., von Davier & Lee, 2019).

On the other hand, nonparametric methods (e.g., Chiu et al., 2018; Chiu & Douglas, 2013) do not use probabilistic item response models; instead, they use an ideal response to define a type of discrepancy function, which will be formally defined in a later section. Such discrepancy functions are defined based on the distance between each item's ideal and actual responses. Intuitively, nonparametric methods directly estimate attribute mastery patterns, which minimize the discrepancy function. Therefore, nonparametric methods do not require a probabilistic item response function. Nonparametric methods exhibit satisfactory statistical properties, such as consistency under certain conditions (Chiu & Köhn, 2019; Wang & Douglas, 2015).

Recently, a general parameter estimation method that can uniformly express parametric and nonparametric methods was developed by Ma, de la Torre, and Xu (2023). The unified estimation method developed by Ma, de la Torre, and Xu (2023) is a loss function-based estimation method for DCMs. If we select cross-entropy for a loss function, its minimization corresponds to maximizing the joint maximum likelihood (Chiu et al., 2016). The distance or discrepancy function in nonparametric methods is a well-known loss function. Additionally, by adding penalty terms to a cross-entropy loss function, we obtain the maximum a posteriori (MAP) estimates, classical Bayesian estimates, to minimize it. These examples indicate that the loss function-based estimation method is flexible and can represent various estimation methods in a unified manner. Furthermore, a unified estimation algorithm for the loss function-based estimation method was available.

However, loss function-based methods exhibit certain limitations. First, these methods only provide point estimates, which may be problematic because we cannot evaluate how point estimates vary due to sampling or estimation variations. Therefore, we cannot evaluate the uncertainty of attribute mastery using the loss function-based method. Furthermore, attribute mastery probabilities for each individual are not expressed in the loss function-based method. This is the same problem that occurs in DCMs' nonparametric estimation method. However, attribute mastery probabilities represent a more nuanced situation than attribute mastery pattern results, with or without mastery. Another limitation of these methods is that prior information on weight parameters in the generalized nonparametric method that defines generalized ideal responses is generally not considered. However, DCM users may have prior knowledge of the test items' conjunctive and disjunctive nature. If so, domain-specific knowledge must be included to improve parameter estimates.

It is not only loss function-based methods that have limitations that need to be addressed; several limitations of previous parametric and nonparametric estimation methods likewise need to be noted. First, parametric estimation methods need to specify data-generating distributions, which determine the likelihood function. The likelihood function provides a connection between data and model parameters such as attribute mastery patterns. Moreover, likelihood functions make it possible to evaluate estimation uncertainty with the asymptotical theory within the maximum likelihood framework or the posterior distribution within the Bayesian framework. However, the data-generating process is not always specified. DCMs are part of the educational measurement model family that need various constraints and limitations, making it difficult to specify the model.

Some of the limitations of the nonparametric methods are the same as those of the loss function-based methods. For instance, current nonparametric methods for DCMs cannot evaluate the uncertainty of attribute mastery estimates. The nonparametric methods for DCMs were developed in studies with small sample sizes (Chiu & Douglas, 2013). Ultimately, the nonparametric method for DCMs can be applied to individuals; however, the parameter estimates need to be evaluated with variability of parameter estimates. Currently, the nonparametric methods simply select attribute mastery patterns to minimize prespecified distance functions so the parameter uncertainty evaluation is not included in the framework. The parameter estimates with nonparametric methods can be changed by small differences in the loss function. One main purpose of DCMs is the diagnosis of individual knowledge. Thus, such variations in parameter estimates due to small differences in the distance functions may be a fundamental problem for application.

To overcome these limitations and extend the previous loss function-based estimation method for DCMs, we employ the generalized Bayesian (GB; Bissiri et al., 2016) method. The usual Bayesian parameter update determines the likelihood function and updates the model parameters in the likelihood with the observed dataset. By contrast, the GB method can express parameter updating with a dataset via loss functions. Therefore, Bayesian inference is applicable to nonparametric-based estimation methods as well as to likelihood-based methods. Moreover, other benefits of the GB method are as follows. First, the GB method originally assumes the $\mathcal{M}$-open setting (Bernardo & Smith, 2009, Chap. 6), which implies that the GB method provides a valid inference even if the assumed model does not match the true data-generating mechanism. Various DCMs have been developed; however, selecting an appropriate item response function that expresses the true data-generating mechanism is not always possible. The GB method does not require an entire data-generating model but instead sets a loss function related to the parameter sets of interest. This means that we do not need to find a correct data-generating model, which is always unknown and often misspecified. We expect the GB method to overcome the practical difficulties of DCMs' applications.

Second, the GB method allows the use of flexible loss functions and priors. The uncertainty of the parameters expressed in the loss functions is easily demonstrated in the generalized posteriors generated using the GB method. In other words, the GB method can handle the amount of uncertainty of attribute mastery estimates. Not only point estimates but also uncertainty variation are important for careful decisions in diagnostic evaluation. The GB method provides a useful tool for addressing the above problems, which both parametric and nonparametric methods have. Furthermore, the generalized posterior is easily obtained using a Markov chain Monte Carlo (MCMC) routine, such as the Metropolis–Hastings method. Third, we can control the relative importance between the dataset and the prior via the learning rate parameter. If the obtained data's quality is questionable, an inference that is completely dependent on the data may lead to inappropriate decisions. In such cases, the data's relative importance can be reduced. The learning rate parameter enables a more flexible inference.

Based on these discussions, we develop a GB method to overcome the limitations of the loss function-based estimation method for DCMs (Ma, de la Torre, & Xu, 2023). The remainder of this article is organized as follows: The second section demonstrates the basic setup of the DCMs and the previous loss function-based estimation method. The third section provides the GB method's fundamentals and its application to DCMs based on their loss functions. Therein, the MCMC algorithm for a generalized posterior is also discussed. The GB method's mathematical properties under certain conditions are discussed in the fourth section. The fifth and sixth sections comprise simulation and real data analysis examples of GB inference for DCMs, wherein we compare previous nonparametric estimation methods in a simulation study. Finally, the seventh section serves as the discussion, where the limitations of the GB inference and future directions of DCMs' estimation methods are discussed.

2. Model setup and previous estimation methods

2.1. Model setup of DCMs

First, we express an individual's attribute mastery pattern using a vector of length $K, \boldsymbol{\alpha}_i \in \{0,1\}^K$, where $i \in \{1,2,\ldots,I\}$. The kth element of the attribute mastery pattern vector $\boldsymbol{\alpha}_i$ is $\alpha_{ik} \in \{0,1\}$, where $k \in \{1,2,\ldots,K\}$; it takes one if individual i masters attribute k, and otherwise, it takes 0. In this study, we assume unconditional attribute mastery patterns, where all possible attribute mastery patterns and the number is $L = 2^K$. Therefore, the $l \in \{1,2,\ldots,L\}$-th attribute mastery pattern can be written as $\boldsymbol{\alpha}_l$. The set of attribute mastery patterns for all individuals is $\mathcal{A} = \{\boldsymbol{\alpha}_i\}_{i=1}^I$. To define the parametric measurement model, we also need to specify the diagnostic relationship between the attributes and item sets.

The diagnostic relationship between the attributes and test items is called the $\boldsymbol{q}$-vector, $\boldsymbol{q}_j \in \{0,1\}^K \setminus \{\mathbf{0}_K\}$, where $j \in \{1,2,\ldots,J\}$; if the kth attribute is required for item $j, q_{jk} = 1$; otherwise $q_{jk} = 0$. Additionally, $\mathbf{0}_K$ is a vector of length K and all its elements are 0. Here, we assume there is no item with $\boldsymbol{q}_j = \mathbf{0}_K$. The Q-matrix (Tatsuoka, 1985) is a $J \times K$ matrix defined by $\left(\boldsymbol{q}_1^\top, \boldsymbol{q}_2^\top, \ldots, \boldsymbol{q}_J^\top\right)^\top$.

Parametric DCMs define their measurement models using attribute mastery patterns and a $\boldsymbol{q}$-vector. For example, one of the most general DCMs, known as the log linear cognitive diagnostic model (LCDM; Henson et al., 2009), uses an item parameter vector $\boldsymbol{\lambda}_j = (\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{j12\ldots K})^\top$, and the measurement model of $X_{ij} = 1$, which is a conditional response probability of individual i for item j, is:

$$P\left(X_{ij} = 1 \mid \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right) = \frac{\exp\left(f\left(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i\right)\right)}{1 + \exp\left(f\left(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i\right)\right)}, \tag{1}$$

where $f\left(\boldsymbol{\lambda}_j, \boldsymbol{\alpha}_i\right)$ is:

$$\begin{aligned} f\left(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i\right) &= \log \frac{P\left(X_{ij} = 1 \mid \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right)}{1 - P\left(X_{ij} = 1 \mid \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right)} \\ &= \lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk} q_{jk} \alpha_{ik} + \sum_{k=1}^{K} \sum_{k'<k} \lambda_{jkk'} q_{jk} q_{jk'} \alpha_{ik} \alpha_{ik'} + \cdots + \lambda_{j12\ldots K} \prod_{k=1}^{K} q_{jk} \alpha_{ik}. \end{aligned} \tag{2}$$

The LCDM has several parameters. The first parameter is the intercept λ_{j0}, which determines the baseline correct item response probability. Attribute mastery patterns that do not master any attributes requiring item j take response probability. The main effect parameters were $\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jK}$; each parameter affected the correct item response probabilities with the corresponding attributes. The first-order interaction parameter $\lambda_{jkk'}$ is the effect of simultaneously mastering the attributes k and k'. Similarly, we introduce the highest interaction term as $\lambda_{j12\ldots K}$. General cognitive diagnosis models that are similar to LCDM have also been proposed in the literature, including the generalized DINA (GDINA) model (de la Torre, 2011) and general diagnostic model (GDM; von Davier, 2008).

As some attributes are not measured by item j, the number of estimated item parameters under LCDM is $2^{\sum_k q_{jk}} \leq 2^K$. Moreover, notably, one-to-one mapping exists between the LCDM item parameters and conditional item response probabilities (Rupp et al., 2010). Therefore, it is convenient to use conditional item response probabilities to develop parameter estimation methods. The same strategy was adopted in previous studies (Yamaguchi & Okada, 2020; Yamaguchi & Templin, 2022b), where DCMs are a restricted version of latent class models (e.g., Rupp & Templin, 2008; Xu & Shang, 2018).

Therefore, let the correct item response probability be parameter θ_{jl}:

$$\theta_{j,\boldsymbol{\alpha}_l} = P\left(X_{ij} = 1 \mid \boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right), \tag{3}$$

Additionally, the attribute mastery mixing parameters $\pi_{\boldsymbol{\alpha}_1}, \pi_{\boldsymbol{\alpha}_2}, \ldots, \pi_{\boldsymbol{\alpha}_L} \in (0,1)$ are defined as $\pi_{\boldsymbol{\alpha}_l} = P\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right)$, satisfying $\sum_l \pi_{\boldsymbol{\alpha}_l} = 1$. From this notation, the complete data likelihood function of the LCDM is:

$$\mathcal{L}\left(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X\right) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{l=1}^{L} \left\{\pi_l \theta_{j,\boldsymbol{\alpha}_l}^{x_{ij}} \left(1 - \theta_{j,\boldsymbol{\alpha}_l}\right)^{1 - x_{ij}}\right\}^{\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)}, \tag{4}$$

where $X = \left\{x_{ij}\right\}_{i,j=1}^{N,J}, \Theta = \left\{\theta_{jl}\right\}_{j,l=1}^{J,L}, \boldsymbol{\pi} = \left(\pi_{\boldsymbol{\alpha}_1}, \pi_{\boldsymbol{\alpha}_2}, \ldots, \pi_{\boldsymbol{\alpha}_L}\right)^\top$, and $\mathcal{I}\left(\cdot\right)$ is an indicator function.

We add some remarks on the correct item response probabilities for item j. First, as mentioned previously, some attribute mastery patterns have the same item response probabilities because of the setting of the $\boldsymbol{q}$ vector. Moreover, some submodels of the LCDM assume fewer parameters than the general LCDM and have parsimonious model forms. The model settings for each item can differ, but we assume that all test items have the same general LCDM form.

Second, the correct item response probabilities for item j exhibit an ordinal relationship: These relationships are known as monotonicity constraints (Xu & Shang, 2018). The formal expression of the monotonicity constraints proposed by Xu and Shang (2018) is:

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq \boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq q_j} \theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\boldsymbol{\alpha}'} \geq \theta_{j,\mathbf{0}_K}, \tag{5}$$

where we write $\boldsymbol{\alpha} \succcurlyeq \boldsymbol{q}_j$ if $\alpha_k \geq q_{jk}, \forall k$; otherwise, $\boldsymbol{\alpha} \not\succcurlyeq \boldsymbol{q}_j$. These constraints imply that the patterns mastering all skills measured in item $j\left(\boldsymbol{\alpha} : \boldsymbol{\alpha} \succcurlyeq \boldsymbol{q}_j\right)$ should have the highest of all the patterns. By contrast, all nonmastering patterns had the lowest correct item response probability. The middle mastering patterns satisfying $\boldsymbol{\alpha}' \not\succcurlyeq \boldsymbol{q}_j$ have response probabilities between these two probabilities.

2.2. Loss function-based parameter estimation

This section introduces the loss function-based parameter estimation method proposed in Ma, de la Torre, and Xu (2023). First, we describe certain elements of the loss function-based method. In this framework, we introduce the length J centroid parameter vector: $\boldsymbol{\mu}_{\boldsymbol{\alpha}_l} = \left(\mu_{1,\boldsymbol{\alpha}_l}, \mu_{2,\boldsymbol{\alpha}_l}, \ldots, \mu_{J,\boldsymbol{\alpha}_l}\right)^\top \in \mathbb{R}^J$. Additionally, a penalty term for the mixing parameter $\pi_{\boldsymbol{\alpha}_l}$ is introduced as $h\left(\pi_{\boldsymbol{\alpha}_l}\right) \in \mathbb{R}$. Furthermore, an element-wise loss function taking item response vector $\boldsymbol{x}_i$ and a centroid parameter vector $\boldsymbol{\mu}_{\boldsymbol{\alpha}_l}$ is expressed as $\ell\left(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}_l}\right)$; its codomain is real positive number $\mathbb{R}^+$. The $\ell\left(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}_l}\right)$ is the individual-level loss function. Therefore, the loss function of the entire dataset is based on the individual-level loss function:

$$\mathcal{L}\left(\mathcal{A}, \boldsymbol{\mu}, \boldsymbol{\pi}\right) = \sum_{l=1}^{L} \sum_{i:\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l} \left\{\ell\left(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}_l}\right) + h\left(\pi_{\boldsymbol{\alpha}_l}\right)\right\}. \tag{6}$$

The second summation takes over the individuals with the attribute mastery pattern $\boldsymbol{\alpha}_l$.

Parameter estimates are obtained to minimize the loss function defined above:

$$\left\{\widehat{\mathcal{A}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\pi}}\right\} = \underset{\mathcal{A}, \boldsymbol{\mu}, \boldsymbol{\pi}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{A}, \boldsymbol{\mu}, \boldsymbol{\pi}\right). \tag{7}$$

Directly minimizing the above loss function is not easy; therefore, we use the iterative update rule instead. In the estimation algorithm, we first set initial parameters $\left\{\boldsymbol{\mu}^{(0)}, \boldsymbol{\pi}^{(0)}\right\}$. When we have parameter estimates at tth iteration, $\left\{\boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)}\right\}$, the following update steps are repeated:

$$\begin{aligned} \text{Step 1}: \left\{\mathcal{A}^{(t+1)}\right\} &= \underset{\mathcal{A}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{A}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)}\right), \\ \text{Step 2}: \left\{\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}\right\} &= \underset{\boldsymbol{\mu}, \boldsymbol{\pi}}{\operatorname{argmin}}\, \mathcal{L}\left(\mathcal{A}^{(t+1)}, \boldsymbol{\mu}, \boldsymbol{\pi}\right). \end{aligned} \tag{8}$$

If the predetermined convergence criterion is satisfied, for example, $\varepsilon > 1 - \sum_i \left(\mathcal{I}\left(\boldsymbol{\alpha}_i^{(t+1)} = \boldsymbol{\alpha}_i^{(t)}\right)\right) / I, 0 < \varepsilon < 1$, the update process is stopped, and the parameter estimates become output: $\left\{\widehat{\mathcal{A}}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\pi}}\right\} = \left\{\mathcal{A}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}\right\}$.

Many previous estimation methods can be viewed as special cases of the general loss function formulation framework. The joint likelihood estimation of the parametric DCM is a first example. In the following, we focus on the deterministic inputs noisy, "and" gate model (DINA model; Junker & Sijtsma, 2001; MacReady & Dayton, 1977; Maris, 1999) as an example because it is well-known and considered the most parsimonious DCM. The loss function further used to obtain the MAP estimation, which is the negative of the sum of the log-likelihood and log-prior density functions, is also presented. Subsequently, a nonparametric classification method (NPC; Chiu & Douglas, 2013; Wang & Douglas, 2015) and generalized NPC (GNPC; Chiu & Köhn, 2019; Chiu et al., 2018) are formulated under the above framework. Furthermore, NPC and GNPC are extended to the GB framework in a later section.

The DINA model is the simplest and most fundamental DCM, which is a special case of the LCDM. The DINA model assumes only the intercept and the highest interaction terms of the LCDM item parameters. Let the subscript set of attributes measured by item j be $\mathcal{K} = \left\{k; q_{jk} = 1, k = 1, 2, \ldots, K\right\}$ and let the LCDM kernel for the DINA model be reduced to:

$$f\left(\boldsymbol{\lambda}_j, \boldsymbol{q}_j, \boldsymbol{\alpha}_i\right) = \lambda_{j0} + \lambda_{j\mathcal{K}} \prod_{k \in \mathcal{K}} \alpha_{ik}. \tag{9}$$

In the conventional DINA formulation, two-item response probabilities are represented by estimating the g_j and slipping s_j parameters:

$$g_j = \frac{\exp(\lambda_{j0})}{1+\exp(\lambda_{j0})}, \tag{10}$$

$$1-s_j = \frac{\exp\left(\lambda_{j0}+\lambda_{j\mathcal{K}}\prod_{k\in\mathcal{K}}\alpha_{ik}\right)}{1+\exp\left(\lambda_{j0}+\lambda_{j\mathcal{K}}\prod_{k\in\mathcal{K}}\alpha_{ik}\right)}. \tag{11}$$

The guessing parameter g_j indicates the chance level of a correct item response for attribute mastery patterns that lack at least one attribute required by item j. The slipping parameter s_j is the incorrect response probability of all-mastering-attribute mastery patterns required by item j. Both g_j and s_j can be represented as functions of the ideal responses

$$\eta_j^{DINA}(\boldsymbol{\alpha}_l) = \prod_{k=1}^{K}\alpha_{lk}^{q_{jk}}. \tag{12}$$

The ideal response represents the response of an individual who belongs to the lth attribute mastery pattern for item j without errors. Then, g_j and s_j are represented as conditional probabilities:

$$g_j = P\left(X_j = 1 \mid \eta_j^{DINA}(\boldsymbol{\alpha}_l) = 0\right), \tag{13}$$

$$s_j = P\left(X_j = 0 \mid \eta_j^{DINA}(\boldsymbol{\alpha}_l) = 1\right). \tag{14}$$

Using the item response probabilities, the centroid parameter under the DINA model is:

$$\theta_{jl} = g_j^{1-\eta_j^{DINA}(\boldsymbol{\alpha}_l)}(1-s_j)^{\eta_j^{DINA}(\boldsymbol{\alpha}_l)}. \tag{15}$$

Assuming a cross-entropy loss, which is $-\log y$, the likelihood-based loss function for the DINA model is:

$$\mathcal{L}(\mathcal{A},\boldsymbol{\mu},\boldsymbol{\pi}) = -\sum_{i=1}^{I}\sum_{l=1}^{L}\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l)\left[\sum_{j=1}^{J}\left\{x_{ij}\log\theta_{jl}+(1-x_{ij})\log\left(1-\theta_{jl}\right)\right\}+\log\pi_{\boldsymbol{\alpha}_l}\right]. \tag{16}$$

We assume $h(\pi_{\boldsymbol{\alpha}_l}) = -\log\pi_{\boldsymbol{\alpha}_l}$. The loss function defined in Equation 16 is equivalent to a negative log complete likelihood function. Therefore, minimizing equation 16 corresponds to maximizing the likelihood function; the minimizers $\left\{\hat{\mathcal{A}},\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\pi}}\right\}$ can be considered as the maximum likelihood estimate.

Subsequently, we examine the NPC and GNPC methods. Following Chiu and Douglas (2013), the loss function in the NPC method is defined by the Hamming distance between the individual item response vector and the ideal response vector:

$$\ell\left(\boldsymbol{x}_i,\boldsymbol{\mu}_{\boldsymbol{\alpha}_l}\right) = \sum_{j=1}^{J}\ell\left(x_{ij},\mu_{j,\boldsymbol{\alpha}_l}\right) = \sum_{j=1}^{J}\left|x_{ij}-\eta_j^{DINA}(\boldsymbol{\alpha}_l)\right|. \tag{17}$$

In the NPC method, the centroid parameter is the ideal response $\mu_{j,\boldsymbol{\alpha}_l}$. The NPC estimates are obtained to minimize Equation 17 for each individual:

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_l}{\operatorname{argmin}}\sum_{j=1}^{J}\left|x_{ij}-\eta_j^{DINA}(\boldsymbol{\alpha}_l)\right|, \forall i. \tag{18}$$

Clearly, the Hamming distance is a loss function, and the NPC method is a loss function-based estimation method.

As introduced in Chiu et al. (2018), the GNPC is a type of generalization that employs DINA-type and deterministic inputs noisy, "or" gate (DINO; Templin & Henson, 2006)-type ideal responses to define a

generalized ideal response. The DINO-type ideal response is

$$\eta_j^{DINO}(\boldsymbol{\alpha}_l) = 1 - \prod_{k=1}^{K} (1-\alpha_{lk})^{q_{jk}}, \tag{19}$$

and $\eta_j^{DINO}(\boldsymbol{\alpha}_l)$ becomes one if pattern l masters at least one attribute required for item j; otherwise, it becomes 0. The generalized ideal response is then defined as

$$\eta_j^{(w)}(\boldsymbol{\alpha}_l) = w_{jl}\eta_j^{DINA}(\boldsymbol{\alpha}_l) + \left(1-w_{jl}\right)\eta_j^{DINO}(\boldsymbol{\alpha}_l), \tag{20}$$

where $w_{jl} \in [0,1]$ is a weight parameter that determines an item's tendency. If the item is more like DINA or conjunctive, w_{jl} is close to one. By contrast, a w_{jl} near zero means that the item is DINO-like or disjunctive in nature. The GNPC assumes a Euclidean distance for its loss function

$$d\left(\boldsymbol{x}_j, \boldsymbol{\eta}^{(w)}(\boldsymbol{\alpha}_l)\right) = \sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)\left(x_{ij} - \eta_j^{(w)}(\boldsymbol{\alpha}_l)\right)^2, \tag{21}$$

where $\boldsymbol{\eta}^{(w)}(\boldsymbol{\alpha}_l) = \left(\eta_1^{(w)}(\boldsymbol{\alpha}_l), \eta_2^{(w)}(\boldsymbol{\alpha}_l), \ldots, \eta_J^{(w)}(\boldsymbol{\alpha}_l)\right)^\top$. The weight parameter is estimated via $\dot{w}_{jl} = 1 - \sum_{i=1}^{I}\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)x_{ij} / \sum_{i=1}^{I}\mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)$. The loss function of the GNPC is

$$\mathcal{L}(\{\mathcal{A}, W\}) = \sum_{j=1}^{J}\sum_{l=1}^{L} d\left(\boldsymbol{x}_j, \boldsymbol{\eta}^{(w)}(\boldsymbol{\alpha}_l)\right) = \sum_{j=1}^{J}\sum_{l=1}^{L}\sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l)\left(x_{ij} - \eta_j^{(w)}(\boldsymbol{\alpha}_l)\right)^2, \tag{22}$$

where $W = \left\{w_{jl}\right\}_{j,l=1}^{J,L}$. The GNPC requires iterative updates of weight w_{jl} and attribute mastery patterns. The detailed update rule is described in Chiu et al. (2018). Note that if η_j^{DINA} and η_j^{DINO} are not distinguished for some items and attribute patterns, the weight value is fixed to a value close to zero or one. See Chiu et al. (2018) for a detailed discussion.

As demonstrated above, parametric and nonparametric estimation methods can be treated in a unified loss function-based framework (C. Ma, de la Torre, & Xu, 2023). However, these loss function-based parameter estimates usually only provide point estimates, and uncertainty quantification in the parameter estimates has been considered less serious. Furthermore, different specifications of the measurement model precipitate significantly different attribute mastery patterns (e.g., Li et al., 2016). However, assessing all possible measurement models for all test items may be difficult. The GNPC is a promising estimation method that can be used in varied situations, even when the measurement model is unknown. However, prior knowledge of the weight parameters in the GNPC is often not considered. These problems can be solved using the GB method introduced in the following section.

3. GB method for DCMs

3.1. Construction of the generalized posterior

The GB method is a decision theory under a model misspecification situation (Bissiri et al., 2016). In other words, the assumed model may not accurately represent the true data-generating process, or the relationship between the model parameters and data may not be described via the assumed model, which is known as the $\mathcal{M}$-open situation (Bissiri et al., 2016, p. 1111). The GB method is a coherent belief update procedure that uses a loss function even in the $\mathcal{M}$-open situation. Thus, the GB method extends the applicability of the typical Bayesian methods, which require a likelihood function.

Let datasets and parameter sets be $\boldsymbol{y}$ and $\boldsymbol{\Theta}$, respectively. Additionally, the loss function and prior distribution are $\ell(\boldsymbol{y};\boldsymbol{\Theta})$ and $p(\boldsymbol{\Theta})$. Then, the generalized posterior of the parameter $p(\boldsymbol{\Theta} \mid \boldsymbol{y})$ is:

$$p(\boldsymbol{\Theta} \mid \boldsymbol{y}) \propto \exp\left(-\omega\ell(\boldsymbol{y};\boldsymbol{\Theta})\right)p(\boldsymbol{\Theta}), \tag{23}$$

where ω is called the learning rate, a tuning parameter that controls a dataset's importance. Methods for determining the learning rate are still being studied (Wu & Martin, 2023); notably, no standard has been established thus far. The generalized posterior function is the result of updating the prior distribution based on the loss function. If we select a negative log-likelihood function for the loss function and ω = 1, the generalized posterior becomes the usual Bayesian posterior function.

Bissiri et al. (2016), pp. 1106–1107) discussed some of the validity requirements for loss functions. First, the solution of the loss function must exist. Second, the loss function must satisfy the following condition:

$$0 < \int \exp\left(-\ell\left(\boldsymbol{y};\boldsymbol{\Theta}\right)\right) p\left(\boldsymbol{\Theta}\right) d\boldsymbol{\Theta} < \infty. \tag{24}$$

Some major loss functions considered in this study, such as the Hamming distance, Euclid distance, or cross-entropy loss, satisfy the above integral conditions. Additionally, Bissiri et al. (2016), p. 1107) identified natural assumptions for deriving a generalized posterior from a loss function. We should also point out that we only need to construct loss functions given a set of data for only the parameter of interest to employ the GB method. In this article, the GB method employed the loss function for attribute mastery patterns. The loss function is based on the GNPC: quadratic of Euclid distance. It satisfies the above conditions and is valid.

3.2. General form of the GB method for DCMs

The general form of the GB method for DCMs can be expressed using Equations 6 and 23:

$$\begin{aligned} p\left(\{\mathcal{A},\boldsymbol{\mu},\boldsymbol{\pi}\} \mid X\right) &\propto \exp\left(-\omega\left\{\mathcal{L}\left(\{\mathcal{A},\boldsymbol{\mu},\boldsymbol{\pi}\}\right)\right\}\right) p\left(\boldsymbol{\mu}\right) p\left(\boldsymbol{\pi}\right), \\ &\propto \exp\left(-\omega \sum_{l=1}^{L} \sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_l} \left\{\ell\left(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}_l}\right) + h\left(\pi_{\boldsymbol{\alpha}_l}\right)\right\}\right) p\left(\boldsymbol{\mu}\right) p\left(\boldsymbol{\pi}\right). \end{aligned} \tag{25}$$

The penalty term was $h\left(\pi_{\boldsymbol{\alpha}_l}\right) = -\log \pi_{\boldsymbol{\alpha}_l}$.

Using the GNPC loss function defined in Equation 22 and adding a penalty term for the mixing parameter $\boldsymbol{\pi}$, the generalized posterior is:

$$p\left(\{\mathcal{A},\boldsymbol{\mu},\boldsymbol{\pi}\} \mid X\right) \propto \exp\left(-\omega \sum_{l=1}^{L} \sum_{i=1}^{I} \left\{\mathcal{I}\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right) \left[\sum_{j=1}^{J} \left(x_{ij} - \eta_j^{(w)}\left(\boldsymbol{\alpha}_l\right)\right)^2\right] - \log \pi_{\boldsymbol{\alpha}_l}\right\}\right) p(W) p\left(\boldsymbol{\pi}\right). \tag{26}$$

Notably, we treat weight W as a parameter and assume a prior instead of a centroid parameter $\boldsymbol{\mu}$ because the centroid parameter $\boldsymbol{\eta}^{(w)}\left(\boldsymbol{\alpha}_l\right)$ is determined by two ideal responses and weight parameters; thus, it is natural. Priors for the mixing parameters and weight parameters are assumed Dirichlet and Beta distributions:

$$p\left(\boldsymbol{\pi}\right) \propto \prod_{l=1}^{L} \pi_l^{\delta_l^0 - 1}, \tag{27}$$

$$p(W) \propto \prod_{j=1}^{J} \prod_{l=1}^{L} w_{jl}^{a_{jl}^0 - 1} \left(1 - w_{jl}\right)^{b_{jl}^0 - 1}, \tag{28}$$

where $\delta_1^0, \delta_2^0, \ldots, \delta_L^0 \geq 0, \sum_l \delta_l^0 = 1$, and $a_{jl}^0, b_{jl}^0 \geq 0$.

The posterior was numerically obtained using MCMC techniques, such as Metropolis–Hastings within the Gibbs sampling method, or MCMC software, such as JAGS (Plummer, 2003) or Stan (Carpenter et al., 2017). The conditional distribution of $\boldsymbol{\alpha}_i$ is categorical:

$$p\left(\boldsymbol{\alpha}_i \mid \boldsymbol{x}_i, W, \boldsymbol{\pi}\right) \propto \prod_{l=1}^{L} r_{il}^{\mathcal{I}\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right)}, r_{il} = \frac{\rho_{il}}{\sum_l \rho_{il}}, \rho_{il} = \exp\left(-\omega \mathcal{I}\left(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l\right) \left[\sum_{j=1}^{J} \left(x_{ij} - \eta_j^{(w)}\left(\boldsymbol{\alpha}_l\right)\right)^2\right] - \log \pi_{\boldsymbol{\alpha}_l}\right). \tag{29}$$

The conditional distribution of the mixing parameters was a Dirichlet distribution:

$$p(\boldsymbol{\pi} \mid X) \propto \prod_{l=1}^{L} \pi_l^{\delta_l^* - 1}, \delta_l^* = \omega \sum_{i=1}^{I} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l) + \delta_l^0. \tag{30}$$

The conditional distribution of the weight parameter is not easily expressed; therefore, its MCMC update was performed using the Metropolis–Hastings method. The candidate was generated by a random walk using a uniform distribution: $w_{jl}^{(\text{cand})} = w_{jl}^{(\text{now})} + u, u \sim \text{Unif}(-0.05, 0.05)$. Using the above distributions and updating rules, the MCMC for a generalized posterior is numerically approximated as follows: The mixing and weight parameters were initialized as $\left\{\boldsymbol{\pi}^{(0)}, W^{(0)}\right\}$, and the hyperparameters were set to $\delta_l^0 = 1, \forall l$ and $a_{jl}^0 = 2, b_{jl}^0 = 1, \forall j, l$ for example. Then, at the tth MCMC iteration $(t = 1, 2, \ldots, T \in \mathbb{N})$, $\boldsymbol{\alpha}_i^{(t+1)}$ is generated from the categorical distribution expressed in Equation 29 with the tth MCMC sample of the parameter set at $\left\{\boldsymbol{\alpha}^{(t)}, W^{(t)}\right\}$. The $(t+1)$th MCMC sample of the mixing parameter is generated from the Dirichlet distribution shown in Equation 30 using $\mathcal{A}^{(t+1)}$. $W^{(t+1)}$ is obtained using the Metropolis–Hastings method.

Under the hyperparameter setting, the Dirichlet distribution for mixing parameters becomes a uniform distribution. This represents a scenario in which we have no information about the population attribute mastery ratio. In this means, the prior of the attribute mastery pattern has almost no information. The mean and SD of the prior of the weight parameter are 0.667 and 0.236, respectively. Under this setting, interval $[0.158, 0.987]$ covers 95% of the support of the parameter. The data analyst expected the items in the test to have a slightly conjunctive nature, which means the items behave more like in the DINA model than in the DINO model. However, the expectation is not particularly strong because the interval covering 95% of the support of the parameter is wide. This interpretation indicates that the prior conveys some information about the weight parameters.

4. Mathematical properties of the proposed method: Consistency of the MAP estimators

First, we formally introduce the estimators under the GB framework and subsequently discuss their statistical behaviors under certain conditions. The appendix provides the full proofs. In this work, we assume that the item responses were generated from the Bernoulli distribution with parameter Θ defined by Equation 3; the attribute mastery patterns were generated from a categorical distribution with a mixing parameter $\boldsymbol{\pi}$. Although several alternatives exist, MAP estimation provides a relatively natural and simple choice. Furthermore, MAP estimators of the GB method $\left(\widehat{\mathcal{A}}, \widehat{\Theta}, \widehat{\boldsymbol{\pi}}\right)$ are estimators of the true parameters $\left(\mathcal{A}^0, \Theta^0, \boldsymbol{\pi}^0\right)$ in the data-generating process. These are obtained by minimizing the loss function of $(\mathcal{A}, \Theta, \boldsymbol{\pi})$ under the constraint imposed by the Q-matrix, as follows:

$$\mathcal{L}(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X) = \sum_{i=1}^{I} \left(\sum_{j=1}^{J} \ell(X_{ij}, \theta_{j,\boldsymbol{\alpha}_i}) + h(\pi_{\boldsymbol{\alpha}_i}) \right) + \sum_{j,l} \log f_{j,\boldsymbol{\alpha}_l}(\theta_{j,\boldsymbol{\alpha}_l}) + \sum_{l=1}^{L} \log g_{\boldsymbol{\alpha}_l}(\pi_{\boldsymbol{\alpha}_l}), \tag{31}$$

where $h(\cdot)$ is a continuous nonincreasing regularization function of the proportion parameters π_α, often taken as $h(\pi) = -\log \pi$; $f_{j,\boldsymbol{\alpha}}$ and $g_{\boldsymbol{\alpha}}$ are the prior density functions of $\theta_{j,\alpha}$ and $\pi_{\boldsymbol{\alpha}}$, respectively. Note that we consider a model sequence indexed by (I, J), where both I and J tend to infinity, while K is held constant.

Several regularity conditions are required to ensure the consistency of MAP estimators. The first assumption is as follows.

Assumption 1. There exists $\delta_1, \delta_2 > 0$ such that

$$\min_{1 \le j \le J} \left\{ \min_{\boldsymbol{\alpha}_l \circ \boldsymbol{q}_j^0 \neq \boldsymbol{\alpha}_{l'} \circ \boldsymbol{q}_j^0} \left(\theta_{j,\boldsymbol{\alpha}_l}^0 - \theta_{j,\boldsymbol{\alpha}_{l'}}^0 \right)^2 \right\} \ge \delta_1,$$

and $\delta_2 \le \min_{j,\alpha} \theta_{j,\alpha}^0 < \max_{j,\alpha} \theta_{j,\alpha}^0 \le 1 - \delta_2$.

The first condition in Assumption 1 serves as an identification condition for local latent classes at each item level. The gap denoted by δ measures the separation between the latent classes, thereby quantifying the signals' strength. The second condition in Assumption 1 keeps the true parameters away from the boundaries of the parameter space to prevent unusual behaviors of the element-wise loss.

Assumption 2 pertains to the discrete structures of $\mathbf{Q}$ and is expressed as the following.

Assumption 2. All proportion parameters π_α are strictly greater than zero, and there exist $\{\delta_J\} \subset (0, \infty)$ such that

$$\min_{1 \le k \le K} \frac{1}{J} \sum_{j=1}^{J} \mathcal{I}\left\{\boldsymbol{q}_j^0 = \boldsymbol{e}_k\right\} \ge \delta_J. \tag{32}$$

This assumption holds that $\mathbf{Q}$ includes an increasing number of identity submatrices, $\mathbf{I}_K$, as J grows. Notably, by attaching the subscript J to the lower bound (32) in Assumption (2), we allow it to decrease to zero as J approaches infinity. As the following theorems show, if the rate at which variable δ_J decreases meets certain mild requirements, the consistency of $\left(\widehat{\mathcal{A}}, \widehat{\Theta}\right)$ can be ensured.

The subsequent assumption concerns the element-wise loss function ℓ.

Assumption 3. The loss function $\ell(X, \theta)$ is twice continuously differentiable in θ on $(0,1)$ and $\exists b_U > b_L > 0$ such that $b_L \le \partial_{\theta^2} \ell(R, \theta) \le b_U$ for θ in a compact subset of $(0,1)$. The total loss (31) is minimized at class means given the subjects' membership, as in, $\widehat{\theta}_{j,\boldsymbol{\alpha}} = \sum_{i=1}^{I} \mathcal{I}\{\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}\} X_{ij} / \sum_{i=1}^{I} \mathcal{I}\{\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}\}$.

Assumption 3 imposes smoothness conditions on the element-wise loss function, rendering it convex. The upper bound of the second derivative is necessary to control the remaining term in the expansion of the first-order condition, and the lower bound allows us to quantify the estimator drift caused by the given priors. For the sample average assumption, we can verify that both ℓ^2 and cross-entropy loss functions satisfy Assumption 3.

Assumption 4 states that the true parameters minimize the element-wise loss functions and quantify the deviations when θ is not a true parameter. This assumption is expressed as follows:

Assumption 4. There exist constants $\eta \ge 2, c > 0$ such that

$$\mathbb{E}[\ell(X_{ij}, \theta)] - \mathbb{E}\left[\ell\left(X_{ij}, \theta^0_{j,\alpha_i^0}\right)\right] \ge c\left|\theta - \theta^0_{j,\alpha_i^0}\right|^{\eta}. \tag{33}$$

Assumption 4 holds for both the ℓ^2 loss and the cross-entropy loss.

Assumption 5 is a technical assumption that allows us to control the effects of prior distributions on the estimators.

Assumption 5. $h(\cdot)$ in (31) is a continuous nonincreasing function of the proportion parameters, and $C > c > 0$ exists such that for any j and $\boldsymbol{\alpha}, C > f_{j,\alpha}, g_\alpha > c$ on a compact parameter subspace of $(0,1)$.

We can verify that the Dirichlet and Beta distributions satisfy this assumption.

Under the aforementioned regularity conditions, we demonstrate the consistency properties of the GB method with constraints for different attribute mastery patterns $\boldsymbol{\alpha}_l$ and $\boldsymbol{\alpha}_{l'}, l \ne l'$ (Ma, de la Torre, & Xu, 2023; Xu, 2017):

$$\left(\boldsymbol{\alpha}_l \circ \boldsymbol{q}_j = \boldsymbol{\alpha}_{l'} \circ \boldsymbol{q}_j\right) \Longrightarrow \left(\theta_{j,\boldsymbol{\alpha}_l} = \theta_{j,\boldsymbol{\alpha}_{l'}}\right), \tag{34}$$

where $\boldsymbol{\alpha} \circ \boldsymbol{q}_j = \left(\alpha_1 \cdot q_{j1}, \ldots, \alpha_K \cdot q_{jk}\right)$ denotes the element-wise product of binary vectors $\boldsymbol{\alpha}$ and $\boldsymbol{q}_j$. This implies that the item response parameter $\theta_{j,\boldsymbol{\alpha}}$ depends only on whether the attribute mastery pattern $\boldsymbol{\alpha}$ contains the required attributes $\mathcal{K}_j := \left\{k \in [K]; q_{jk} = 1\right\}$ for item j.

Based on the above five assumptions, we can derive consistent results for the GB method. The following main theorem first validates the clustering consistency of the GB method under the constraint (34), providing a bound for its convergence rate in recovering the attribute mastery patterns.

Theorem 1 (Clustering Consistency). *Consider* $\left(\widehat{\mathcal{A}},\widehat{\Theta},\widehat{\boldsymbol{\pi}}\right) = \arg\min_{(\mathcal{A},\Theta,\boldsymbol{\pi})} \mathcal{L}\left(\mathcal{A},\Theta,\boldsymbol{\pi} \mid X\right)$ *under the constraint (34). When* $I,J \to \infty$ *jointly, suppose* $\sqrt{J} = O\left(I^{1-c}\right)$ *for some small constant* $c \in (0,1)$ *. Under Assumption 1 to Assumption 5. the clustering error rate is:*

$$\frac{1}{I}\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i^0\right\} = o_p\left(\frac{(\log J)^{\tilde{\varepsilon}/\eta}}{\delta_J (J)^{1/\eta}}\right), \tag{35}$$

where for a small positive constant $\tilde{\varepsilon} > 0$.

Theorem 1 bounds the error of the estimator $\hat{\mathcal{A}}$, which establishes the clustering consistency of the MAP estimators of the GB method, allowing the rate δ_J to go to zero. Notably, the scaling condition only assumes that J goes to infinity jointly with I, but at a slower rate.

The following result demonstrates that the MAP estimator of the item parameters can be uniformly estimated consistently as $I,J \to \infty$:

Theorem 2 (Item parameters consistency). *Under Assumptions 1 to 5 and the scaling conditions given in Theorem 1, we have the following uniform consistency result for all* $j \in [J]$ *and* $\boldsymbol{\alpha} \in \{0,1\}^K$:

$$\max_{j,\alpha} \left|\widehat{\theta}_{j,\alpha} - \theta_{j,\alpha}^0\right| = o_p\left(\frac{1}{\sqrt{I^{1-\tilde{c}}}}\right) + o_p\left(\frac{(\log J)^{\tilde{\varepsilon}/\eta}}{\delta_J (J)^{1/\eta}}\right), \tag{36}$$

where $\tilde{c}$ and $\tilde{\varepsilon}$ are small positive constants.

On the first error term, the condition $\pi_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$ ensures that with probability one, there are enough samples within each class to provide accurate estimates of item parameters. Notably, $\grave{c}$ the first error term arises because the number of parameters approaches infinity jointly with the sample size I, which causes a slight deviation from the optimal error rate of $O_p\left(1/\sqrt{I}\right)$. The maximum deviation $\max_{j,\alpha}\left|\widehat{\theta}_{j,\alpha} - \theta_{j,\alpha}^0\right|$ is also affected by the classification error. This is indicated in the second error term $o_p\left((\log J)^{\tilde{\varepsilon}}/\delta_J\sqrt{J}\right)$.

We can easily establish the consistency of the mixing parameter estimator $\hat{\boldsymbol{\pi}}$. When $h(\pi) = -\log\pi$, the mixing parameters will be estimated as the sample average form $\sum_i \mathcal{I}\left\{\boldsymbol{\alpha}^0 = \boldsymbol{\alpha}\right\}/I$, which converge in probability to π_{α}^0 because of the clustering consistency.

Corollary 1 (Proportion parameters consistency). *Under Assumptions 1 to 5 and the scaling conditions given in Theorem 1, when* $h(\pi)$ *is taken as* $-\log\pi$*, we have* $\widehat{\pi}_{\alpha} \xrightarrow{P} \pi_{\alpha}^0$.

5. Simulation study

This section compares the previous (G)NPC and the corresponding GB methods using the loss functions in NPC and GNPC, named as GBNPC and GBGNPC, respectively. This simulation study primarily aims to assess the behavior of the GB method's parameter estimates under finite small sample and item situations. As the GBNPC and GNPC are based on loss function in nonparametric methods, the most interesting parameters are attribute mastery patterns. In this simulation study, we mainly focus on the comparisons of the point estimates from these methods. To represent the uncertainty of the estimates, we also present attribute mastery probabilities using the GBGNPC and GBNPC methods, which indicate the benefit of the proposed method against the nonparametric methods.

The code for this simulation study is available on the Open Science Framework (OSF) webpage: https://osf.io/sau6j/.

5.1. Simulation settings

Five factors are manipulated in the simulations. All factors had two conditions; hence, $2^5 = 32$ simulation settings were used. The first factor was the data-generating model: DINA or general DCM (e.g., LCDM). The DINA model condition is a simpler data-generating situation, whereas the general DCM model is

Table 1. The four-attribute Q-matrix

Item	Attribute			
	1	2	3	4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	0	0	0
6	0	1	0	0
7	0	0	1	0
8	0	0	0	1
9	1	1	0	0
10	1	0	1	0
11	1	0	0	1
12	0	1	1	0
13	0	1	0	1
14	0	0	1	1
15	1	1	1	0
16	1	1	0	1
17	1	0	1	1
18	0	1	1	1
19	1	1	1	1

more complex. The second factor was the Q-matrix; four or five attributes are listed in Tables 1 and 2. Table 1 contains 19 items: eight simple items (i.e., measuring only one attribute), six items measuring two attributes, five items requiring three attributes, and the most complex item measuring all four attributes. Table 2 lists 30 items: eight simple items, ten items measuring two attributes, and ten items measuring three attributes.

Sample size was the third factor, with 30 or 300 participants assumed. The sample size setting of 30 participants mimicked classroom size. The sample size of 300 participants was 10 times larger than that of other classroom settings. The fourth condition was attribute correlation: independent ($\rho = 0$) or highly correlated ($\rho = 0.8$). The independent attribute condition was unrealistic but represented an ideal condition. The highly correlated condition was more realistic because the DCMs application indicated a high correlation among attributes (e.g., von Davier, 2008). The fifth condition was item quality. The high-item-quality condition indicates a high description of all nonmastering attributes and all perfectly mastering attributes. Under the high-item-quality condition, the correct response probability of the all nonmastering pattern was 0.1 and that of the all-mastering pattern was 0.9. On the other hand, in the low-item-quality condition, the corresponding probabilities were 0.3 and 0.7. The correct item response probabilities of the intermediate mastering patterns are generated based on Yamaguchi and Templin (2022b) or Yamaguchi and Templin (2022a).

The data generation process used herein was similar to those in previous studies, such as Chiu and Douglas (2013), Yamaguchi and Templin (2022b), and Yamaguchi and Templin (2022a). First, for each individual, we generated a continuous latent variable vector $\tilde{\boldsymbol{\alpha}}_i = (\tilde{\alpha}_{i1}, \ldots, \tilde{\alpha}_{iK})^\top$ from K-dimensional normal distributions with zero means and compound symmetry covariance with a correlation of 0 or

Table 2. The five-attribute Q-matrix

Item	Attribute					Item	Attribute				
	1	2	3	4	5		1	2	3	4	5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

0.8, and variances of 1. Subsequently, the continuous latent variable vector $\grave{\alpha}_{i1}$ was converted into an attribute mastery pattern. More precisely, if $\tilde{\alpha}_{ik}$ was greater than $\Phi(k/(1+K))^{-1}, \alpha_{ik} = 1$; otherwise, $\alpha_{ik} = 0$, where $\Phi(\cdot)^{-1}$ is the inverse cumulative normal distribution function. The simulated item responses were randomly generated using these attribute mastery patterns, an assumed data-generating model (DINA or general DCM), and item response probabilities. As mentioned in the previous section, the parameters of the priors in the GB method were set to $a_{jl}^0 = 2, b_{jl}^0 = 1, \forall j, l$ and $\delta_l^0 = 1, \forall l$. The step size of the Metropolis update was fixed at 0.05. A one-chain MCMC with 1,000 iterations was employed. The first 500 iterations are discarded as the burn-in period; therefore, 500 MCMC samples were used to approximate the posterior distributions.

The main target parameter is attribute masteries, and they are categorical latent variables. However, common MCMC convergence criteria, such as Gelman-Rubin's $\grave{R}$, are for continuous variables, which means the indicators may not be applicable to categorical variables. Therefore, performing a convergence check of categorical variables in MCMC is not easy in this context. Instead of directly checking for the convergence of attribute mastery, we calculated the average correlations of the attribute mastery probabilities, which we estimated for the first and second halves of the MCMC iterations after the burn-in period. If the estimated results of the attribute mastery probabilities with the first half after the burn-in period are consistent with those of the later MCMC iterations, we consider the attribute mastery results to be stable.

The attribute mastery pattern of the i-individual was calculated based on the posterior attribute mastery probabilities. If the probability of the kth attribute was <0.5, the attribute was considered mastered. Each estimation method was evaluated using two attribute mastery recovery indices: attribute-level agreement ratio (AAR) and pattern-level agreement ratio (PAR). AAR and PAR were calculated as follows:

$$\text{AAR}_k = \frac{1}{IM} \sum_{m=1}^{M} \sum_{i=1}^{I} \mathcal{I}\left(\hat{\alpha}_{ik}^{(m)} = \alpha_{ik}^{(\text{True})}\right), \forall k \tag{37}$$

Table 3. The average correlations of attribute mastery probabilities estimated by first and second halves of MCMC iterations after the burn-in period

Data generating model	Sample size	Attribute correlation	Item quality	GBGNPC		GBNPC	
				Three attributes	Four attributes	Three attributes	Four attributes
DINA	30	0	High	.994	.994	.998	.998
			Low	.984	.983	.993	.994
		0.8	High	.996	.995	.998	.998
			Low	.984	.983	.995	.995
	300	0	High	.997	.998	.998	.999
			Low	.992	.993	.993	.995
		0.8	High	.999	.999	.999	.999
			Low	.992	.993	.995	.996
General	30	0	High	.990	.989	.996	.997
			Low	.981	.980	.994	.995
		0.8	High	.994	.994	.998	.998
			Low	.982	.982	.995	.995
	300	0	High	.997	.997	.997	.997
			Low	.992	.993	.994	.995
		0.8	High	.999	.999	.999	.999
			Low	.991	.993	.996	.996

Note: GBGNPC, generalized Bayesian method with generalized nonparametric loss function; GBNPC, generalized Bayesian method with nonparametric loss function.

$$\text{PAR} = \frac{1}{IM}\sum_{m=1}^{M}\sum_{i=1}^{I}\mathcal{I}\left(\hat{\boldsymbol{\alpha}}_i^{(m)} = \boldsymbol{\alpha}_i^{(\text{True})}\right), \tag{38}$$

where $\hat{\boldsymbol{\alpha}}_i^{(m)} = \left(\hat{\alpha}_{i1}^{(m)}, \hat{\alpha}_{i2}^{(m)}, \ldots, \hat{\alpha}_{iK}^{(m)}\right)^{\top}$ is an estimate of the attribute mastery pattern for individual i in the mth simulation, and $\boldsymbol{\alpha}_i^{(\text{True})} = \left(\alpha_{i1}^{(\text{True})}, \alpha_{i2}^{(\text{True})}, \ldots, \alpha_{iK}^{(\text{True})}\right)^{\top}$ is the true attribute vector of individual i, where M is the total number of simulations, which is $M = 100$.

5.2. Results

Table 3 shows the results, which indicate correlations >0.98. Therefore, this result can be interpreted as an indication that the MCMC iterations were stable and attribute mastery can be estimated from the MCMC samples after the burn-in period.

Figures 1 and 2 present the simulation results of the DINA data generation with four- and five-attribute Q-matrix conditions, respectively. In this simulation, the AARs and PARs of the two Q-matrix conditions demonstrated similar tendencies; therefore, our discussion here focuses on the four-attribute Q-matrix condition. The high-item-quality conditions presented in the four left panels of Figure 1 indicated that all four estimation methods provide high AARs and PARs. The low-item-quality conditions presented in the four right panels of Figure 1 indicated lower AARs and PARs than the high-item-quality conditions, and the low-item-quality conditions exhibited some differences among the

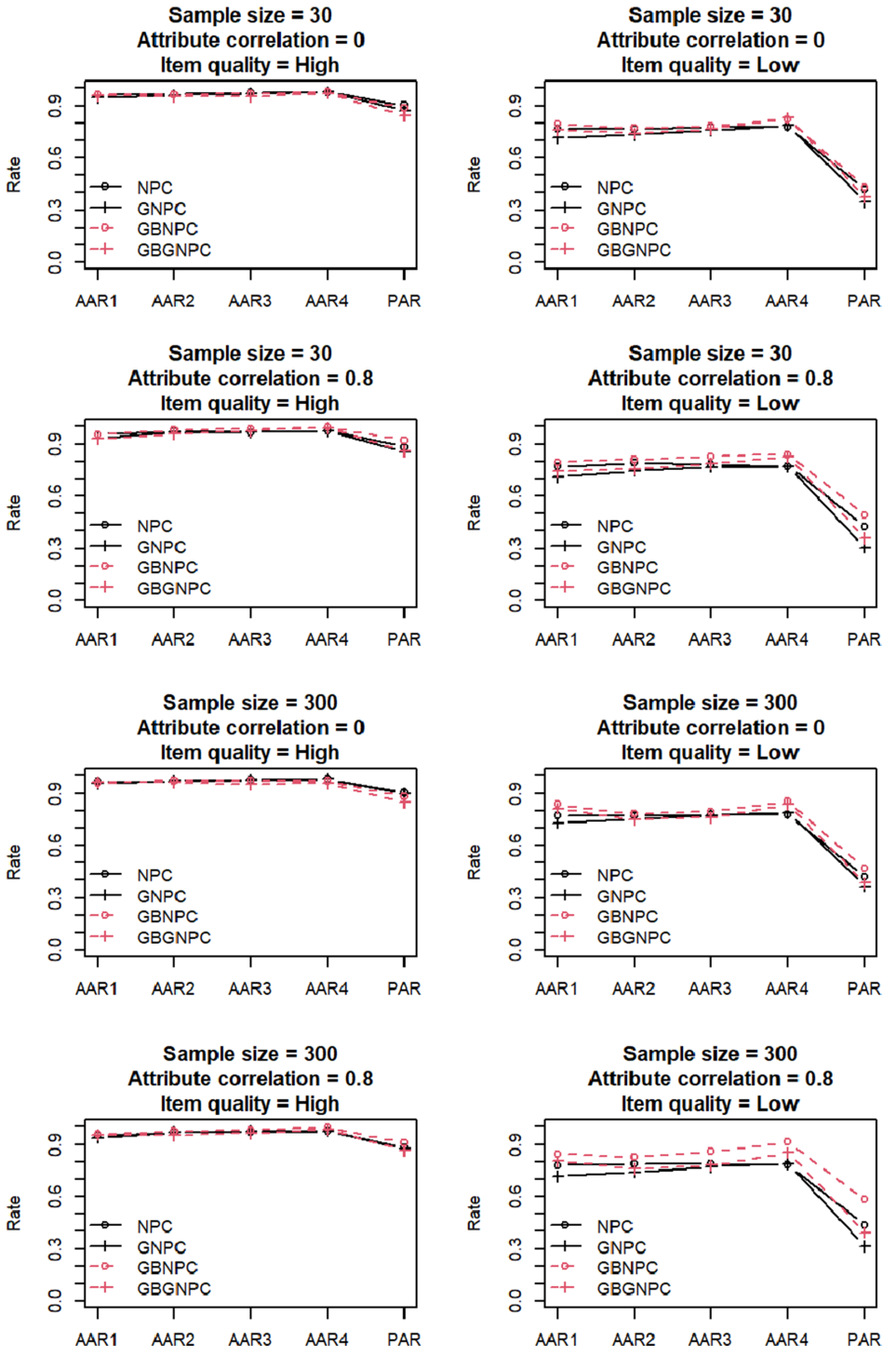

Figure 1. Simulation results of the DINA data generation with four-attribute **Q**-matrix conditions.

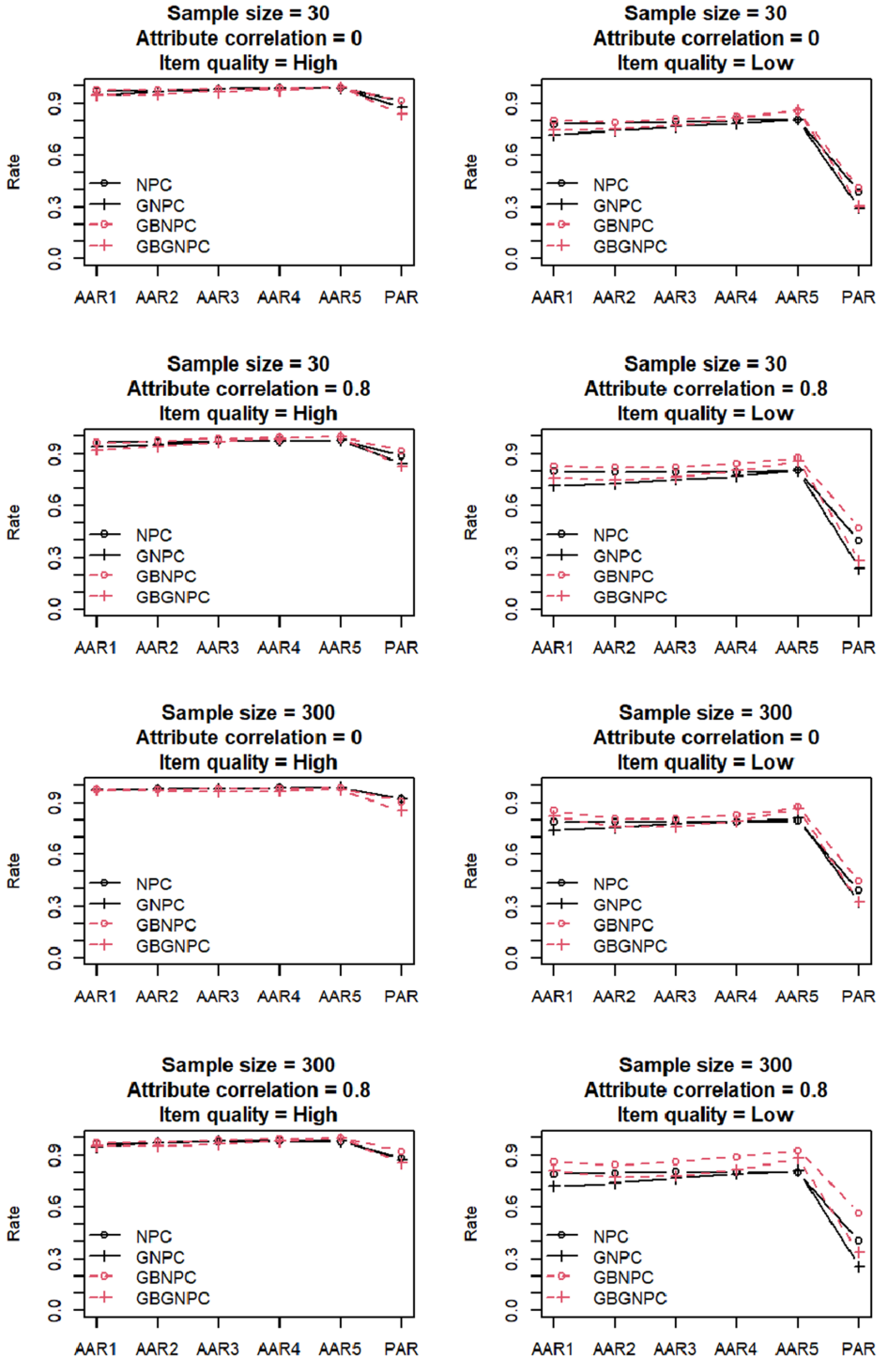

Figure 2. Simulation results of the DINA data generation with five-attribute **Q**-matrix conditions.

four estimation methods. Attribute correlations under low-item-quality conditions affected AARs and PARs more significantly. Furthermore, GBNPC and GBGNPC demonstrated higher AARs and PARs than the corresponding NPC and GNPC methods under the 30-sample size, 0.8 attribute correlation, and low-item-quality conditions. Interestingly, GBNPC exhibited the highest AARs and PARs under the 300-sample size, 0.8 attribute correlation, and low-item-quality conditions. Moreover, under these conditions, GBGNPC had similar AARs and PARs to the NPC, and the GNPC produced the least optimal result.

Figures 3 and 4 present the results of the general DCM data generation with four- and five-attribute Q-matrix conditions, respectively. Again, the AARs and PARs of the two Q-matrix conditions exhibited similar patterns; hereafter, we predominantly focus on results of the four-attribute Q-matrix conditions. Under high-item-quality conditions, GNPC and GBGNPC outperformed NPC and GBNPC. Furthermore, under high-attribute correlation conditions, GBGNPC was superior to GNPC; the same pattern was observed between GBNPC and NPC under the same conditions. In the low-item-quality conditions presented in the four right panels of Figure 3, GBGNPC and GBNPC tended to have higher AARs and PARs than GNPC and NPC. In particular, a sample size of 300, a high-attribute correlation, and low-item-quality conditions indicated better AARs and PARs for GBGNPC and GBNPC than GNPC or NPC.

We also checked attribute mastery probabilities of the GBGNPC and GBNPC methods that represented uncertainty of parameter estimates. Figures 5 and 6 represent box plots of average attribute mastery probabilities of the four- and five-attribute conditions under the DINA model-based data-generating process. Interestingly, the GBNPC method tended to show higher average attribute mastery probabilities than the GBGNPC. The differences between the GBGNPC and GBNPC methods were relatively small in the first attribute but the discrepancy became larger as the attribute number increased. The later attributes were more difficult to master and the number of individuals mastering them was small. These tendencies also occurred in the general data-generating process situations, which are shown in Figures 7 and 8. These posterior probabilities of attribute mastering represent estimation uncertainty, so we can carefully check the attribute mastery status. For example, the attribute mastery probabilities around cutoff values might represent indeterminacy of mastery or nonmastery. Such uncertainty quantification results cannot be obtained through the GNPC or NPC methods.

In summary, NPC and GBNPC tended to have higher AARs and PARs under DINA data generation, low-item-quality conditions, and high-attribute correlations. However, GBGNPC was sometimes similar to NPC under the DINA data generation conditions, whereas GNPC was the least optimal. By contrast, under general DCM data generation conditions, GBGNPC and GNPC performed better than GBNPC and GNPC for high-quality items. For low-quality items, GBGNPC and GBNPC performed better. Based on these results, GBGNPC appears the optimal choice for attribute mastery estimation. If the DINA type item response mechanism is confirmed, GBNPC is the optimal choice among the four estimation methods from the perspective of attribute recovery.

The possible reason for the superiority of the GBGNPC over the GNPC is prior settings. In our simulation setting, sample sizes were relatively small in the situations in which the nonparametric methods were employed. Under such conditions, estimation of weight parameters might be difficult for the GNPC, especially under the low-item-quality conditions. The GBGNPC, on the other hand, assumed priors for the weight parameters, and the prior conveyed information of item characteristics and succeeded in estimating attribute mastery patterns. Another reason may be that the GBGNPC can deal with uncertainty in parameter estimation. This means that the GNPC uses parameter estimates to minimize the loss function, which simply selects the attribute mastery pattern that provides the minimum value of loss function without considering the second or third best attribute mastery patterns. By contrast, the GBGNPC can consider and use the second-best attribute mastery pattern for estimating attribute mastery probabilities. If these considerations are correct, even if we use noninformation priors for the GNPC method, the GBGNPC may remain superior. The effects of prior settings are also an important topic for detailed research in future studies.

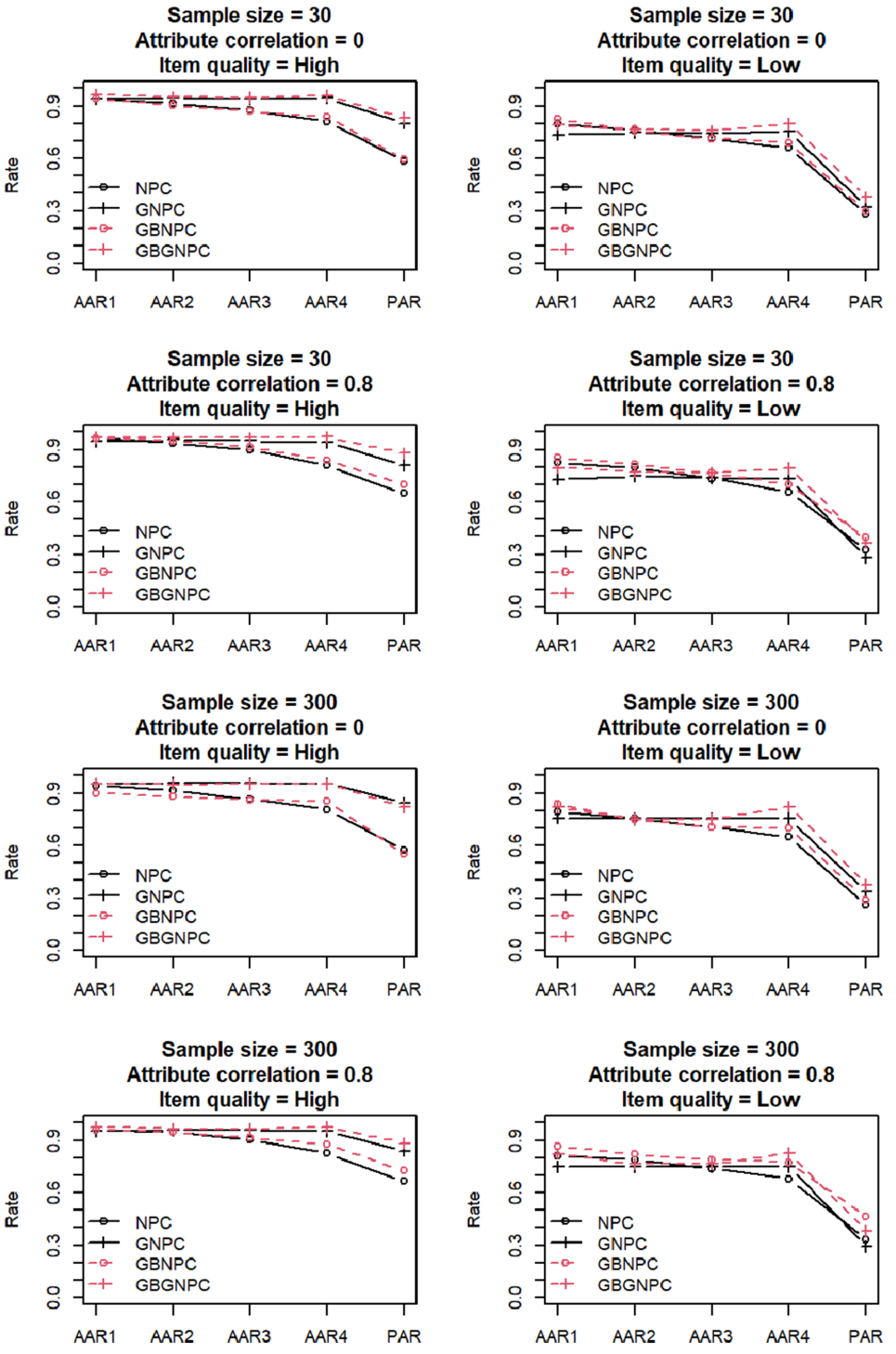

Figure 3. Simulation results of the general DCM data generation with four-attribute Q-matrix conditions.

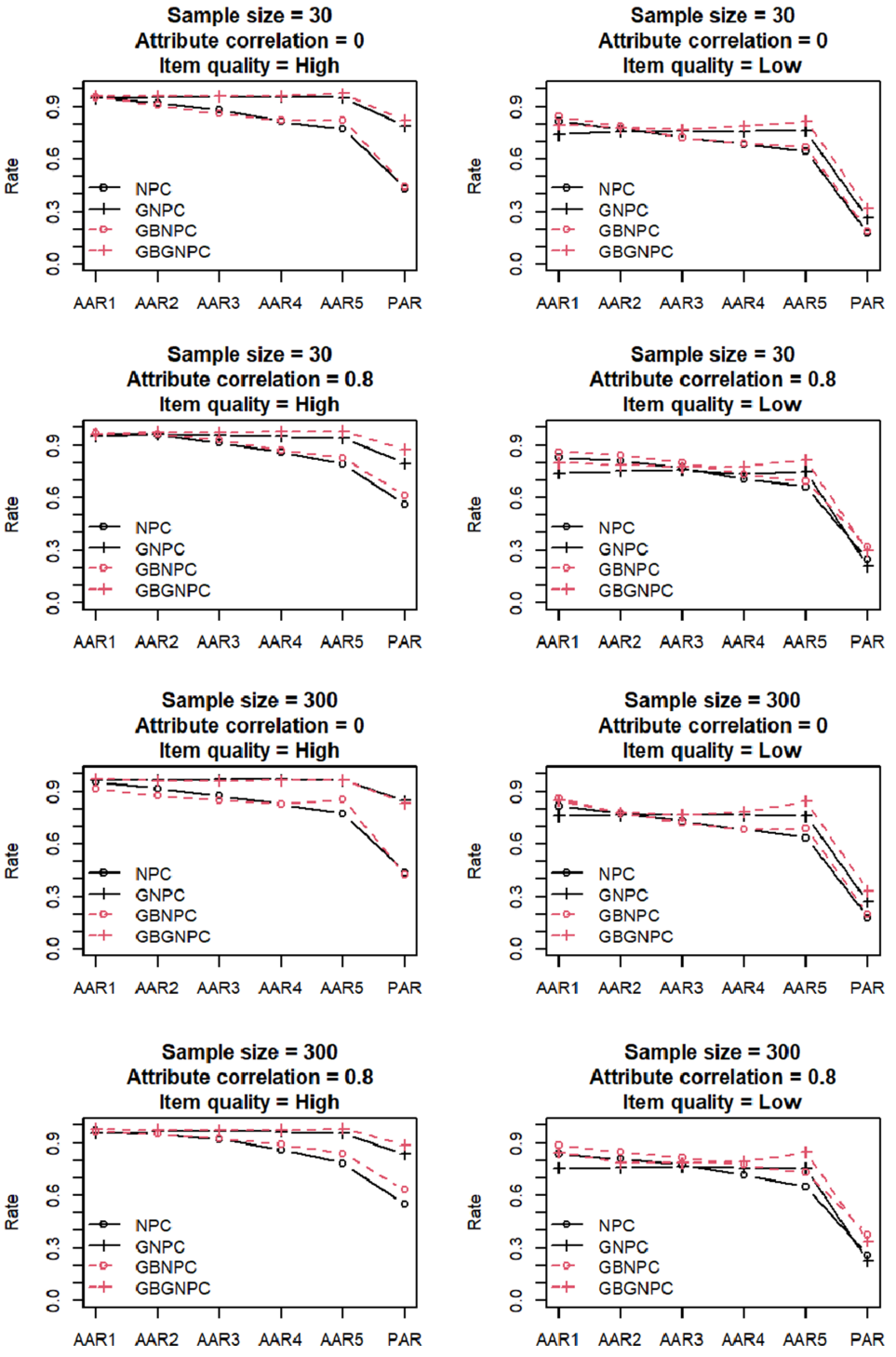

Figure 4. Simulation results of the general DCM data generation with five-attribute Q-matrix conditions.

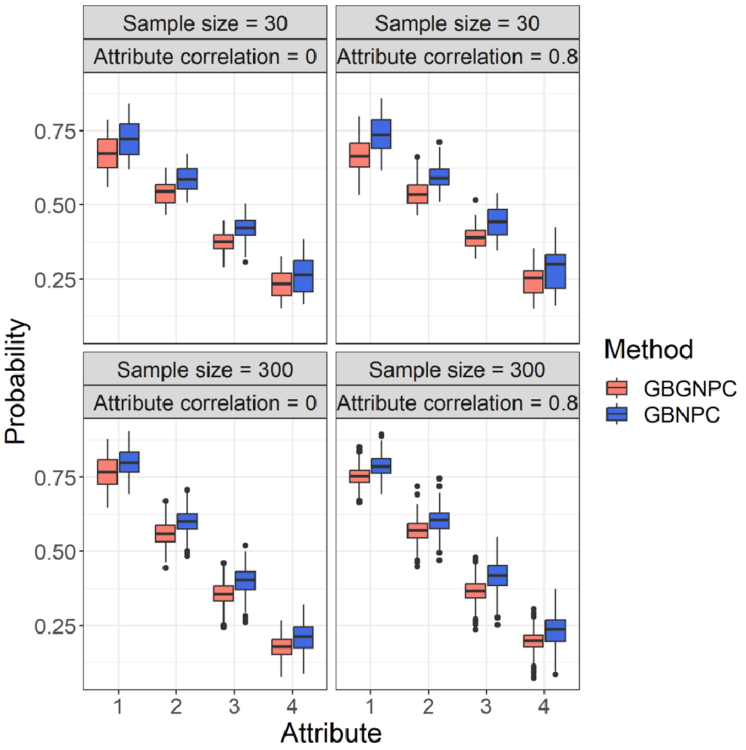

Figure 5. Box plots of attribute mastery probabilities of the DINA data generation with four-attribute Q-matrix conditions.

In addition, the effects of manipulated factors are discussed here. First, attribute correlation affected attribute mastery recovery. The GB method provided better attribute mastery recovery than the nonparametric methods. The nonparametric loss could not include such information, but the generalized posteriors have such information based on the data. The attributes generally correlate with each other, making the GB methods generally better than the nonparametric methods. We did not explicitly include a loss related to the attribute correlations, but the GB method allows us to include a loss term of the attribute correlations or attribute structure. This may be a good extension for constructing a loss function for the GB method.

Second, item quality also affected attribute mastery recovery results. The GB methods showed better results than the nonparametric methods, especially under low-item-quality conditions. Under such conditions, prior information might help to improve attribute recovery. This means the GB method can utilize not only the loss function but also prior information. This makes the GB method the preferred method compared to the current nonparametric methods, which cannot do this. Thus, based on the simulation study, the GB methods are always better than the nonparametric methods from the perspective of attribute mastery recovery.

6. Real data example

The real data example aimed to compare the four estimation methods used in the simulation study and examine how these estimation methods provide different attribute mastery results. This real data comparison provided an example of the behavior of the proposed GB method for DCMs.

To show the superiority of their proposed methods, Ma and Jiang (2021) used *k*-fold cross-validation with the log marginal likelihood. From our understanding, the log marginal likelihood does not contain individual parameters that are attribute mastery patterns. In the cross-validation procedure, model

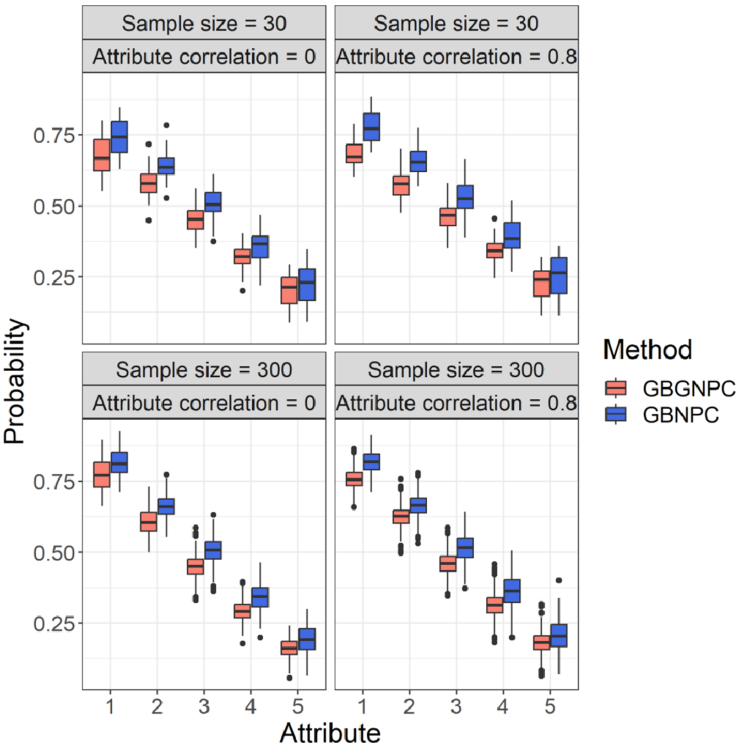

Figure 6. Box plots of attribute mastery probabilities of the DINA data generation with five-attribute Q-matrix conditions.

parameters estimated with a training dataset are plugged in to calculate the log marginal likelihood of the test dataset. In our context, the loss functions in the GB method and nonparametric methods do not contain model parameters and only estimate attribute mastery patterns that relate to individuals. Therefore, the attribute mastery patterns in a training data set are not contained in a test dataset. Exploring the appropriate quantitative evaluation for the GB method in the DCMs is an important direction for future research.

6.1. Data analysis settings

The Examination of the Certificate of Proficiency in English (ECPE) data were selected as an example. ECPE data have been analyzed in various previous studies, such as Templin and Hoffman (2013) and Templin and Bradshaw (2014). The ECPE data contained 2,922 responses for 28 items. Table 4 presents a 28×3 Q-matrix that assumes three attributes: Morphosyntactic (α_1), cohesive (α_2), and lexical rules (α_3). The settings of the GB methods were the same as those used in previous simulations. One difference was that we employed GNPC and NPC estimates as initial values for GBGNPC and GBNPC. The data analysis code can be obtained from the OSF webpage https://osf.io/sau6j/.

6.2. Results

The same correlations as in the simulation study were calculated. Again, the correlations of the three attributes with the GBNPC and GBGNPC methods were all greater than 0.99. This indicated that the MCMC iterations for attribute mastery were stable.

Table 6 lists the frequencies and ratios of the attribute mastery patterns for the four estimation methods. Several differences are observed in Table 6. First, GBGNPC and GBNPC estimated the pattern

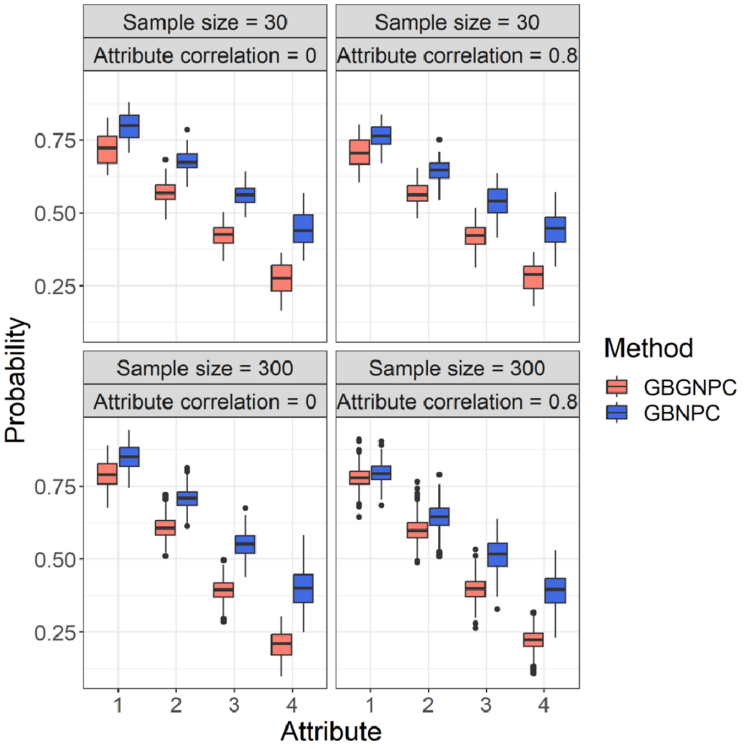

Figure 7. Box plots of attribute mastery probabilities of the general data generation with four-attribute **Q**-matrix conditions.

Table 4. The Q-matrix of ECPE data

Item	Attribute			Item	Attribute		
	Morphosyntactic rules: α_1	Cohesive rules: α_2	Lexical rules: α_3		Morphosyntactic rules: α_1	Cohesive rules: α_2	Lexical rules: α_3
1	1	1	0	15	0	0	1
2	0	1	0	16	1	0	1
3	1	0	1	17	0	1	1
4	0	0	1	18	0	0	1
5	0	0	1	19	0	0	1
6	0	0	1	20	1	0	1
7	1	0	1	21	1	0	1
8	0	1	0	22	0	0	1
9	0	0	1	23	0	1	0
10	1	0	0	24	0	1	0
11	1	0	1	25	1	0	0
12	1	0	1	26	0	0	1
13	1	0	0	27	1	0	0
14	1	0	0	28	0	0	1

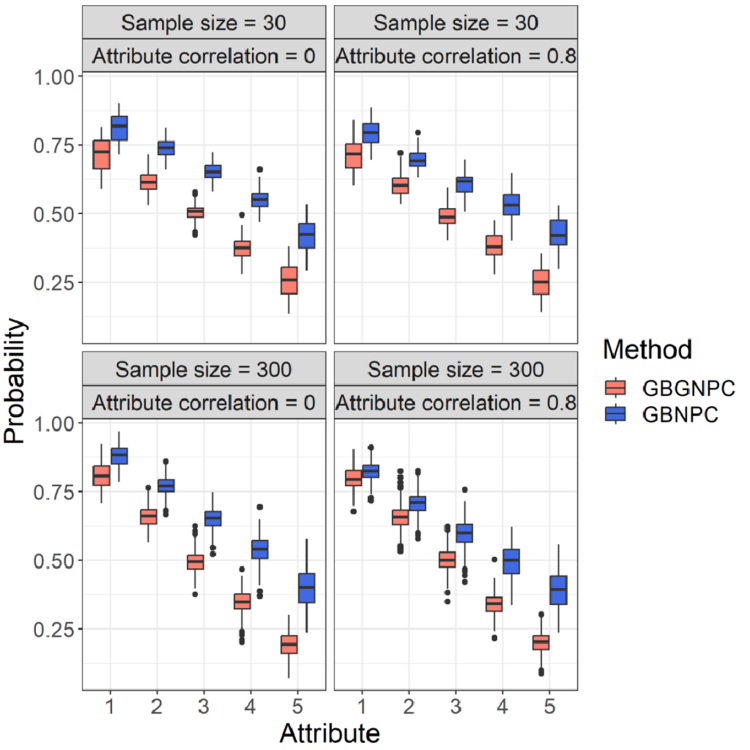

Figure 8. Box plots of attribute mastery probabilities of the general data generation with five-attribute Q-matrix conditions.

Table 5. Means and SDs of posterior attribute mastery probabilities for GBGNPC and GBNPC methods

Estimation method	Attribute	Mean	SD
GBGNPC	Morphosyntactic rules: α_1	.551	.388
	Cohesive rules: α_2	.985	.077
	Lexical rules: α_3	.939	.198
GBNPC	Morphosyntactic rules: α_1	.807	.326
	Cohesive rules: α_2	.978	.103
	Lexical rules: α_3	.949	.187

(001) to be lower than GNPC and NPC estimates. Second, as pattern (011) indicates, the frequency of pattern (011) for GBGNPC was the highest (1203), that for the GNPC was the second (955), that for the NPC was the third (522), and that for the GBNPC was the last (386). The GBGNPC and GBNPC produced lower frequencies than the GNPC and NPC for patterns (100), (101), and (110). The final difference is indicated in pattern (111). GBGNPC and GNPC had relatively smaller numbers than GBNPC and NPC.

Table 5 shows the means and *SD*s of the attribute mastery probabilities for the GBGNPC and GBNPC methods. The attribute mastery probability for the first attribute (Morphosyntactic rules) of GBGNPC was mean = .551 (SD = .388) and that of GBNPC was mean = .807 (SD = .326). The discrepancy was the largest among the three attributes. The attribute mastery probabilities for the second (cohesive rules) and third (lexical rules) attributes using the GBGNPC and GBNPC methods were higher than 0.90 so these attributes tended to be mastered.

Table 6. Frequencies and ratios of the estimated attribute mastery patterns with the four estimation methods

Pattern	GBGNPC		GBNPC		GNPC		NPC	
	Frequency	Ratio	Frequency	Ratio	Frequency	Ratio	Frequency	Ratio
000	24	.008	35	.012	29	.010	44	.015
001	2	.001	3	.001	155	.053	91	.031
010	88	.030	64	.022	88	.030	82	.028
011	1201	.411	384	.131	955	.327	522	.179
100	3	.001	3	.001	38	.013	27	.009
101	0	.000	0	.000	82	.028	87	.030
110	45	.015	36	.012	157	.054	96	.033
111	1559	.534	2397	.820	1418	.485	1973	.675

Note: GBGNPC, generalized Bayesian method with generalized nonparametric loss function; GB-NPC, generalized Bayesian method with nonparametric loss function; GNPC, generalized nonparametric method; NPC, nonparametric method.

Table 7. Contingency table of the estimated attribute mastery patterns by GBGNPC and GNPC

GBGNPC	GNPC							
	000	001	010	011	100	101	110	111
000	18	1	0	0	5	0	0	0
001	0	1	0	0	1	0	0	0
010	7	1	54	0	9	0	17	0
011	4	152	34	924	11	6	41	29
100	0	0	0	0	3	0	0	0
101	0	0	0	0	0	0	0	0
110	0	0	0	0	5	0	40	0
111	0	0	0	31	4	76	59	1389

Note: GBGNPC, generalized Bayesian method with generalized nonparametric loss function; GNPC, generalized nonparametric method.

Table 7 shows the estimated attribute mastery patterns of GBGNPC and GNPC. A large portion of the GBGNPC pattern (011) corresponds to patterns (001), (001), and (010) of the GNPC. Furthermore, patterns (011), (100), (101), and (110) of the GNPC correspond to pattern (111) of the GBGNPC. From these results, the GBGNPC tended to overestimate the number of attributes compared with the GNPC.

Table 8 presents the GBNPC's and NPC's estimated attribute mastery patterns. The results in Table 8 are similar to those of GBGNPC and GNPC. For example, patterns (000), (001), (010), and (010) with the NPC are sometimes estimated as pattern (011) in GBGNPC. Furthermore, patterns (000) to (110) in the NPC were classified as pattern (111) in the GBGNPC. Therefore, the GBNPC overestimates the number of attributes compared with the NPC.

We checked individual differences between the GBGNPC and GNPC methods. Table 9 shows that some individuals indicated the largest pattern discrepancy of attribute mastery between GBGNPC and GNPC methods. The GBGNPC and the GNPC provided $\boldsymbol{\alpha} = (0,1,1)$ and $\boldsymbol{\alpha} = (1,0,0)$, respectively. The response patterns did not indicate systematic tendency but the sum scores of the individuals ranged from 11 to 15, which meant they could answer more than half of the test items. The maximum subscores for attributes one, two, and three were 13, 6, and 18, respectively, so the individuals in Table 9 received half points out of the maximum total subscores. In addition, the sum scores of the individuals ranged from

Table 8. Contingency table of the estimated attribute mastery patterns by GBNPC and NPC

GBNPC	NPC							
	000	001	010	011	100	101	110	111
000	26	5	1	0	3	0	0	0
001	0	3	0	0	0	0	0	0
010	10	0	37	0	8	0	9	0
011	7	54	32	278	3	0	10	0
100	0	0	0	0	3	0	0	0
101	0	0	0	0	0	0	0	0
110	0	0	2	0	3	0	31	0
111	1	29	10	244	7	87	46	1973

Note: GBNPC, generalized Bayesian method with nonparametric loss function; NPC, nonparametric method.

Table 9. Individual differences in estimated patterns for GBGNPC and GNPC methods, response patterns, sum- and subscores, and attribute mastery probabilities

ID	Attribute mastery pattern		Response pattern	Sum-score	Subscore			Attribute mastery probability		
	GBGNPC	GNPC			α_1	α_2	α_3	α_1	α_2	α_3
813	011	100	10001001101001000000011101100	11	5	3	6	.130	.882	.516
1060	011	100	10000111010111000111010001101	14	7	3	9	.418	.956	.864
2378	011	100	11101101111100100100000001011	15	7	3	9	.420	.996	.982
2607	011	100	10001100001010001100101011100	11	5	3	7	.110	.874	.556

Note: GBGNPC, generalized Bayesian method with a generalized nonparametric loss function; GBNPC, generalized Bayesian method with a nonparametric loss function.

Table 10. Generalized posterior of attribute mastery pattern by GBNPC and NPC

Estimation method	ID	Attribute mastery pattern							
		000	100	010	110	001	101	011	111
GBGNPC	813	.102	.016	**.264**	.102	0	0	**.504**	.012
	1060	.008	.022	.020	.086	.006	.008	**.548**	**.302**
	2378	.002	0	.004	.012	0	.002	**.574**	**.406**
	2607	.104	.012	**.246**	.082	.010	0	**.530**	.016
GBNPC	813	**.282**	.012	**.316**	**.250**	.006	0	.106	.028
	1060	.022	.008	.004	.016	0	.004	.066	**.880**
	2378	.006	.002	.004	.016	.002	.002	.072	**.896**
	2607	**.520**	.028	.082	.054	.022	0	**.244**	.050

Note: GBGNPC: generalized Bayesian method with a generalized nonparametric loss function, GBNPC: generalized Bayesian method with a nonparametric loss function.

11 to 15, which is about half the maximum sum-score of 28. Thus, the pattern $(1,0,0)$ might saliently underestimate the latent attributes, making the pattern $(0,1,1)$ possibly more likely. Furthermore, some attribute mastery probabilities were close to the cutoff value 0.5. For example, the mastery probability of the third attribute for the ID 813 individual was 0.516. Additionally, the mastery probabilities of the first attribute for the ID 1060 and 2378 individuals were 0.418 and 0.420. Furthermore, the attribute mastery probability of the third attribute for the ID 2607 individual was 0.556. These values might indicate the mastery of corresponding attributes was not strongly supported. The posterior probabilities for the proposed GBGNPC and GBNPC methods can be used in such cases. However, this is not possible if we use typical nonparametric methods.

The attribute mastery probabilities information provides estimation uncertainty of mastery and nonmastery of an attribute for an individual. It may be better to empathize even if we judge an individual to have mastered an attribute as the mastery might be just slightly over the cutoff value. The nonparametric methods cannot provide such information. Therefore, it may be better to introduce the third category representing the midpoint between mastery and nonmastery in DCM applications.

In addition to each attribute mastery probability, we also added posterior attribute mastery pattern probabilities with the GBGNPC and GBNPC methods in Table 10. The individual's posterior attribute pattern probabilities represent the relative possibilities of attribute mastery patterns. From Table 10, we can see that some attribute patterns showed almost the same posterior probabilities. For example, individual ID 2378 indicated relatively high posterior probabilities 0.574 and 0.406 for (011) and (111) according to the GBGNPC method. A similar tendency was shown by the ID 1060 students with the GBGNPC method. The posterior based on the GBNPC method provided more nuanced estimates for the ID 813 student. This individual had similar posterior probabilities for (000), (010), and (110), with values of $0.282, 0.316$, and 0.250, respectively. It may not be better to provide diagnostic feedback with such unstable posterior probabilities. Previous nonparametric methods cannot provide such uncertainty information, which can be used for careful diagnosis of attribute mastery.

7. Discussions and future directions

This study extends the loss function-based estimation method proposed by Ma, de la Torre, and Xu (2023) to the GB method, which considers estimation uncertainty and prior knowledge. The proposed estimation method can be used for any type of loss function and has great flexibility. This study's contribution is that the proposed method provides a novel approach for estimating the DCMs' parameters. The GB method is flexible because we can select any type of loss function and consider the uncertainty of the parameter estimation. Furthermore, the proposed method relaxes the assumption of the typical Bayesian method, which requires a likelihood function. The theoretical analysis revealed consistent results for the proposed GB method under mild regularity conditions. Additionally, the simulation study revealed that the GB method improved attribute mastery recoveries compared to previous nonparametric methods. The real data example indicated that the proposed GB method with the nonparametric loss function tended to overestimate attribute mastery compared to the nonparametric methods.

The theoretical results not only guarantee the consistency of the MAP estimation results but also give convergence rate results, which is helpful in characterizing the finite sample estimate errors. All these results are new to the literature and provide theoretical justification for using the nonparametric methods and the proposed GB approach. Moreover, the theoretical results in the paper are established for the general loss function under the proposed assumptions. It covers popular loss functions, such as the GNPC and log-likelihood loss functions, which are used in Ma, de la Torre, and Xu (2023).

One interesting future research problem is to establish consistent results for other Bayesian estimators, such as expected a posteriori (EAP). However, this is a more challenging question as it involves deriving the limiting distribution of the Bayesian posterior distribution. Intuitively, given our theoretical results of MAP, EAP would also be consistent, but technically this is not easy to determine and needs the development of new mathematical tools. Moreover, Assumption 2 may be further relaxed to allow

for some latent attribute mastery patterns that do not exist in the population. In particular, if we know which attribute mastery patterns have zero probability, such as in hierarchical DCMs, then our theoretical results would still apply. However, if this information is unknown, while some latent attribute mastery patterns have zero probability, the model itself may have some identifiability issues under the nonparametric DCM setting. This is another interesting topic for future study.

Another future research direction is to explore how to determine the learning rate from data, especially under the $\mathcal{M}$-open setting. Intuitively, the learning rate controls the relative importance between prior information and the loss function. We can set a relatively small value for the learning rate if we have enough prior information about the attribute mastery distribution and use several new items whose nature we do not know. In this case, we put relatively great importance on the prior information rather than the obtained data. However, it may not be realistic to set the learning rate greater than one. Such a high learning rate would amplify the effect of the loss function but might indicate an overreliance on the data. It may not be suitable for the $\mathcal{M}$-open setting that the data-generating process is unknown. Therefore, we need to explore how to determine the learning rate from data.

As mentioned previously, no scholarly agreement exists regarding how to determine the learning rate, which is an important topic for future research especially in the DCM context. In particular, data-driven learning rate determination procedures were studied in Wu and Martin (2023), where several selection methods such as the SafeBayes algorithm based on the cumulative log-loss (Grünwald & van Ommen, 2017), information gain perspective (Holmes & Walker, 2017), modified weighted likelihood bootstrap approach (Lyddon et al., 2019), and the approximate achievement of nominal frequentist coverage probability (Syring & Martin, 2019) were compared. However, all of these methods have different foundations, and we need to explore which one is most appropriate for the DCM context.

Another topic that requires further investigation is model data fit evaluation. From our understanding, the GB method avoids explicit model representation in the framework. Therefore, the model evaluation scheme is not included in the procedure of the GB method. This is also true for the GBGNPC method proposed in this study. Therefore, future research needs to explore what kind of statistics can be used for model data fit. In particular, previously developed methods of model data fit assessment in psychometrics and Bayesian data analysis could be employed in our setting. Following Sinharay (2006), discrepancy measures such as observed score distribution, point biserial correlation, and statistical measures of association among the item pairs could be used for posterior predictive model checking (PPMC). For further details on PPMC methods for Bayesian networks and IRT models, see also Sinharay (2006) and Sinharay (2016). Moreover, PPMC for person fit (Sinharay, 2015) would also provide an important measure to assess the model fit for the attribute mastery patterns at the personal level, which is often of interest in cognitive diagnosis.

As a final note about the choice of estimation methods, it is necessary to consider estimation time. The GB method employs an MCMC procedure, so it has a longer estimation time than that of the nonparametric methods. In our simulation, the estimation times were less than ten seconds, so it is not irritatingly time consuming. However, if we need immediate feedback, the time difference between the two kinds of methods may be crucial. We also need to consider estimation time for the requirement of real data analysis.

Data availability statement. The data analysis code is available in the Open Science Framework page: https://osf.io/sau6j/.

Funding statement. This work was supported by JSPS KAKENHI 20H01720, 21H00936, 22K13810, 23H00985, 23H00065, and 24K00485.

Competing interests. The authors declare no conflicts of interest.

References

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory*. John Wiley & Sons.

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(5), 1103–1130. https://doi.org/10.1111/rssb.12158

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

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*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225–250. https://doi.org/10.1007/s00357-013-9132-9

Chiu, C.-Y., & Köhn, H.-F. (2019). Consistency theory for the general nonparametric classification method. *Psychometrika*, *84*, 830–845. https://doi.org/10.1007/s11336-019-09660-x

Chiu, C.-Y., Köhn, H.-F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika*, *81*, 1069–1092. https://doi.org/10.1007/s11336-016-9534-9

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*, 355–375. https://doi.org/10.1007/s11336-017-9595-4

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476. https://doi.org/10.3102/1076998615595403

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

Grünwald, P., & van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, *12*(4), 1069–1103. https://doi.org/10.1214/17-BA1085

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*, 191–210. https://doi.org/10.1007/s11336-008-9089-5

Holmes, C. C., & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, *104*(2), 497–503. https://doi.org/10.1093/biomet/asx010

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. https://doi.org/10.1177/01466210122032064

Li, H., Hunter, C. V., & Lei, P.-W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, *33*(3), 391–409. https://doi.org/10.1177/0265532215590848

Lyddon, S. P., Holmes, C. C., & Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, *106*(2), 465–478. https://doi.org/10.1093/biomet/asz006

Ma, C., de la Torre, J., & Xu, G. (2023). Bridging parametric and nonparametric methods in cognitive diagnosis. *Psychometrika*, *88*(1), 51–75. https://doi.org/10.1007/s11336-022-09878-2

Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, *88*(1), 175–207. https://doi.org/10.1007/s11336-022-09867-5

Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement*, *45*(2), 95–111. https://doi.org/10.1177/0146621620977681

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*(2), 99–120. https://doi.org/10.2307/1164802

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212. https://doi.org/10.1007/BF02294535

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (*124*, pp. 1–10). https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*(1), 1–33. https://doi.org/10.3102/10769986031001001

Sinharay, S. (2015). Assessment of person fit for mixed-format tests. *Journal of Educational and Behavioral Statistics*, *40*(4), 343–365. https://doi.org/10.3102/1076998615589128

Sinharay, S. (2016). Bayesian model fit and model comparison. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume 2: Statistical tools* (pp. 379–394). Chapman and Hall/CRC. https://doi.org/10.1201/b19166

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Syring, N., & Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika*, *106*(2), 479–486. https://doi.org/10.1093/biomet/asy054

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*(1), 55–73. https://doi.org/10.3102/10769986010001055

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*, 317–339. https://doi.org/10.1007/s11336-013-9362-0
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287
Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. https://doi.org/10.1348/000711007X193957
von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer.
Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, *80*, 85–100. https://doi.org/10.1007/s11336-013-9372-y
Wu, P.-S., & Martin, R. (2023). A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, *18*(1), 105–132. https://doi.org/10.1214/21-BA1302
Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, *45*(2), 675–707. https://doi.org/10.1214/16-AOS1464
Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*(523), 1284–1295. https://doi.org/10.1080/01621459.2017.1340889
Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, *45*(5), 569–597. https://doi.org/10.3102/1076998620911934
Yamaguchi, K., & Templin, J. L. (2022a). Direct estimation of diagnostic classification model attribute mastery profiles via a collapsed Gibbs sampling algorithm. *Psychometrika*, *87*(4), 1390–1421. https://doi.org/10.1007/s11336-022-09857-7
Yamaguchi, K., & Templin, J. L. (2022b). A Gibbs sampling algorithm with monotonicity constraints for diagnostic classification models. *Journal of Classification*, *39*(1), 24–54. https://doi.org/10.1007/s00357-021-09392-7

1. Appendix

Proofs of Theorems 1 and 2

A.1. Preparation for the Proofs

In this appendix, we provide some basic tools and introduce helpful notations for the proofs of Theorems 1 and 2. The proofs are presented in the subsequent sections.

Motivated by the constraint (34), we introduce the concept of a "local" latent class at the item level. Considering item j with q-vector $\boldsymbol{q}_j$, the constraint (34) divides the collection of attribute mastery profiles $\boldsymbol{\alpha}$, which is $\{0,1\}^K$, based on an equivalence relationship where $\boldsymbol{\alpha}_l \sim_j \boldsymbol{\alpha}_{l'}$ is defined by $\boldsymbol{\alpha}_l \circ \boldsymbol{q}_j = \boldsymbol{\alpha}_{l'} \circ \boldsymbol{q}_j$; here, the subscript $\sim_j$ emphasizes that the equivalence relationship is determined by the jth item $\boldsymbol{q}_j$. On this basis, we introduce a function $\xi : \{0,1\}^K \times \{0,1\}^K \to \mathbb{N}$ where $\xi\left(\boldsymbol{q}_j, \boldsymbol{\alpha}_l\right) = \xi\left(\boldsymbol{q}_j, \boldsymbol{\alpha}_{l'}\right)$ is equivalent to $\boldsymbol{\alpha}_l \circ \boldsymbol{q}_j = \boldsymbol{\alpha}_{l'} \circ \boldsymbol{q}_j$. This function assigns numbers to the equivalent classes induced by item j based on specific rules. In the following context, we refer to $\xi\left(\boldsymbol{q}_j^0, \boldsymbol{\alpha}\right)$ as the local latent class of $\boldsymbol{\alpha}$ induced by item j. It is straightforward to verify that the number of local latent classes induced by item j, denoted by $\left|\xi\left(\boldsymbol{q}_j, \{0,1\}^K\right)\right|$, is equal to $L_j = 2^{K_j}$. Here, $K_j = \sum_{k=1}^K q_{jk}^0$ represents the number of latent attributes required for item j; consequently, the range of function ξ satisfies $\xi\left(\boldsymbol{q}_j, \{0,1\}^K\right) = \left[L_j\right] := \left\{1, \ldots, L_j\right\}$. As the local latent classes are identified up to permutations on $\left[L_j\right]$, owing to their categorical nature, the mapping rules between $\xi\left(\boldsymbol{q}_j, \{0,1\}^K\right)$ and $\left[L_j\right]$ need not be completely specified in our discussion.

For brevity, we use the general notation $\mathbf{Z} = \left(z_{ij}\right)$ to denote the collection of local latent classes for all items $j \in [J]$ and subjects $i \in [I]$, where z_{ij} represents $\xi\left(\boldsymbol{q}_j^0, \boldsymbol{\alpha}_i\right)$. Given that $\xi\left(\boldsymbol{q}_j^0, \boldsymbol{\alpha}_l\right) = \xi\left(\boldsymbol{q}_j^0, \boldsymbol{\alpha}_{l'}\right)$ implies $\theta_{j,\boldsymbol{\alpha}_l} = \theta_{j,\boldsymbol{\alpha}_{l'}}$ by the definition of ξ, we express $\theta_{j,\boldsymbol{\alpha}_i}$ as $\theta_{j,z_{ij}}$ to directly incorporate the constraint (34) into the loss function (31). For notational simplicity, we may write $\theta_{j,z_{ij}}$ as θ_{j,z_i}. Consequently, we define:

$$P_{ij} = P\left(X_{ij} = 1\right) = \theta_{j,z_i^0}^0. \tag{A.1}$$

Then, the loss function (31) can be rewritten as:

$$\mathcal{L}\left(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X\right) = \sum_{i=1}^I \left(h\left(\pi_{\boldsymbol{\alpha}_i}\right) + \sum_{j=1}^J \ell\left(X_{ij}, \theta_{j,z_i}\right) \right) + \sum_{j,a} \log f_{ja}\left(\theta_{ja}\right) + \sum_{\boldsymbol{\alpha}} \log g_{\boldsymbol{\alpha}}\left(\pi_{\boldsymbol{\alpha}}\right), \tag{A.2}$$

where $a \in [L_j]$. Observe that $X_{ij}^2 = X_{ij}$, and $\mathbb{E}[X_{ij}] = P_{ij}$, we denote the expectation of the above $\mathcal{L}(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X)$ by $\overline{\mathcal{L}}(\mathcal{A}, \Theta, \boldsymbol{\pi}) := \mathbb{E}[\mathcal{L}(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X)]$.

Notably, $\mathbf{Z} = (z_{ij})$ is determined only by $\mathcal{A}$ because $\mathbf{Q}^0$ is known. In the subsequent context, the quantities determined by the latent attribute profiles $\mathcal{A}$ are sometimes denoted by the superscript $\mathcal{A}$ to emphasize their relationships with $\mathcal{A}$. Considering an arbitrary $\mathcal{A}$, we denote it as:

$$\mathcal{L}(\mathcal{A}) = \inf_{\Theta, \pi} \mathcal{L}\left(\mathcal{A}, \Theta, \boldsymbol{\pi}^{(\mathcal{A})} \mid X\right) = \mathcal{L}\left(\mathcal{A}, \widehat{\Theta}^{(\mathcal{A})}, \widehat{\boldsymbol{\pi}}^{(\mathcal{A})} \mid X\right), \tag{A.3}$$

$$\overline{\mathcal{L}}(\mathcal{A}) = \overline{\mathcal{L}}\left(\mathcal{A}, \overleftarrow{\Theta}^{(\mathcal{A})}, \widehat{\boldsymbol{\pi}}^{(\mathcal{A})}\right), \tag{A.4}$$

where $\left(\widehat{\Theta}^{(\mathcal{A})}, \widehat{\boldsymbol{\pi}}^{(\mathcal{A})}\right) := \arg\min_{\Theta, \pi} \mathcal{L}(\mathcal{A}, \Theta, \boldsymbol{\pi} \mid X)$ and the definition of $\overleftarrow{\Theta}^{(\mathcal{A})}$ is provided later. Notably, $\left(\overleftarrow{\Theta}^{(\mathcal{A})}, \widehat{\boldsymbol{\pi}}^{(\mathcal{A})}\right)$ may not minimize $\overline{\mathcal{L}}(\mathcal{A}, \Theta, \boldsymbol{\pi})$ for a given $\mathcal{A}$. Then, under any realization of $\mathcal{A}$, if the prior distribution of Θ is uniform, the following equations hold for any local latent class $a \in [L_j]$:

$$\widehat{\theta}_{ja}^{(\mathcal{A})} = \frac{\sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{(\mathcal{A})} = a\right\} X_{ij}}{\sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{(\mathcal{A})} = a\right\}}, \overleftarrow{\theta}_{ja}^{(\mathcal{A})} = \frac{\sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{(\mathcal{A})} = a\right\} P_{ij}}{\sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{(\mathcal{A})} = a\right\}}. \tag{A.5}$$

To derive (A.5), note the sum $\sum_{j=1}^{J} \sum_{i=1}^{I} \ell\left(X_{ij}, \theta_{j, z_i}\right)$ equals the sum $\sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i = a} \ell\left(X_{ij}, \theta_{ja}\right)$. When estimating $\widehat{\theta}_{ja}$, we focus on minimizing $\sum_{z_i = a} \ell\left(X_{ij}, \theta_{ja}\right)$. By substituting $\mathbb{E}[X_{ij}] = P_{ij}$ into (A.5), we find that $\mathbb{E}\left[\hat{\theta}_{ja}\right] = \overleftarrow{\theta}_{ja}$ holds for any (j, a). In the following section, we use the second formula in (A.5) to define $\overleftarrow{\Theta}^{(\mathcal{A})}$ given in (A.4). When the prior distribution is not a uniform distribution, $\widehat{\theta}_{ja}$ is obtained from minimizing $\sum_{z_i = a} \ell\left(X_{ij}, \theta_{ja}\right) + \log f_{ja}\left(\theta_{ja}\right)$, where $f_{ja}\left(\theta_{ja}\right)$ is the prior density of θ_{ja}. To avoid ambiguity, we denote $\hat{\theta}_{ja} := \arg\min_{\theta} \sum_{z_i = a} \ell\left(X_{ij}, \theta_{ja}\right) + \log f_{ja}\left(\theta_{ja}\right)$, and $\tilde{\theta}_{ja} := \arg\min_{\theta} \sum_{z_i = a} \ell\left(X_{ij}, \theta_{ja}\right)$. It is clear that $\mathbb{E}\left[\tilde{\theta}_{ja}\right] = \overleftarrow{\theta}_{ja}$.

Before discussing the details of our proof, we provide technical remarks to simplify the discussion. Notably, although we assume that the latent attributes have proportion parameters $\boldsymbol{\pi}^0$, they are still treated as unknown but fixed parameters that need to be estimated. As all the proportion parameters $\pi_{\boldsymbol{\alpha}}^0$ are strictly greater than zero, with the probability converging to 1, $\varepsilon_1 > 0$ exists such that $\min_{\boldsymbol{\alpha}} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}_i^0 = \boldsymbol{\alpha}\right\} \geq I\varepsilon_1$. Subsequently, we use this fact interchangeably with the first condition in Assumption 2.

The second point concerns the compact parameter space specified in Assumptions 2 and 3. Some loss functions may exhibit unusual behavior near the boundary of the parameter space. Although Assumption 2 confines the true item parameters to a compact subset within $(0, 1)$, the estimated item responses can still approach zero or one, making theoretical analysis more difficult. For any pair (j, a), θ_{ja} lies within $[\delta_2, 1 - \delta_2]$. We add a condition to $\hat{\mathcal{A}}$, stating that there exists an $\varepsilon_2 > 0$ such that for each $\boldsymbol{\alpha}$, the sum $\sum_{i=1}^{I} \mathcal{I}\{\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}\}$ is at least $I\varepsilon_2$. With a probability approaching one, this constraint is satisfied by the true latent attribute mastery patterns $\mathcal{A}^0$. With this constraint, for any pair (j, a), the probability that $\left|\tilde{\theta}_{ja} - \theta_{ja}\right|$ exceeds t can be bounded by $2\exp\left(-I\varepsilon_2 t^2\right)$ using Hoeffding's inequality. Thus, the probability that $\max_{j,a} \left|\tilde{\theta}_{ja} - \theta_{ja}\right|$ exceeds t is less than $J2^{K+1} \exp\left(-I\varepsilon_2 t^2\right)$. Based on the scaling condition in Theorem 1, $\max_{j,a} \left|\tilde{\theta}_{ja} - \theta_{ja}\right| = o_p(1)$, implying that with probability converging to 1, all the $\tilde{\theta}_{ja}$ values fall within $[\delta_2/2, 1 - \delta_2/2]$. Based on this result, we assume that in the later content, the estimators $\left(\widehat{\mathcal{A}}, \widehat{\Theta}\right)$ are obtained by minimizing the total loss (A.2), under the constraints that $\min_{\boldsymbol{\alpha}} \sum_i \mathcal{I}\{\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}\} \geq I\varepsilon_2$ and $\widehat{\Theta} \subset [\delta_3, 1 - \delta_3]$, for two small positive constants $\varepsilon_2, \delta_3 > 0$.

The third comment concerns how to quantify the effect of prior density f_{ja} on the corresponding estimator $\widehat{\theta}_{ja}$. Actually, under the smoothness and shape constraints given in Assumption 5 and Assumption 3, the additional term $\log f_{ja}\left(\theta_{ja}\right)$ might cause the estimator $\hat{\theta}_{ja}$ to have a $O_p\left(1/\sqrt{I}\right)$ level drift from the sample average form $\tilde{\theta}_{ja}$ given in (A.6). By considering the Taylor expansion formula, we have:

$$\begin{aligned}
\log f_{ja}\left(\theta_{ja}\right) + \sum_{z_{ij} = a} \ell\left(X_{ij}, \theta_{ja}\right) &= \log f_{ja}\left(\theta_{ja}\right) + \sum_{z_{ij} = a} \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) + \left(\sum_{z_{ij} = a} \partial_{\theta} \ell\left(X_{ij}, \tilde{\theta}_{ja}\right)\right)\left(\theta_{ja} - \tilde{\theta}_{ja}\right) \\
&\quad + \frac{1}{2} \partial_{\theta^2} \ell\left(X_{ij}, \dot{\theta}_{ja}\right)\left(\theta_{ja} - \tilde{\theta}_{ja}\right)^2, \\
&= \log f_{ja}\left(\theta_{ja}\right) + \sum_{z_{ij} = a} \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) + \frac{1}{2} \sum_{z_{ij} = a} \partial_{\theta^2} \ell\left(X_{ij}, \dot{\theta}_{ja}\right)\left(\theta_{ja} - \tilde{\theta}_{ja}\right)^2,
\end{aligned}$$

where $\dot{\theta}_{ja}$ is between $\tilde{\theta}_{ja}$ and θ_{ja} according to the mean value theorem; the second equality holds owing to Assumption 3. According to the above equation, we can find that $\widehat{\theta}_{ja} = \arg\min_{\theta \in [\delta_3, 1 - \delta_3]} \log f_{ja}(\theta) + \sum_{z_{ij} = a} \partial_{\theta^2} \ell\left(X_{ij}, \dot{\theta}_{ja}\right)\left(\theta - \tilde{\theta}_{ja}\right)^2 / 2$. Note

that if we take $\theta = \tilde{\theta}_{ja}$, $\log f_{ja}(\theta) + \sum_{z_{ij}=a} \partial_{\theta^2} \ell\left(X_{ij}, \dot{\theta}_{ja}\right)\left(\theta - \tilde{\theta}_{ja}\right)^2/2 = \log f_{ja}\left(\tilde{\theta}_{ja}\right)$. Based on Assumption 5, there exists a constant $C > 0$ such that $\left|\log f_{ja}\left(\tilde{\theta}_{ja}\right)\right| \le \sup_{\theta \in [\delta_3, 1-\delta_3]} \left|\log f_{ja}(\theta)\right| < C$, implying the following:

$$
\begin{aligned}
2C &\ge \frac{1}{2} \sum_{z_{ij}=a} \partial_{\theta^2} \ell\left(X_{ij}, \dot{\theta}_{ja}\right)\left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \\
&\ge \frac{b_L}{2} \sum_{z_{ij}=a} \left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 = \frac{b_L i_{ja}}{2}\left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2,
\end{aligned}
$$

where $i_{ja} := \sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{(\mathcal{A})} = a\right\}$. Thus, a constant $\check{C} > 0$ exists such that for any pair (j,a), we have:

$$
\left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \le \check{C} i_{ja}^{-1}.
$$

This inequality will be used several times afterwards. In the theoretical analysis of the estimators above, uniform bounds related to the quantities of element-wise loss $\ell(\cdot,\cdot)$ and prior densities f_{ja} are frequently used. The existence of these uniform bounds requires restricting the parameter space of the item response parameters to a compact subspace. Therefore, discussing the compact parameter subspace of the item parameters is necessary.

A.2. Outline of the first half of the proof

Step 1: Express the upper bound of $|\overline{\mathcal{L}}(\mathcal{A}) - \mathcal{L}(\mathcal{A})|$ in terms of $(b/2)\cdot\left(\sum_{j=1}^{J}\sum_{a=1}^{L_j} i_{ja}\left(\tilde{\theta}_{ja} - \overleftarrow{\theta}_{ja}\right)^2\right) + |\mathbb{E}[Y] - Y| + O_p(J)$, where $Y := \sum_i \sum_j \ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}^{(\mathcal{A})}\right)$ depending on Y and $\overleftarrow{\Theta}^{(\mathcal{A})}$ under $\mathcal{A}$, b is the upper bound of the second-order derivative of the $\ell(\cdot,\cdot)$.

Step 2: Bound $\sum_j \sum_i i_{ja}\left(\hat{\theta}_{ja} - \overleftarrow{\theta}_{ja}\right)^2$ and $|Y - \mathbb{E}[Y]|$ separately to obtain a uniform convergence rate $\sup_{\mathcal{A}} |\overline{\mathcal{L}}(\mathcal{A}) - \mathcal{L}(\mathcal{A})| = o_p(\delta_{IJ})$

Step 3: Based on the definition of $\hat{\mathcal{A}}$, it follows that $0 \le \overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right) - \overline{\mathcal{L}}\left(\mathcal{A}^0\right) \le 2\sup_{\mathcal{A}} |\overline{\mathcal{L}}(\mathcal{A}) - \mathcal{L}(\mathcal{A})| = o_p(\delta_{IJ})$, which controls the deviation $\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right) - \overline{\mathcal{L}}\left(\mathcal{A}^0\right)$.

In some classical statistical inference contexts, consistent results for the parameters of interest are typically established through the uniform convergence of random functions associated with these parameters. For instance, if $\sup_{\boldsymbol{\theta}\in\Theta} |\widehat{\ell}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \xrightarrow{P} 0$, and if we further assume that ℓ has a unique minimum $\tilde{\boldsymbol{\theta}}$ on Θ, $\arg\min_{\Theta} \widehat{\ell}(\boldsymbol{\theta}) =: \widehat{\boldsymbol{\theta}} \xrightarrow{P} \tilde{\boldsymbol{\theta}}$ under some regularity conditions. The regular conditions may vary across settings. Considering $\mathcal{A}$ as the parameter to be estimated, the primary aim of the first three steps is to demonstrate that $\mathcal{A}$ minimizes the expected loss and establishes a uniform convergence result for its random loss function of $\mathcal{A}$.

A.3. Outline of the second half of the proof

Step 4: Define $I_{a,b}^{j} = \sum_{i=1}^{I} \mathcal{I}\left\{z_{ij}^{0} = a\right\}\mathcal{I}\left\{\widehat{z}_{ij} = b\right\}, a, b \in [L_j]$ to represent the samples with the wrong local latent class assignments. Derive some upper bounds for the quantities based on $I_{a,b}^{j}$ using $\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right) - \overline{\mathcal{L}}\left(\mathcal{A}^0\right)$ with the help of the identification assumptions.

Step 5: Bound the $\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{\boldsymbol{\alpha}}_i \ne \boldsymbol{\alpha}^0\right\}$ using the quantities based on $I_{a,b}^{j}$ with the help of the discrete structure of the Q-matrix, then obtain the desired classification error rate.

Assumptions 1–5 are the regularity conditions for achieving clustering consistency based on the uniform convergence results established in the first half of the proof. We have provided further details regarding the assumptions in later proofs.

A.4. First Half of the Proof of Theorem 1

Step 1. The idea of decomposing $\overline{\mathcal{L}}(\mathcal{A}) - \mathcal{L}(\mathcal{A})$ is to consider:

$$
\begin{aligned}
\ell\left(X_{ij}, \hat{\theta}_{j,z_i}\right) - \mathbb{E}\left[\ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right)\right] &= \left(\ell\left(X_{ij}, \widehat{\theta}_{j,z_i}\right) - \ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right)\right) \\
&+ \left(\ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right) - \mathbb{E}\left[\ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right)\right]\right).
\end{aligned}
$$

The variability in the first term of the right-hand side mainly emerges from the fluctuation in $\left|\widehat{\theta}_{ja} - \overleftarrow{\theta}_{ja}\right|$, while the randomness in the second term is attributable to the stochastic nature of X_{ij}.

Lemma 1. *Let* $\left(X_{ij}; 1 \le i \le I, 1 \le j \le J\right)$ *denote independent Bernoulli trials with parameters* $\left(P_{ij}; 1 \le i \le I, 1 \le j \le J\right)$. *In a general latent class model, given arbitrary latent attribute mastery patterns* $\mathcal{A}$,

$$\left|\overline{\mathcal{L}}(\mathcal{A}) - \mathcal{L}(\mathcal{A})\right| \le \frac{b}{2} \cdot \left(\sum_{j=1}^{J} \sum_{a=1}^{L_j} i_{ja} \left(\widehat{\theta}_{ja} - \overline{\theta}_{ja}\right)^2 \right) + \left|Y - \mathbb{E}(Y)\right| + O_{p(J)}, \tag{A.6}$$

where $Y = \sum_{j=1}^{J} \sum_{i=1}^{I} \ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right)$ *is a random variable depending on* $\mathcal{A}$ *and* L_j *denotes the number of the distinct local latent classes induced by* $\boldsymbol{q}_j$ *for item* j.

Proof. By noting the decomposition that we mentioned at the beginning of Step 1, $\left| Y - \mathbb{E}[Y] \right|$ is easy to check. It is sufficient for us to prove that

$$0 \le \sum_i \sum_j \left(\ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right) - \ell\left(X_{ij}, \widehat{\theta}_{j,z_i}\right) \right) \le \frac{b}{2} \cdot \left(\sum_{j=1}^{J} \sum_{a=1}^{L_j} i_{ja} \left(\tilde{\theta}_{ja} - \overleftarrow{\theta}_{ja}\right)^2 \right) + O_p(J).$$

□

The first inequality is clear by the definition of $\tilde{\theta}_{j,z_i}$ (minimizing the loss). For the second part, using the mean value theorem for second-order derivatives, we obtain

$$\begin{aligned}
&\sum_i \sum_j \left(\ell\left(X_{ij}, \overleftarrow{\theta}_{j,z_i}\right) - \ell\left(X_{ij}, \tilde{\theta}_{j,z_i}\right) \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\ell\left(X_{ij}, \overleftarrow{\theta}_{ja}\right) - \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\partial_\theta \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) \left(\overleftarrow{\theta}_{ja} - \tilde{\theta}_{ja}\right) + \frac{1}{2} \partial_{\theta^2}\left(X_{ij}, \dot{\theta}_{ja}\right) \left(\tilde{\theta}_{ja} - \widehat{\theta}_{ja}\right)^2 \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\frac{1}{2} \partial_{\theta^2}\left(X_{ij}, \dot{\theta}_{ja}\right) \left(\overleftarrow{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \right) \\
&\le \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\frac{b_U}{2} \left(\overleftarrow{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \right) = \frac{b_U}{2} \left(\sum_{j=1}^{J} \sum_{a=1}^{L_j} i_{ja} (\tilde{\theta}_{ja} - \tilde{\theta}_{ja})^2 \right),
\end{aligned} \tag{A.7}$$

where $\dot{\theta}_{ja}$ is between $\tilde{\theta}_{ja}$ and $\overleftarrow{\theta}_{ja}$ according to the mean value theorem. The third equality holds true since by Assumption 3, we have

$$\sum_{z_i=a} \partial_\theta \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) = 0.$$

Similarly, using $\left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \le \tilde{C} i_{ja}^{-1}$, we have:

$$\begin{aligned}
&\sum_i \sum_j \left(\ell\left(X_{ij}, \widehat{\theta}_{j,z_i}\right) - \ell\left(X_{ij}, \tilde{\theta}_{j,z_i}\right) \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\ell\left(X_{ij}, \widehat{\theta}_{ja}\right) - \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\partial_\theta \ell\left(X_{ij}, \tilde{\theta}_{ja}\right) \left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right) + \frac{1}{2} \partial_{\theta^2}\left(X_{ij}, \dot{\theta}_{ja}\right) \left(\tilde{\theta}_{ja} - \widehat{\theta}_{ja}\right)^2 \right) \\
&= \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\frac{1}{2} \partial_{\theta^2}\left(X_{ij}, \dot{\theta}_{ja}\right) \left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \right) \\
&\le \sum_{j=1}^{J} \sum_{a=1}^{L_j} \sum_{z_i=a} \left(\frac{b_U}{2} \left(\widehat{\theta}_{ja} - \tilde{\theta}_{ja}\right)^2 \right) = \sum_{j=1}^{J} \sum_{a=1}^{L_j} \frac{b_U \tilde{C}}{2} \le \left(b_U \tilde{C} 2^K\right) J,
\end{aligned}$$

which concludes the proof of this lemma. □

Lemma 2. *The following event happens with a probability of at least* $1-\delta$,

$$\max_{\mathcal{A}}\left\{\sum_{j=1}^{J}\sum_{a=1}^{L_j} i_{ja}\left(\tilde{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\right\} < \frac{1}{2}\left(I\log 2^K + J2^K\log\left(\frac{I}{2^K}+1\right)-\log\delta\right).$$

Proof. Under any realization of $\mathcal{A}$, each $\tilde{\theta}_{ja}$ is an average of i_{ja} independent Bernoulli random variables $x_{1j},\ldots,x_{Ij}$ with mean $\overleftarrow{\theta}_{ja}$. By applying the Hoeffding inequality, we have

$$P\left(\tilde{\theta}_{ja}\geq\overleftarrow{\theta}_{ja}+t\right)\leq\exp\left(-2i_{ja}t^2\right), P\left(\tilde{\theta}_{ja}\leq\overleftarrow{\theta}_{ja}-t\right)\leq\exp\left(-2i_{ja}t^2\right). \tag{A.8}$$

□

Notably, considering a fixed $\mathcal{A}$, each $\tilde{\theta}_{ja}$ can take values only in the finite set $\{0,1/i_{ja},2/i_{ja},\ldots,1\}$ of cardinality $i_{ja}+1$. We denote this range of $\tilde{\theta}_{ja}$ by $\tilde{\Theta}^{ja}$ and the range of the matrix $\tilde{\Theta}=\left(\tilde{\theta}_{ja}\right)$ by $\tilde{\Theta}$. Subsequently, $P\left(\tilde{\theta}_{ja}=v\right)\leq\exp\left(-2i_{ja}\left(v-\overleftarrow{\theta}_{ja}\right)^2\right)$ for any $v\in\tilde{\Theta}^{j,a}$. As each of the $J\times 2^K$ entries in $\tilde{\Theta}$, $\tilde{\theta}_{ja}$ can independently take on $i_{ja}+1$ different values, there is $|\tilde{\Theta}|=\prod_j\prod_{a=1}^{L_j}\left(i_{ja}+1\right)$ with constraint $\sum_{a=1}^{L_j} i_{ja}=I$. As $L_j=2^{K_j}\leq 2^K$, we have $\prod_{a=1}^{L_j}\left(i_{ja}+1\right)\leq\left(1+I/2^K\right)^{2^K}$. Denote $\tilde{\Theta}_\varepsilon=\left\{\dot{\Theta}\in\tilde{\Theta}:\sum_j\sum_a i_{ja}\left(\dot{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\geq\varepsilon\right\}$, $\dot{\Theta}_\varepsilon\subseteq\tilde{\Theta}$, and

$$\begin{aligned}
P\left(\sum_j\sum_{a=1}^{L_j} i_{ja}\left(\tilde{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\geq\varepsilon\right) &= \sum_{\dot{\Theta}\in\tilde{\Theta}_\varepsilon} P\left(\tilde{\Theta}=\dot{\Theta}\right)\\
&\leq \sum_{\dot{\Theta}\in\tilde{\Theta}_\varepsilon}\prod_j\prod_a\exp\left(-2i_{ja}\left(\dot{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\right)\\
&= \sum_{\dot{\Theta}\in\tilde{\Theta}_\varepsilon}\exp\left(-2i_{ja}\sum_j\sum_a\left(\dot{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\right)\\
&\leq \sum_{\dot{\Theta}\in\tilde{\Theta}_\varepsilon}\exp\left(-2\varepsilon\right)\leq\left|\tilde{\Theta}\right|e^{-2\varepsilon}\leq\left(\frac{I}{2^K}+1\right)^{J2^K}e^{-2\varepsilon}.
\end{aligned} \tag{A.9}$$

The above result holds for any fixed $\mathcal{A}$ when we apply a union bound over all the $\left(2^K\right)^I$ possible assignments of $\mathcal{A}$ to obtain

$$P\left(\max_{\mathcal{A}}\left\{\sum_j\sum_a i_{ja}\left(\tilde{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\right\}\geq\varepsilon\right)\leq 2^{KI}\left(\frac{I}{2^K}+1\right)^{J2^K}e^{-2\varepsilon}. \tag{A.10}$$

Take $\delta=2^{KI}\left(\frac{I}{2^K}+1\right)^{J2^K}e^{-2\varepsilon}$; then, $2\varepsilon=I\log 2^K+J2^K\log\left(1+I/2^K\right)-\log\delta$. This concludes the proof of lemma 2. □

Lemma 3. *Define the random variable* $Y=\sum_i\sum_j\ell\left(X_{ij},\overleftarrow{\theta}^{(\mathcal{A})}_{j,z_i}\right)$, *and denote* $Y_{ij}=\ell\left(X_{ij},\overleftarrow{\theta}_{j,z_i}\right)$. *Note that* $\overleftarrow{\theta}_{ja}\in[\delta_2,1-\delta_2]$ *and* $\ell(\cdot,\cdot)$ *are continuous on* θ *in* $(0,1)$. *Since continuous functions on the compact set are bounded, a constant* $U>0$ *exists such that* $\left|\ell\left(X_{ij},\overleftarrow{\theta}_{j,z_i}\right)\right|\leq U,\forall(i,j)$. *By applying Hoeffding's inequality to bound* $|Y-\mathbb{E}[Y]|$ *for any realization of* $\mathcal{A}$, *we have:*

$$P\left(|Y-\mathbb{E}[Y]|\geq\varepsilon\right)\leq 2\exp\left\{-\frac{\varepsilon^2}{(4U^2)IJ}\right\}. \tag{A.11}$$

With the help of Lemma 2 and Lemma 3, subsequently, we prove the following proposition:

Proposition 1. *Under the following scaling for some small positive constant* $c>0$,

$$\sqrt{J}=O\left(I^{1-c}\right)$$

we have $max_{\mathcal{A}}\ |\overline{\mathcal{L}}(\mathcal{A})-\mathcal{L}(\mathcal{A})|=o_p\left(\delta_{IJ}\right)$ where $\delta_{IJ}=I\sqrt{J}(\log J)^{\bar{\varepsilon}}$ for a small positive $\bar{\varepsilon}>0$.

Proof. First, note that under the given scaling condition, $J=o\left(I\sqrt{J}\right)$. Combining the results of Lemma 2 and Lemma 3, since there are $\left(2^K\right)^I$ possible assignments of $\mathcal{A}$, we apply the union bound to obtain

$$
\begin{aligned}
P\left(\max_{\mathcal{A}}\left|\overline{\mathcal{L}}(\mathcal{A})-\mathcal{L}(\mathcal{A})\right|\geq 3\varepsilon\delta_{IJ}\right) &\leq \left(2^K\right)^I P\left[\left\{\sum_j\sum_a i_{ja}\left(\tilde{\theta}_{ja}-\overleftarrow{\theta}_{ja}\right)^2\geq\varepsilon\delta_{IJ}\right\}\cup\left\{\left|Y-\mathbb{E}[Y]\right|\geq\varepsilon\delta_{IJ}\right\}\right] \\
&\quad + P\left(\max_{\mathcal{A}}\sum_i\sum_j\left(\ell\left(X_{ij},\widehat{\theta}_{j,z_i}\right)-\ell\left(X_{ij},\tilde{\theta}_{j,z_i}\right)\right)\geq J(\log)^{\tilde{\varepsilon}}\right) \\
&\leq \exp\left(I\log\left(2^K\right)+J2^K\log\left(\frac{I}{2^K}+1\right)-2\varepsilon\delta_{IJ}\right) \\
&\quad + 2\exp\left(I\log\left(2^K\right)-\frac{\varepsilon^2\delta_{IJ}^2}{4U^2IJ}\right) \\
&\quad + P\left(\max_{\mathcal{A}}\sum_i\sum_j\left(\ell\left(X_{ij},\widehat{\theta}_{j,z_i}\right)-\ell\left(X_{ij},\tilde{\theta}_{j,z_i}\right)\right)\geq J(\log)^{\tilde{\varepsilon}}\right). \qquad \text{(A.12)}
\end{aligned}
$$

□

For the third term, note that the following inequality holds for any given $\mathcal{A}$:

$$
\sum_i\sum_j\left(\ell\left(X_{ij},\hat{\theta}_{j,z_i}\right)-\ell\left(X_{ij},\tilde{\theta}_{j,z_i}\right)\right)\leq\left(b_U\tilde{C}2^K\right)J
$$

For the second term on the right-hand side of the aforementioned display to converge to zero, we set $\delta_{IJ}=I\sqrt{J}(\log J)^{\tilde{\varepsilon}}$ for a small positive constant $\tilde{\varepsilon}$. Moreover, under this δ_{IJ}, for the first term to converge to zero as I,J increase, the scaling $\sqrt{J}=O\left(I^{1-c}\right)$ given in the theorem results in $P\left(\max_{\mathcal{A}}\left|\overline{\mathcal{L}}(\mathcal{A})-\mathcal{L}(\mathcal{A})\right|\geq\varepsilon\delta_{IJ}\right)=o(1)$, which implies the result in Proposition 1.

Step 3. (A.5) implies that $\overleftarrow{\theta}_{j,z_i}^{(\mathcal{A}^0)}=P_{ij}$, which means that if we plug into the true latent attribute mastery pattern $\mathcal{A}^0$, the estimators will be the corresponding true parameters. According to this property, the following lemma indicates that $\mathcal{A}^0$ minimizes expected loss.

Lemma 4. *By Assumption 4,* $\mathbb{E}\left[\ell\left(X_{ij},\overleftarrow{\theta}_{j,z_i}\right)\right]-\mathbb{E}\left[\ell\left(X_{ij},\theta^0_{j,z_i^0}\right)\right]\geq c\left(\overleftarrow{\theta}_{j,z_i}-\theta^0_{j,z_i^0}\right)^\eta$ *for some* $\eta\geq 2, c>0$, *then we have*

$$
\overline{\mathcal{L}}(\mathcal{A})-\overline{\mathcal{L}}\left(\mathcal{A}^0\right)\geq c\cdot\left(\sum_i\sum_j\left(P_{ij}-\overleftarrow{\theta}_{j,z_i}\right)^\eta\right)\geq 0. \qquad \text{(A.13)}
$$

Notably, while Lemma 4 holds for any $\mathcal{A}$, it also holds for the estimator $\widehat{\mathcal{A}}$, then

$$
0\leq\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right)-\overline{\mathcal{L}}\left(\mathcal{A}^0\right)=\left[\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right)-\mathcal{L}\left(\widehat{\mathcal{A}}\right)\right]+\left[\mathcal{L}\left(\widehat{\mathcal{A}}\right)-\mathcal{L}\left(\mathcal{A}^0\right)\right]+\left[\mathcal{L}\left(\mathcal{A}^0\right)-\overline{\mathcal{L}}\left(\mathcal{A}^0\right)\right]. \qquad \text{(A.14)}
$$

As $\widehat{\mathcal{A}}=\arg\min_{\mathcal{A}}\mathcal{L}(\mathcal{A})$, we have $\mathcal{L}\left(\widehat{\mathcal{A}}\right)-\mathcal{L}\left(\mathcal{A}^0\right)\leq 0$. Substituting this into A.14, we can derive that

$$
0\leq\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right)-\overline{\mathcal{L}}\left(\mathcal{A}^0\right)\leq 2\sup_{\mathcal{A}}\left|\overline{\mathcal{L}}(\mathcal{A})-\mathcal{L}(\mathcal{A})\right|=o_p\left(\delta_{IJ}\right)
$$

A.5. Second Half of the Proof of Theorem 1

By applying Hölder's inequality, we have

$$
(IJ)^{1-\frac{\eta}{2}}\left(\sum_i\sum_j\left(P_{ij}-\overleftarrow{\theta}_{j,z_i}\right)^2\right)^{\frac{\eta}{2}}\leq\sum_i\sum_j\left(P_{ij}-\overleftarrow{\theta}_{j,z_i}\right)^\eta=o_p\left(\delta_{IJ}\right).
$$

By letting $(IJ)^{1-\eta/2}(S)^{\eta/2}=\delta_{IJ}$, we can check that $\sum_i\sum_j\left(P_{ij}-\overleftarrow{\theta}_{j,\widetilde{z}_i}\right)^2=o_p(S)$ where $S:=I(J)^{1-1/\eta}(\log J)^{2\tilde{\varepsilon}/\eta}$. In the following, we derive a lower bound for $\sum_i\sum_j\left(P_{ij}-\overleftarrow{\theta}_{j,z_i}\right)^2$ because it is easier to work with than $\left(P_{ij}-\overleftarrow{\theta}_{j,z_i}\right)^\eta$.

Step 4. Motivated by Assumption 2, we define $\mathcal{J}:=\left\{j\in[J];\exists k\in[K]\text{ s.t. }\boldsymbol{q}_j^0=\boldsymbol{e}_k\right\}$, which represents the set of all items j that depend on only one latent attribute. Notably, $\forall j\in\mathcal{J}$, $\left|\left\{\boldsymbol{\alpha}\circ\boldsymbol{q}_j^0;\boldsymbol{\alpha}\in\{0,1\}^K\right\}\right|=2$, as $\boldsymbol{q}_j$ only contains one required latent attribute, then $\xi\left(\boldsymbol{q}_j^0,\boldsymbol{\alpha}\right)\in\{1,2\}$ for all $j\in\mathcal{J}$. Without loss of generality, we assume that if $\boldsymbol{\alpha}\circ\boldsymbol{q}_j^0\neq\mathbf{0}$, then let $\xi\left(\boldsymbol{q}_j^0,\boldsymbol{\alpha}\right)=2$, otherwise, let $\xi\left(\boldsymbol{q}_j^0,\boldsymbol{\alpha}\right)=1$. Further, we assume that $\theta^0_{j,2}>\theta^0_{j,1},\forall j\in\mathcal{J}$, which aligns with the concept that subjects possessing

the required latent attribute tend to perform better. For any $j \in \mathcal{J}$, define

$$I^j_{a,b} := \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} = a\right\} \mathcal{I}\left\{\widehat{z}_{ij} = b\right\}, (a,b) \in \{1,2\}^2. \tag{A.15}$$

Note $P_{ij} = \mathcal{I}\left\{z^0_{ij} = 2\right\} \theta^0_{j,2} + \mathcal{I}\left\{z^0_{ij} = 1\right\} \theta^0_{j,1}$ and $I^j_{2,2} + I^j_{1,2} = \sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 2\right\}$, $I^j_{2,1} + I^j_{1,1} = \sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 1\right\}$. By using (A.5), there are

$$\begin{aligned} \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2} &= \frac{\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 2\right\} P_{ij}}{\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 2\right\}} = \frac{\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 2\right\} \left(\mathcal{I}\left\{z^0_{ij} = 2\right\} \theta^0_{j,2} + \mathcal{I}\left\{z^0_{ij} = 1\right\} \theta^0_{j,1}\right)}{\sum_{i=1}^{I} \mathcal{I}\left\{\widehat{z}_{ij} = 2\right\}} \\ &= \frac{I^j_{2,2}\theta^0_{j,2} + I^j_{1,2}\theta^0_{j,1}}{I^j_{2,2} + I^j_{1,2}}; \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1} = \frac{I^j_{2,1}\theta^0_{j,2} + I^j_{1,1}\theta^0_{j,1}}{I^j_{2,1} + I^j_{1,1}}. \end{aligned} \tag{A.16}$$

Under $\widehat{\mathcal{A}}$, we impose a natural constraint $\overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2} > \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}, \forall j \in \mathcal{J}$ on $\widehat{\mathcal{A}}$ for identifiability purpose. This constraint does not change the previous results as $\theta^0_{j,2} > \theta^0_{j,1}$ allows $\mathcal{L}\left(\widehat{\mathcal{A}}\right) - \mathcal{L}\left(\mathcal{A}^0\right) \le 0$ in (A.14) still holds; thus, $\overline{\mathcal{L}}\left(\widehat{\mathcal{A}}\right) - \overline{\mathcal{L}}\left(\mathcal{A}^0\right) = o_p\left(\delta_{IJ}\right)$ still holds under this constraint. Combining $\overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2} > \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}$ and $\theta^0_{j,2} > \theta^0_{j,1}$, there is

$$\overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2} > \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1} \Longleftrightarrow \left(I^j_{2,2}I^j_{1,1} - I^j_{1,0}I^j_{0,1}\right)\theta^0_{j,2} > \left(I^j_{2,2}I^j_{1,1} - I^j_{1,0}I^j_{0,1}\right)\theta^0_{j,1} \Longleftrightarrow I^j_{2,2}I^j_{1,1} > I^j_{2,1}I^j_{1,2}. \tag{A.17}$$

From (A.15), we can obtain

$$\begin{aligned} \left|\theta^0_{j,1} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}\right| &= \frac{I^j_{2,1}\left(\theta^0_{j,2} - \theta^0_{j,1}\right)}{I^j_{2,1} + I^j_{1,1}}, \left|\theta^0_{j,2} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2}\right| = \frac{I^j_{1,2}\left(\theta^0_{j,2} - \theta^0_{j,1}\right)}{I^j_{2,2} + I^j_{1,2}}, \\ \left|\theta^0_{j,2} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}\right| &= \frac{I^j_{1,1}\left(\theta^0_{j,2} - \theta^0_{j,1}\right)}{I^j_{2,1} + I^j_{1,1}}, \left|\theta^0_{j,1} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2}\right| = \frac{I^j_{2,2}\left(\theta^0_{j,2} - \theta^0_{j,1}\right)}{I^j_{2,2} + I^j_{1,2}}. \end{aligned}$$

Therefore,

$$\begin{aligned} &\sum_{j=1}^{J}\sum_{i=1}^{I}\left(P_{ij} - \overleftarrow{\theta}_{j,\widehat{z}_i}\right)^2 \ge \sum_{j\in\mathcal{J}}\sum_{i=1}^{I}\left(P_{ij} - \overleftarrow{\theta}_{j,\widehat{z}_i}\right)^2 \\ &= \sum_{j\in\mathcal{J}}\left(I^j_{1,1}\left(\theta^0_{j,1} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}\right)^2 + I^j_{2,1}\left(\theta^0_{j,2} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,1}\right)^2 + I^j_{1,2}\left(\theta^0_{j,1} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2}\right)^2 + I^j_{2,2}\left(\theta^0_{j,2} - \overleftarrow{\theta}^{(\widehat{\mathcal{A}})}_{j,2}\right)^2\right) \\ &= \sum_{j\in\mathcal{J}}\left(\frac{I^j_{1,1}(I_{2,1})^2 + I^j_{2,1}(I_{1,1})^2}{\left(I^j_{2,1} + I^j_{1,1}\right)^2} + \frac{I^j_{1,2}(I_{2,2})^2 + I^j_{2,2}(I_{1,2})^2}{\left(I^j_{2,2} + I^j_{1,2}\right)^2}\right)\left(\theta^0_{j,2} - \theta^0_{j,1}\right)^2 \\ &= \sum_{j\in\mathcal{J}}\left(\frac{I^j_{2,1}I^j_{1,1}}{I^j_{2,1} + I^j_{1,1}} + \frac{I^j_{2,2}I^j_{1,2}}{I^j_{2,2} + I^j_{1,2}}\right)\left(\theta^0_{j,2} - \theta^0_{j,1}\right)^2 \\ &\ge \delta\sum_{j\in\mathcal{J}}\left(\frac{I^j_{2,1}I^j_{1,1}}{I^j_{2,1} + I^j_{1,1}} + \frac{I^j_{2,2}I^j_{1,2}}{I^j_{2,2} + I^j_{1,2}}\right) \\ &\ge \frac{1}{2}\delta\sum_{j\in\mathcal{J}}\left(\min\left\{I^j_{2,1}, I^j_{1,1}\right\} + \min\left\{I^j_{2,2}, I^j_{1,2}\right\}\right). \end{aligned} \tag{A.18}$$

The second inequality holds since by Assumption 1, $\left(\theta^0_{j,2} - \theta^0_{j,1}\right)^2 \ge \delta$. One ideal scenario is that for most $j \in \mathcal{J}$, $\min\left\{I^j_{2,1}, I^j_{1,1}\right\} + \min\left\{I^j_{2,2}, I^j_{1,2}\right\} = I^j_{2,1} + I^j_{1,2} = \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} \ne \widehat{z}_{ij}\right\}$; thus the misclassification error for the local latent classes could be bounded relatively tight. The following result confirms this intuition.

Lemma 5. *Define the following random set depending on the estimated latent attribute mastery patterns $\hat{A}$ under constraint $\overleftarrow{\theta}^{(\mathcal{A})}_{j,2} > \overleftarrow{\theta}^{(\mathcal{A})}_{j,1}, \forall j \in \mathcal{J}$:*

$$\mathcal{J}_0 = \left\{j \in \mathcal{J}; I^j_{2,1} < I^j_{1,1}, I^j_{1,2} < I^j_{2,2}\right\};$$
$$\mathcal{J}_1 = \left\{j \in \mathcal{J}; I^j_{2,1} < I^j_{1,1}, I^j_{1,2} > I^j_{2,2}\right\};$$
$$\mathcal{J}_2 = \left\{j \in \mathcal{J}; I^j_{2,1} > I^j_{1,1}, I^j_{1,2} < I^j_{2,2}\right\},$$

then under Assumption 1 and Assumption 2, there are $|\mathcal{J}_1| = o_p(S/I)$, $|\mathcal{J}_2| = o_p(S/I)$.

Proof. If $j \in \mathcal{J}_1$, $\min\left\{I^j_{2,1}, I^j_{1,1}\right\} + \min\left\{I^j_{2,2}, I^j_{1,2}\right\} = I^j_{2,1} + I^j_{2,2} = \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} = 2\right\}$. Under Assumption 2,

$$\sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} = 2\right\} \geq I\varepsilon$$

then

$$\begin{aligned} & P\left(|\mathcal{J}_1| \geq \frac{S}{\delta I}\right) \\ & \leq P\left(\sum_{j \in \mathcal{J}_1} I^j_{2,1} + I^j_{2,2} \geq \frac{S}{\delta I} \cdot I\varepsilon\right) \\ & \leq P\left(\sum_i \sum_j \left(P_{ij} - \overleftarrow{\theta}_{j,\widehat{z}_i}\right)^2 \geq \frac{\varepsilon S}{2}\right). \end{aligned}$$

By noting $\sum_i \sum_j \left(P_{ij} - \overleftarrow{\theta}_{j,\widehat{z}_i}\right)^2 = o_p(S)$, then $|\mathcal{J}_1| = o_p(S/I)$. Similar arguments yield $|\mathcal{J}_2| = o_p(S/I)$, which concludes the proof of Lemma 5. □

Note (A.17) implies that $\min\left\{I^j_{2,1}, I^j_{1,1}\right\} + \min\left\{I^j_{2,2}, I^j_{1,2}\right\} \neq I^j_{1,1} + I^j_{2,2}, \forall j \in \mathcal{J}$, thus $\mathcal{J} = \mathcal{J}_0 \cup \mathcal{J}_1 \cup \mathcal{J}_2$. Lemma 5 implies that when δ_J goes to 0 with a mild rate, the number of elements in $\mathcal{J}_0$ dominates the number of elements in $\mathcal{J}_1 \cup \mathcal{J}_2$; thus for most $j \in \mathcal{J}$, $\min\left\{I^j_{2,1}, I^j_{1,1}\right\} + \min\left\{I^j_{2,2}, I^j_{1,2}\right\}$ should be $I^j_{2,1} + I^j_{1,2} = \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} \neq \widehat{z}_{ij}\right\}$, which represents the number of subjects with the incorrectly assigned local latent classes.

Step 5. (A.18) implies that $\sum_i \sum_j \left(P_{ij} - \overleftarrow{\theta}_{j,\widehat{z}_i}\right)^2 \geq \delta \sum_{j \in \mathcal{J}_0} \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} \neq \widehat{z}_{ij}\right\}/2$. Next we focus on obtaining a lower bound of $\sum_{j \in \mathcal{J}_0} \sum_{i=1}^{I} \mathcal{I}\left\{z^0_{ij} \neq \widehat{z}_{ij}\right\}$ to control the classification error rate $I^{-1} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\}$.

Motivated by Assumption 2, for each latent attribute k, denote j^1_k as the smallest integer j such that item j has a $\boldsymbol{q}$-vector $\boldsymbol{e}_k$, and denote j^2_k as the second smallest integer j such that $\boldsymbol{q}_j = \boldsymbol{e}_k$, etc. For positive integer m, denote

$$\mathcal{B}^m = \{j^m_1, \ldots, j^m_K\}. \tag{A.19}$$

For each $k \in \{1, \ldots, K\}$, denote

$$J_{\min} = \min_{1 \leq k \leq K} \left|\left\{j \in \mathcal{J}_0; \boldsymbol{q}^0_j = \boldsymbol{e}_k\right\}\right|, \tilde{J}_{\min} = \min_{1 \leq k \leq K} \left|\left\{j \in \mathcal{J}; \boldsymbol{q}^0_j = \boldsymbol{e}_k\right\}\right|. \tag{A.20}$$

Then, we have that $\mathcal{B}^m \cap \mathcal{B}^l = \varnothing$ for any $m \neq l$, thus

$$\begin{aligned} & \sum_{i=1}^{I} \sum_{j \in \mathcal{J}_0} \mathcal{I}\left\{\xi\left(\boldsymbol{q}^0_j, \boldsymbol{\alpha}^0_i\right) \neq \xi\left(\boldsymbol{q}^0_j, \widehat{\boldsymbol{\alpha}}_i\right)\right\} \\ & \geq \sum_{i=1}^{I} \sum_{m=1}^{J_{\min}} \sum_{j \in \mathcal{B}^m} \mathcal{I}\left\{\xi\left(\boldsymbol{q}^0_j, \boldsymbol{\alpha}^0_i\right) \neq \xi\left(\boldsymbol{q}^0_j, \widehat{\boldsymbol{\alpha}}_i\right)\right\} \\ & = J_{\min} \sum_{i=1}^{I} \sum_{k=1}^{K} \mathcal{I}\left\{\xi\left(\boldsymbol{e}_k, \boldsymbol{\alpha}^0_i\right) \neq \xi\left(\boldsymbol{e}_k, \widehat{\boldsymbol{\alpha}}_i\right)\right\}. \end{aligned} \tag{A.21}$$

The last inequality holds since $\sum_{k=1}^{K} \mathcal{I}\left\{\xi\left(\boldsymbol{e}_k, \boldsymbol{\alpha}^0_i\right) \neq \xi\left(\boldsymbol{e}_k, \widehat{\boldsymbol{\alpha}}_i\right)\right\} \geq \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\}$. Note (A.21) implies $o_p(S/I) \geq J_{\min} I^{-1} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\}$. For simplicity,

$$\gamma_J = \frac{S}{IJ} = J^{-\frac{1}{\eta}} (\log J)^{\frac{\xi}{\eta}}.$$

Note that (32) in Assumption 2 implies that $|\mathcal{J}|/J \geq \tilde{J}_{\min}/J \geq \delta_J$ and $J_{\min} \geq \tilde{J}_{\min} - |\mathcal{J}_1 \cup \mathcal{J}_2|$; by plugging these results into $o_p(S/I) \geq J_{\min} I^{-1} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\}$, we can obtain

$$o_p\left(\frac{S}{I}\right) + |\mathcal{J}_1 \cup \mathcal{J}_2| \geq \frac{\tilde{J}_{\min}}{I} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\} \geq \frac{J\delta_J}{I} \sum_{i=1}^{I} \mathcal{I}\left\{\boldsymbol{\alpha}^0_i \neq \widehat{\boldsymbol{\alpha}}_i\right\}.$$

From Lemma 5, we have $|\mathcal{J}_i| = o_p(S/I)$ for $i = 1,2$, which implies that $|\mathcal{J}_1 \cup \mathcal{J}_2| = o_p(S/I)$. By substituting this into the above inequality, we can conclude that

$$o_p\left(\frac{S}{I}\right) \geq \frac{J\delta_J}{I}\sum_{i=1}^{I}\mathcal{I}\left\{\boldsymbol{\alpha}_i^0 \neq \widehat{\boldsymbol{\alpha}}_i\right\},$$

which is equivalent to $I^{-1}\sum_{i=1}^{I}\mathcal{I}\left\{\boldsymbol{\alpha}_i^0 \neq \widehat{\boldsymbol{\alpha}}_i\right\} = o_p(\gamma_J/\delta_J)$. The proof of this theorem is complete.

The inequality (32) in Assumption 2 bridges between the misclassification error for the local latent classes and the misclassification error for the latent attribute mastery patterns $\widehat{\mathcal{A}}$ by using the inequality $\sum_{k=1}^{K}\mathcal{I}\left\{\xi\left(\boldsymbol{e}_k, \boldsymbol{\alpha}_i^0\right) \neq \xi\left(\boldsymbol{e}_k, \widehat{\boldsymbol{\alpha}}_i\right)\right\} \geq \mathcal{I}\left\{\boldsymbol{\alpha}_i^0 \neq \widehat{\boldsymbol{\alpha}}_i\right\}$.

A.6. Proof of Theorem 2

For notational simplicity, denote $i_{ja}^0 = \sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}$. Thus, Assumption 2 implies that $\forall \boldsymbol{\alpha} \in \{0,1\}^K, \sum_{i=1}^{I}\mathcal{I}\left\{\boldsymbol{\alpha}_i^0 = \boldsymbol{\alpha}\right\} \geq \varepsilon I$ and

$$i_{ja}^0 \geq \frac{2^K}{2^{K_j}} I\varepsilon \geq I\varepsilon. \tag{A.22}$$

Recall that

$$\tilde{\theta}_{ja} = \frac{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}}.$$

Rewrite θ_{ja}^0 as similar form

$$\theta_{ja}^0 = \frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}\theta_{ja}^0}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}} = \frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}P_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}}.$$

By triangle inequality, we have

$$\begin{aligned}
&\max_{j,a}\left|\tilde{\theta}_{ja} - \theta_{ja}^0\right| \\
&= \max_{j,a}\left|\frac{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}} - \frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}P_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}}\right| \\
&\leq \max_{j,a}\left|\frac{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}} - \frac{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}}\right| \\
&\quad + \max_{j,a}\left|\frac{\sum_{i=1}^{I}\mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}} - \frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}}\right| \\
&\quad + \max_{j,a}\left|\frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}X_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}} - \frac{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}P_{ij}}{\sum_{i=1}^{I}\mathcal{I}\left\{z_{ij}^0 = a\right\}}\right| \\
&\equiv \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3.
\end{aligned}$$

Thereafter, we analyze these three terms separately. For the first term,

$$\mathcal{I}_1 \leq \max_{j,a}\left(\sum_i \mathcal{I}\left\{\widehat{z}_{ij} = a\right\}X_{ij}\right)\cdot\frac{\sum_i\left|\mathcal{I}\left\{\widehat{z}_{ij} = a\right\} - \mathcal{I}\left\{z_{ij}^0 = a\right\}\right|}{i_{ja}^0\sum_i \mathcal{I}\left\{\widehat{z}_{ij} = a\right\}}$$

$$\leq \max_{j,a} \frac{\sum_i \left|\mathcal{I}\left\{\widehat{z}_{ij}=a\right\}-\mathcal{I}\left\{z_{ij}^0=a\right\}\right|}{i_{ja}^0}$$
$$\leq \frac{1}{\varepsilon I}\sum_i \mathcal{I}\left\{\boldsymbol{\alpha}_i^0 \neq \widehat{\boldsymbol{\alpha}}_i\right\} = o_p\left(\frac{\gamma_J}{\delta_J}\right).$$

The last inequality holds since $\forall j \in [J], j \in [L_j], \sum_i \left|\mathcal{I}\left\{\widehat{z}_{ij}=a\right\}-\mathcal{I}\left\{z_{ij}^0=a\right\}\right| \leq \sum_i \mathcal{I}\left\{\boldsymbol{\alpha}_i^0 \neq \widehat{\boldsymbol{\alpha}}_i\right\}$. For the second term, we have

$$\mathcal{I}_2 = \max_{j,a} \frac{\sum_i \left|X_{ij}\left(\mathcal{I}\left\{\widehat{z}_{ij}=a\right\}-\mathcal{I}\left\{z_{ij}^0=a\right\}\right)\right|}{i_{ja}^0} \leq \max_{j,a} \frac{\sum_i \left|\mathcal{I}\left\{\widehat{z}_{ij}=a\right\}-\mathcal{I}\left\{z_{ij}^0=a\right\}\right|}{i_{ja}^0}.$$

For the same reason as $\mathcal{I}_1 \xrightarrow{P} 0$, we can also conclude that $\mathcal{I}_2 = o_p(\gamma_J/\delta_J)$; thus, $\mathcal{I}_1 + \mathcal{I}_2 = o_p(\gamma_J/\delta_J)$. For the third term, we apply Hoeffding's inequality for bounded random variables and obtain

$$P\left(\frac{\sum_i \mathcal{I}\left\{z_{ij}^0=a\right\}\left(X_{ij}-P_{ij}\right)}{i_{ja}^0} \geq t\right) \leq 2\exp\left(-2i_{ja}^0 t^2\right) \leq 2\exp\left(-2\varepsilon I t^2\right).$$

Note the number of (j,a) pairs less than or equal to $J2^K$ under Assumption 2, we have for $\forall t > 0$,

$$P(\mathcal{I}_3 \geq t) \leq J2^{K+1}\exp\left(-2\varepsilon I t^2\right). \tag{A.23}$$

Notably, 2^{K+1} remains a constant as K is fixed. By choosing $t = 1/\sqrt{I^{1-\tilde{c}}}$ for a small $\tilde{c} > 0$, the tail probability in A.23 converges to zero when the scaling condition $\sqrt{J} = O\left(I^{1-c}\right)$ holds. This implies that $\mathcal{I}_3 = o_p\left(1/\sqrt{I^{1-\tilde{c}}}\right)$. Bringing together the preceding results, we have

$$\max_{j,a}\left|\tilde{\theta}_{ja}-\theta_{ja}^0\right| = o_p\left(\frac{\gamma_J}{\delta_J}\right) + o_p\left(\frac{1}{\sqrt{I^{1-\tilde{c}}}}\right).$$

Note for any (j,a), we have $\left(\tilde{\theta}_{ja}-\widehat{\theta}_{ja}\right)^2 \leq \check{C} i_{ja}^{-1}$ and $i_{ja} \geq I\varepsilon_2$ with the probability approaching 1 ; thus $\max_{j,a}\left|\tilde{\theta}_{ja}-\widehat{\theta}_{ja}\right| = O_p\left(I^{-1/2}\right)$. Therefore,

$$\max_{j,a}\left|\widehat{\theta}_{ja}-\theta_{ja}^0\right| \leq \max_{j,a}\left|\tilde{\theta}_{ja}-\theta_{ja}^0\right| + \max_{j,a}\left|\tilde{\theta}_{ja}-\hat{\theta}_{ja}\right| = o_p\left(\frac{\gamma_J}{\delta_J}\right) + o_p\left(\frac{1}{\sqrt{I^{1-\tilde{c}}}}\right).$$

□

Cite this article: Yamaguchi, K., Liu, Y. and Xu, G. (2025). Generalized Bayesian method for diagnostic classification models. *Psychometrika*, 1–38. https://doi.org/10.1017/psy.2024.20