Received: 9 November 2022 | Accepted: 14 August 2023

DOI: 10.1111/bmsp.12322

the british psychological society
promoting excellence in psychology

ARTICLE

A Bayesian nonparametric approach for handling item and examinee heterogeneity in assessment data

Tianyu Pan[1] | **Weining Shen**[1] | **Clintin P. Davis-Stober**[2] | **Guanyu Hu**[3]

[1]Department of Statistics, University of California, Irvine, California, USA

[2]Department of Psychological Sciences, University of Missouri – Columbia, Columbia, Missouri, USA

[3]Department of Biostatistics and Data Science, Center for Spatial Temporal Modeling for Applications in Population Sciences, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Correspondence
Guanyu Hu, 1200 Pressler Street, Houston, TX 77030, USA.
Email: guanyu.hu@uth.tmc.edu

Abstract

We propose a novel nonparametric Bayesian item response theory model that estimates clusters at the question level, while simultaneously allowing for heterogeneity at the examinee level under each question cluster, characterized by a mixture of binomial distributions. The main contribution of this work is threefold. First, we present our new model and demonstrate that it is identifiable under a set of conditions. Second, we show that our model can correctly identify question-level clusters asymptotically, and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a $\sqrt{n}$ rate (up to a log term). Third, we present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. Compared to the existing methods, our model manages to reveal the multi-dimensionality of the examinees' proficiency level in handling different types of questions parsimoniously by imposing a nested clustering structure. The proposed model is evaluated via a series of simulations as well as apply it to an English proficiency assessment data set. This data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

KEYWORDS

IRT model, model averaging, nonparametric Bayesian method, posterior contraction rate, Rasch model

1 | INTRODUCTION

Item response theory (IRT; Mislevy & Verhelst, 1990; Rost, 1990) was developed to better understand the mechanism behind an examinee answering a question (item) correctly and to evaluate the discrepancies between examinees or items. The majority of IRT models assume that the response of an examinee to a question is probabilistic (Thomas, 2011), governed by a latent parameter. For example, an accuracy parameter is between 0 and 1 if the response is true or false. In general, the latent probabilistic

parameter relies on two main factors: the ability of an examinee and the difficulty of a question. Among all IRT models, the Rasch model (Rasch, 1993) is one of the best-known and most widely used models thanks to its elegant format and interpretability. When applied to a dichotomous matrix, where each column represents a test question and each row is an examinee's response, the Rasch model can be encapsulated as a logistic regression that models the (i,j)th entry of the matrix using the linear predictor that contrasts the ability of the ith examinee and the difficulty of the jth question. While our proposed framework builds upon the Rasch model, future work could consider extensions to other IRT frameworks such as two-parameter logistic (Muraki, 1992) and three-parameter logistic models (Rouse et al., 1999; Zumbo et al., 1997). We refer readers to Thomas (2011) for a more comprehensive review on the history of IRT models.

Despite the usefulness of the IRT models, there is evidence that two key assumptions, *unidimensionality* and *local independence* (Andrich & Marais, 2019), may be violated in common applications (Bell et al., 1988; Keith, 1987; Kreiner & Christensen, 2007; Marais & Andrich, 2008), that is, the ability of an examinee cannot be entirely represented by a single parameter (unidimensionality) and the responses remain dependent after imposing the IRT model structure (local independence). From a statistical point of view, violations of these assumptions are intertwined. When the mechanism behind the true data-generating process cannot be recovered by using a single ability or difficulty parameter that corresponds to an examinee or a question, it undoubtedly introduces a conditional dependency on the responses due to the lack of fit. Intuitively, unidimensionality is a proper assumption to unveil the Guttman pattern (Andrich & Marais, 2019), that is, the examinees who can correctly answer difficult questions with a certain probability should be able to answer easy questions with a higher probability. Nevertheless, this probably neglects the fact that some examiness could be more expert in handing certian questions rather than other questions. Throughout this paper, we use the term *heterogeneity* to describe the phenomenon of an examinee being able to correctly answer classes of questions based on content type that is not captured by a unidimensional ability or difficulty parameter. A popular approach for accommodating multidimensional IRT data is to impose mixing structures simultaneously on the ability of examinees and the difficulty of questions, which generates a mixture of Rasch models as a result. This idea has been investigated using frequentist (Alexeev et al., 2011; Rost, 1990) and Bayesian (Bolt et al., 2002; Hu et al., 2020; Jang et al., 2018; Miyazaki & Hoshino, 2009; Sen et al., 2019) approaches. Indeed, these methods enjoy more robustness and interpretability by allowing heterogeneity in the ability of examinees and the difficulty of questions, yet they are limited by the structure of the Rasch model; for example, these two main factors have to be linearly contrasted in each Rasch model mixture. Technically, using only the first-order information of the two main factors while ignoring possible interactions can lead to lack of fit when the two main factors truly interact with one another in complex ways. Bartolucci et al. (2017) solves this problem by integrating the two main factors into a single accuracy parameter for each question while assuming a global mixing structure on the accuracy parameters and letting them share a common sorting order over the mixing components. The resutls from Bartolucci et al. (2017) can be interpreted as those examinees who have higher accuracy when answering questions of that same type. In addition, such construction introduces heterogeneity and dependency between the two main factors by quantifying an examinee's proficiency on a question using a mixture of accuracy parameters.

From a technical point of view, a mixing structure on the ability of the examinees and the difficulty of the questions induces a binomial mixture model with each logit-transformed accuracy parameter expressed as a linear combination of the two main factors. Therefore, the interpretability of the two main factors is highly dependent on the identifiability of the binomial mixture model induced by the mixing structure. However, it is worth noting that none of the aforementioned works address the identifiability problem in the context of a binomial mixture distribution, which is crucial since it is not difficult to discover that the density functions are equivalent between $\text{Bernoulli}(0.5)$ and $0.5 \times \text{Bernoulli}(0.1) + 0.5 \times \text{Bernoulli}(0.9)$. If the model is not identifiable, researchers cannot expect a fast mixing when performing the Gibbs sampling, and even question the necessity of introducing a mixing structure, so let alone ensure the consistency on the mixing parameters and interpret the results.

Motivated by the works introduced above, we propose a multidimensional IRT model based on a nonparametric Bayesian procedure, which is termed the 'averaged constrained binomial mixture' (ACBM). The objective of this paper is to relax the independence assumption implied by the Rasch model, to model the heterogeneity, and to address the identifiability issue. Before we present our model, we might first imagine a pattern of heterogeneity at both the examinee level and the question level whereby examinees form several groups in each hypothetical cluster of the questions according to their proficiency in handling questions of this type, while the grouping pattern could differ between the question clusters. To model this idea, we consider the following steps. First, given a dichotomous response matrix, we aim to discover a partition over all questions such that in each question cluster, an examinee's responses to these questions can be characterized by a binomial distribution, of which the accuracy parameter (i.e., the examinee's proficiency in answering these questions) follows a mixing distribution. Technically, our proposed Bayesian model is essentially a prior over all mixing distributions given a partition on these questions, jointly with a prior over all possible partitions on these questions. The key novelties of our proposed model are in interpretation and theoretical guarantees of identifiability and posterior consistency. Our model can infer a partition on questions that reveals information at both levels. At the question level, in each question cluster, the examinees are automatically distinguished by their proficiency in tackling these questions using a mixing distribution. This allows us to discover a more complex accuracy pattern than just simple Guttman patterns (Andrich & Marais, 2019). At the examinee level, each mixing accuracy parameter under a given question cluster represents the proficiency of a specific group of examinees in handling these questions, which provides information for 'precision education', that is, statistical evidence to the test maker to identify the examinees who are not skilled in solving a certain question type. This information, in turn, is helpful in designing and implementing additional questions of this type. In addition, the identifiability of our model is ensured by putting a dynamic upper bound on the number of mixing components in each question cluster, where the upper bound is determined by the size of the question cluster. We later address the identifiability of our model in Lemma 1. Our method can be tractably applied using a Markov chain Monte Carlo (MCMC) sampling algorithm for realization and is able to capture the true question partition and estimate the mixing parameters at a $\sqrt{n}$ rate (up to a log term), with n defined as the number of examinees, thanks to the rapid developments of Bayesian analysis over the past 30 years, in both computational (Ishwaran & James, 2001; Neal, 2000) and theoretical (Ghosal & Van der Vaart, 2017; Guha et al., 2019; Ho & Nguyen, 2016; Nobile, 1994; Shen et al., 2013) research.

The rest of this article is organized as follows. A motivation and interpretation of our model is given in Section 2. The convergence results are presented in Section 3, with proofs deferred to the Appendix S1. In Section 4 we outline the MCMC sampling algorithm and introduce statistics to summarize the posterior samples obtained. The simulation study that validates our theoretical results and compares the performance between our model and the Rasch model is provided in Section 5. We carry out a real data analysis applying our model to an English language assessment data set in Section 6. In Section 7 we discuss several possible ways to generalize our model, which paves the way for future study. For ease of exposition, proofs, computation algorithms, and additional technical results are given in Appendix S1.

2 | MOTIVATION

Consider an $n \times D$ dichotomous matrix $\mathbf{X}$, whose (i,j)th entry is a random variable, denoted by X_{ij}, which takes value 1 if the ith examinee answered the jth question correctly and 0 otherwise. We let $\mathscr{D} \equiv \{1, 2, \ldots, D\}$ be an index set on the total number of items, $\mathscr{C}$ be a partition of $\mathscr{D}$, and $c \in \mathscr{C}$ be a cluster of items defined by the partition $\mathscr{C}$. To illustrate the connection between $\mathscr{C}$ and $\mathscr{D}$ using an example, we take D to be 3 and a partition $\mathscr{C}$ of $\mathscr{D}$ can be $\mathscr{C} = \{\{1,2\},\{3\}\}$. For this simple example, there are just two clusters, $c_1 = \{1,2\}$ and $c_2 = \{3\}$. Our model construction is motivated by the following example. Suppose a professor creates a test to evaluate her students, with the questions being classified into multiple clusters, each cluster being an element $c \in \mathscr{C}$. Each question cluster c can ideally distinguish different

types of students, of which the total number of types is given by $K^{(c)}$. Moreover, the kth type indicates the proficiency of a specific group of students in answering a certain question type c (i.e., heterogeneity), which can further be quantified by an accuracy parameter $\theta_k^{(c)}$, for $1 \le k \le K^{(c)}$. Inspired by this idea, we propose the following hierarchical order to model this structural clustering pattern:

$$\begin{aligned} X_{1j}, \ldots, X_{nj} &\overset{\text{i.i.d}}{\sim} p_F^{(c)}, \quad \forall j \in c, \ \forall c \in \mathscr{C}, \\ p_F^{(c)}(x) &= \sum_{k=1}^{K^{(c)}} w_k^{(c)} \times (1-\theta_k^{(c)})^{1-x} \times (\theta_k^{(c)})^{x}, \quad \text{for } x = 0 \text{ or } 1, \\ (w_1^{(c)}, \ldots, w_{K^{(c)}}^{(c)}) &\sim \text{Dir}(K^{(c)}, (\alpha, \ldots, \alpha)), \\ \theta_k^{(c)} &\overset{\text{i.i.d}}{\sim} \text{Beta}(a_0, b_0), \quad \text{for } k = 1, \ldots, K^{(c)}, \\ K^{(c)} &\sim q_0^{(c)}(K^{(c)}; \gamma), \\ \mathscr{C} &\sim m(\mathscr{C}), \end{aligned} \tag{1}$$

where $\text{Dir}(K, (\alpha, \ldots, \alpha))$ refers to a Dirichlet distribution with K categories and a concentration parameter $\alpha > 0$, $\text{Beta}(a_0, b_0)$ denotes a beta distribution whose shape parameters are a_0 and b_0, $m(\mathscr{C})$ denotes a probability mass function over all possible partitions on $\mathscr{D}$, $q_0^{(c)}(\cdot; \gamma)$ denotes the Poisson distribution parameterized by γ and truncated between 1 and $(|c|+1)/2$, with $|\cdot|$ referring to the cardinality of a set rather than its absolute value. Note that the truncation on $q_0^{(c)}(\cdot; \gamma)$ refers to the maximum limit of a question cluster c in dividing the students into different levels by their proficiency in answering a certain type of questions. This limitation relies on the question cluster size $|c|$. The term $(|c|+1)/2$ is the upper limit on the value $K^{(c)}$ and is a function of the number of items within the cluster c. This upper limit can be interpreted as the 'resolution' of the question cluster. The superscript (c) highlights that the mixing weights and the number of mixtures are allowed to vary across question clusters. The main interest of our Bayesian model is to infer $\{w_i^{(c)}\}_{i=1}^{K^{(c)}}$, $\{\theta_i^{(c)}\}_{i=1}^{K^{(c)}}$, $K^{(c)}$, for all $c \in \mathscr{C}$ and $\mathscr{C}$ using the posterior distribution, which is expressed as follows:

$$\begin{aligned} &\pi\left(\left\{\{w_i^{(c)}\}_{i=1}^{K^{(c)}}, \{\theta_i^{(c)}\}_{i=1}^{K^{(c)}}, K^{(c)}\right\}_{c \in \mathscr{C}}, \mathscr{C} | \mathbf{X}\right) \\ &\propto \prod_{i=1}^{n} \prod_{c \in \mathscr{C}} \prod_{j \in c} \left(\sum_{k=1}^{K^{(c)}} w_k^{(c)} \times (1-\theta_k^{(c)})^{1-X_{ij}} \times (\theta_k^{(c)})^{X_{ij}} \right) \\ &\quad \times \pi_w\left(\{w_k^{(c)}\}_{k=1}^{K^{(c)}} | K^{(c)}\right) \times \pi_B\left(\{w_k^{(c)}\}_{k=1}^{K^{(c)}} | K^{(c)}\right) \\ &\quad \times \pi_K\left(K^{(c)}\right) \times m(\mathscr{C}), \end{aligned} \tag{2}$$

where π_w, π_B, and π_K are the density functions of the Dirichlet distribution, the beta distribution, and the Poisson distribution, respectively. Empirically, we integrate out $\{w_k^{(c)}\}_{k=1}^{K^{(c)}}$ and $K^{(c)}$ following the procedure introduced in Green and Richardson (2001) and Miller and Harrison (2018) to reduce the C parameter dimensionality for efficient sampling.

The main feature of our model lies in the dependency between the clustering structures at two levels. Specifically, the mixing distribution at the examinee level is allowed to vary between question groups, whereas most existing methods (Bolt et al., 2002; Hu et al., 2020; Jang et al., 2018; Miyazaki & Hoshino, 2009; Sen et al., 2019) suggest that the mixing distribution at the examinee level is invariant regardless of the heterogeneity at the question level. We conclude three main benefits for this feature of our model. Our model provides more interpretable results (e.g., heterogeneity among examinees for each question type) compared with heuristically assuming mixing distributions on the two main factors separately, which can hardly be interpreted in this way. Such structural heterogeneity allows us to discover more complicated patterns than the Guttman pattern. For example, in a mathematics test, it is reasonable to believe that some examinees are more proficient in algebra questions than they are in geometric questions. Other examinees are more expert in solving geometric questions but not good

at algebra questions. This cannot be explained by the Guttman mechanism but can be justified by the heterogeneity. In addition, we demonstrate that our model is identifiable when our chosen upper bound on the number of mixtures is imposed for each question cluster. Indeed, such an upper bound could inevitably lead to information loss when the size of a question cluster is small, but it intuitively makes sense because one cannot distinguish the proficiency of examinees at a certain type of questions only using very few of them. Moreover, the identifiability of our model further contributes to the identification of the true clustering structure on the questions, while providing $\sqrt{n}$ (up to a log term) estimates to the mixing weights and parameters in each question cluster under mild conditions.

In the next section we present the technical details of our proposed method and introduce the posterior consistency results.

3 | CONVERGENCE RESULTS

3.1 | Notation

In addition to the notation introduced in Section 2, throughout the rest of this paper, we will use $f(\cdot)$ or f to denote a function $f(x)$ if it only takes a single argument. We let $X_j^{(c)}$ denote the cluster section of random vector X_j, where X_j refers to the dichotomous response of the jth examinee. A similar definition applies to $x^{(c)}$ with respect to vector $\mathbf{x}$. We also let

$$p_{F_0}^{(c)}(x) = \sum_{k=1}^{K_0^{(c)}} w_{k,0}^{(c)} \times (1-\theta_{k,0}^{(c)})^{1-x} \times (\theta_{k,0}^{(c)})^{x} \tag{3}$$

be the true density function associated with a single question in cluster c. The corresponding true numbers of mixtures, mixing weights, and componentwise parameters of $p_{F_0}^{(c)}(x)$ are therefore $K_0^{(c)}$, $\{w_{k,0}^{(c)}\}_{k=1}^{K_0^{(c)}}$, and $\{\theta_{k,0}^{(c)}\}_{k=1}^{K_0^{(c)}}$. Analogously, a density function sampled following (1) is denoted by $p_F^{(c)}(x)$, whose number of mixtures, mixing weights, and componentwise parameters are $K^{(c)}$, $\{w_k^{(c)}\}_{k=1}^{K^{(c)}}$, and $\{\theta_k^{(c)}\}_{k=1}^{K^{(c)}}$. We denote the true partition by $\mathscr{C}_0$ and any partition sampled following $m(\cdot)$ by $\mathscr{C}$. For random vector X_j, we let $p_{F_0}(\mathbf{x})$ and $p_F(\mathbf{x})$ be the true density function and a sampled density function following (1), respectively. We also let P_0 and $\mathbb{P}_{F_0}$ be the probability distribution and the probability measure induced by p_{F_0}, and the expectation taken under $\mathbb{P}_{F_0}$ is denoted by $\mathbb{P}_{F_0}[f]$ or $\mathbb{P}_{F_0} f$.

For succinctness, we denote the prior on p_F in (1) by Π and let $\Pi^{(c)}$ for all $c \in \mathscr{C}$ be the prior on $p_F^{(c)}$ given $\mathscr{C}$. We define the set of all possible binomial mixtures as

$$\begin{aligned}\mathscr{P}^{(c)} = \bigcup_{K=1}^{+\infty} \mathscr{P}^{(c)}(K) \equiv \Bigg\{ p_F^{(c)}(x^{(c)}) : p_F^{(c)}(x^{(c)}) = \sum_{k=1}^{K} w_k^{(c)} (1-\theta_k^{(c)})^{|c|-\|x^{(c)}\|_1} \times (\theta_k^{(c)})^{\|x^{(c)}\|_1}, \\ w_k^{(c)} \in (0,1), \text{ for } k=1,\ldots,K, \ \{\theta_k^{(c)}\}_{k=1}^{K} \text{ are distinct} \Bigg\},\end{aligned} \tag{4}$$

where the operator $\|\cdot\|_1$ refers to the ℓ_1 norm when applied to a vector, and in our situation it is equivalent to a summation function as $x^{(c)}$ is a non-negative dichotomous vector. We point out that the prior $\Pi^{(c)}$ is essentially supported on $\mathscr{P}^{(c)}$ by truncating K between 1 and $(|c|+1)/2$. For every $\mathscr{C}$ in the support of $m(\mathscr{C})$, we let $\Pi_{\mathscr{C}} \equiv \prod_{c\in\mathscr{C}} \Pi^{(c)}$ denote the prior on $\prod_{c\in\mathscr{C}} p_F^{(c)}$. It follows that

$$\Pi = \sum_{\mathscr{C}\in\mathcal{S}} m(\mathscr{C}) \Pi_{\mathscr{C}}, \tag{5}$$

where $\mathcal{S}$ refers to the collection of all possible partitions of $\mathscr{D}$. For every $c \in \mathscr{C}$ and every $\mathscr{C} \in \mathcal{S}$, we use $p_{F_0}^{(c)}(x^{(c)}) \equiv \int p_{F_0}(\mathbf{x}) \mathrm{d}x^{(\mathscr{D}\setminus c)}$ to denote the marginal density of p_{F_0} on the subvector $x^{(c)}$ and $p_F^{(c)}(x^{(c)})$ to denote the joint density function on $x^{(c)}$ given $p_F^{(c)}(x)$, with similar definition to (3).

3.2 | Convergence results

We begin with the interpretations of the following three assumptions:

(A1) The true partition $\mathscr{C}_0$ is in the support of $m(\mathscr{C})$.

(A2) For all $c \in \mathscr{C}_0$, the following properties hold for $p_{F_0}^{(c)}$:

$$\begin{aligned} &\text{Distinctness: } 0 < \theta_{i,0}^{(c)} \neq \theta_{j,0}^{(c)} < 1, \text{ for } 1 \le i \neq j \le K_0^{(c)}. \\ &\text{Non-trivial weights: } w_{k,0}^{(c)} > 0, \text{ for every } 1 \le k \le K_0^{(c)}. \\ &\text{Bounded component number: } K_0^{(c)} \le \frac{|c|+1}{2}. \end{aligned} \tag{6}$$

(A3) For every $\mathscr{C}$ in the support of $m(\mathscr{C})$, the true density has at least $\epsilon_0 > 0$ distance from the best estimation induced by $\mathscr{C}$ with respect to the Kullback–Leibler divergence,[1] that is,

$$\left\| \prod_{c \in \mathscr{C}_0} p_{F_0}^{(c)}(x^{(c)}) - \prod_{c \in \mathscr{C}} p_{F^*}^{(c)}(x^{(c)}) \right\|_1 > \epsilon_0,$$

where $p_{F^*}^{(c)}(x^{(c)})$, for all $c \in \mathscr{C}$, is obtained by minimizing $\mathrm{KL}\left(p_{F_0}^{(c)}(x^{(c)}); p_F^{(c)}(x^{(c)})\right)$, with respect to $p_F^{(c)}(x^{(c)}) \in \mathscr{P}^{(c)}$.

The first two assumptions are quite standard. Assumption (A1) is necessary to ensure our model is correctly specified. The only eye-catching part of Assumption (A2) is the constraint on the number of components under each question cluster. This constraint aims to guarantee that the binomial mixture part in our proposed model is identifiable, as noted by Teicher (1963). Later, in Lemma 1, we will prove that the binomial mixture model is first-order identifiable. Assumption (A3) is a weaker condition than the general identifiability assumption. In fact, the density function $\prod_{c \in \mathscr{C}} p_F^{(c)}(x^{(c)})$ is non-identifiable for some instances of $\left\{ \{w_k^{(c)}, \theta_k^{(c)}\}_{c \in \mathscr{C}}, \mathscr{C} \right\}$, for example, the corresponding density functions are identical given different parameterizations

$$\{\{w_1^c = 1, \theta_1^{(c)} = 0.5\}_{c=\{1,2\}}, \{w_1^c = 1, \theta_1^{(c)} = 0.5\}_{c=\{3\}}\}$$

and $\{\{\theta_1^{(c)} = 1, \theta_1^{(c)} = 0.5\}_{c=\{1,2,3\}}\}$ when $D = 3$. Therefore, one cannot directly apply Doob's theorem (Doob, 1949) but needs to resort to Assumption (A3) and Theorem 1 for the consistency on $\mathscr{C}$.

Theorem 1. Assume (A1) and (A3) are satisfied. Then

$$\Pi(\mathscr{C} = \mathscr{C}_0 | X_1, \ldots, X_n) \to 1 \quad a.s.\ P_0. \tag{7}$$

Theorem 1 states that our proposed model can correctly identify the latent question partition asymptotically; for example, when the number of examinees increases, we can expect that the question clustering configuration sampled from the posterior distribution of our model will eventually converge to the true question partition.

[1] The Kullback–Leibler divergence between $f(x)$ and $g(x)$ is defined as

$$\mathrm{KL}(f(\mathbf{x}); g(\mathbf{x})) = \int \log\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) f(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

To pave the way to our next theorem, we prove that the binomial mixture model is first-order identifiable, which is defined as follows.

Lemma 1. (First-order identifiability) Assume that $\{\theta_i\}_{i=1}^K$ are distinct, that $\{\alpha_i\}_{i=1}^K$ and $\{\beta_i\}_{i=1}^K$ are real-valued coefficients, and that (A3) holds. Suppose that

$$\sum_{i=1}^{K} \alpha_i f(y|, \theta_i, n) + \sum_{i=1}^{K} \beta_i \frac{\partial f}{\partial \theta_i} f(y|, \theta_i, n) = 0, \tag{8}$$

where $f(y|, \theta_i, n) = \binom{n}{y}(1-\theta_i)^{n-y}(\theta_i)^y$ *and*

$$\begin{aligned}\frac{\partial f}{\partial \theta_i} f(y|, \theta_i, n) = \quad & \mathbb{1}(1 \le y \le n) n \binom{n-1}{y-1}(1-\theta_i)^{n-1-(y-1)}(\theta_i)^{y-1} \\ & - \mathbb{1}(0 \le y \le n-1) n \binom{n-1}{y}(1-\theta_i)^{n-1-y}(\theta_i)^y.\end{aligned}$$

Then $\alpha_1 = \beta_1 = \ldots = \alpha_K = \beta_K = 0$, *if* $1 \le K \le (n+1)/2$.

Lemma 1 is indispensable for estimating, at a $\sqrt{n}$ rate, the true mixing weights and the true componentwise parameters under each question cluster. The results of Theorem 1 and Lemma 1 yield our final result.

Theorem 2. Assume (A1)–(A3) are satisfied. Then the proposed model can estimate the true parameters a posteriori, given a contraction rate ϵ_n,

$$\begin{aligned}\Pi(\{p_F^{(c)}: |w_{\sigma^{(c)}(i)}^{(c)} - w_{i,0}^{(c)}| \lesssim M'\epsilon_n, \|\theta_{\sigma^{(c)}(i)}^{(c)} - \theta_{i,0}^{(c)}\|_2 \lesssim M'\epsilon_n, \quad \text{for } i = 1, \ldots, K^{(c)}, \\ K^{(c)} = K_0^{(c)}, \forall c \in \mathscr{C}, \mathscr{C} = \mathscr{C}_0\} | X_1, \ldots, X_n) \to 1, \quad a.s.\ P_0,\end{aligned} \tag{9}$$

where $M' > 0$ *is a universal constant,* $\epsilon_n = (\log(n))^t/\sqrt{n}$, *for any* $t > 1$.

Theorem 2 indicates that our model can detect the heterogeneity in examinees while identifying the true question partition.

4 | BAYESIAN INFERENCE AND ALGORITHM

We begin with the outline of our posterior sampling algorithm, which consists of the following four steps,

1. For each column (question), calculate the marginal likelihood when all students possess the same accuracy in answering this question, for example marginalizing the binomial likelihood over the beta prior in Equation (1).
2. Conditioning on the row (examinee) assignment under each column (question) cluster, update the column assignment by enumerating from column 1 to column D.
3. Conditioning on the column (question) assignment, update the row (examinee) assignment under each column (question) cluster by enumerating from row 1 to row n a total of n_{rep} times.
4. Loop between steps 2 and 3 n_{iter} times to approach the stationary distribution.

The detailed algorithm is deferred to Appendix S1. It is worth pointing out that step 2 is essentially Algorithm 1 proposed by Neal (2000), if each column (question) is treated as an 'individual', whose parameter is the rowwise partition on the examinees. To update the rowwise partition in step

3, we modify the aforementioned Algorithm 1 to accommodate the upper bound on the number of (student) mixtures under each column (question) cluster, which leads to a sampling scheme which is slightly different from theorem 4.1 of Miller and Harrison (2018). Based on the findings in our simulation studies, we notice that $n_{\text{iter}} = 200$ and $n_{\text{rep}} = 400$ are sufficient to obtain trustworthy posterior samples when $n \leq 1000$ and $D \leq 80$. In addition, for the hyper-parameters defined in (1), we let both a_0 and b_0 be 0.01 and γ equal to 1 such that the prior information is sufficiently non-informative but new clusters still have sufficient probability of being generated for both column (question) and row (examinee) in practice. The probability mass function $m(\mathscr{C})$ is chosen as the exchangeable partition probability function (EPPF; Pitman, 2006) of the mixture of finite mixtures model (MFM; Miller & Harrison, 2018) to ensure a closed form on the full conditional distribution when sampling. The most time-consuming task ($n = 1,000$ and $D = 80$) among the simulation studies and the real data analysis takes approximately 6 hours to finish after being assigned to a server with 94.24 GB RAM, 24 processing cores, operating at 3.33GHz.

To summarize the posterior samples, we use the following three statistics to estimate the column (question) partition, the rowwise (examinee) partition under each column cluster, and the componentwise accuracy under each column cluster. The column (question) partition is estimated using Dahl's estimate (Dahl, 2006), defined as

$$\begin{aligned} \hat{\ell} &= \underset{1\leq \ell \leq M}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\delta_{ij}(\mathscr{C}^{\text{Col}}(\ell)) - \hat{\pi}^{\text{Col}}_{ij}\right\}^2, \\ \hat{\mathscr{C}}^{\text{Col}} &= \mathscr{C}^{\text{Col}}(\hat{\ell}), \end{aligned} \tag{10}$$

where M is the number of MCMC iterations after burn-in, $\mathscr{C}^{\text{Col}}(\ell)$ refers to the column assignment at the ℓth iteration after burn-in, $\delta_{ij}(\mathscr{C}^{\text{Col}}(\ell))$ is an indicator function, defined as $\mathbb{1}(\mathscr{C}^{\text{Col}}_i(\ell) = \mathscr{C}^{\text{Col}}_j(\ell))$, with $\mathscr{C}^{\text{Col}}_i(\ell)$ denoting the clustering assignment of the ith column, and $\hat{\pi}^{\text{Col}}_{ij}$ is obtained by averaging $\delta_{ij}(\mathscr{C}^{\text{Col}}(\ell))$ over post-burn-in MCMC samples, namely, $\hat{\pi}^{\text{Col}}_{ij} = \frac{1}{M}\sum_{\ell=1}^{M}\delta_{ij}(\mathscr{C}^{\text{Col}}(\ell))$. The column partition summarized by Dahl's estimate is believed to be the most representative one as it minimizes the entrywise ℓ_2-distance between the self-concordance matrix of a given partition and the probability matrix $\hat{\pi}^{\text{Col}}_{ij}$ that any pair of columns i and j are clustered together. The rowwise (student) partition is then summarized from the iterations where the column (question) partition is equal to $\hat{\mathscr{C}}^{\text{Col}}$,

$$\begin{aligned} \hat{\ell} &= \underset{1\leq \ell \leq M;\mathscr{C}^{\text{Col}}(\ell)=\hat{\mathscr{C}}^{\text{Col}}}{\operatorname{argmin}} \sum_{d=1}^{D}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\delta_{ij}(\mathscr{C}^{\text{Row};d}(\ell)) - \hat{\pi}^{\text{Row};d}_{ij}\right\}^2, \\ \hat{\mathscr{C}}^{\text{Row};d} &= \mathscr{C}^{\text{Row};d}(\hat{\ell}), \quad \text{for } d = 1,\ldots,D, \end{aligned} \tag{11}$$

where $\mathscr{C}^{\text{Row};d}(\ell)$ refers to the row assignment of the dth column at the ℓth iteration after burn-in and $\hat{\pi}^{\text{Row};d}_{ij}$ is defined in a similar way with $\hat{\pi}^{\text{Col}}_{ij}$ for the dth column. It can be expected that ties happen for $\hat{\mathscr{C}}^{\text{Row};d}$ for all $d \in c$ and for all $c \in \hat{\mathscr{C}}^{\text{Col}}$ by definition. Analogous to the idea behind $\hat{\mathscr{C}}^{\text{Col}}$, $\hat{\mathscr{C}}^{\text{Row};d}$ looks for an iteration such that the squared ℓ_2-distance is minimized averaged over all columns. The componentwise accuracy under each column cluster is then estimated using a posterior mean given $\hat{\mathscr{C}}^{\text{Col}}$ and $\hat{\mathscr{C}}^{\text{Row};d}$ for $d = 1,\ldots,D$.

5 | SIMULATION

We study our proposed method using four data-generating processes (DGPs) and compare the result of our model with that given by the Rasch model (Rasch, 1993). The Rasch model is realized using the *tam* package in R. The four DGPs are designed to mimic the situations when the data are generated under our proposed model or the Rasch model, given increasing number of examinees ($n = 100,\ 300,\ 1000$).

We proceed by outlining the first four DGPs, deferring the details to Appendix S1. The first two DGPs are designed under our model,

- *DGP1*. Twenty questions are divided into five column (question) clusters, with three large question clusters and the remaining two questions individually forming two question clusters. Under each column (question) cluster, the accuracy within each mixture stays identical and the mixture number satisfies the constraint.
- *DGP2*. Sixty questions are divided into five column (question) clusters, with three large question clusters and the remaining questions individually forming two question clusters. Under each column (question) cluster, the accuracy within each mixture stays identical and the mixture number satisfies the constraint.

The last two DGPs are generated following the Rasch model, that is,

$$\begin{aligned} X_{i,j} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{i,j}), \quad \text{for } i=1,\ldots,n \text{ and } j=1,\ldots,D, \\ \theta_{i,j} &= \frac{\exp\{\xi_i - \psi_j\}}{1+\exp\{\xi_i - \psi_j\}}, \end{aligned} \tag{12}$$

with the two DGPs being presented as follows:

- *DGP3*. Twenty questions are divided into two column (question) clusters by letting ψ_j take a value from $\{-0.5, 0.5\}$ and three row clusters by letting ξ_i randomly take a value from $\{-2, 0, 2\}$, following the DGP defined in (12).
- *DGP4*. Sixty questions are divided into two column (question) clusters by letting ψ_j take a value from $\{-0.5, 0.5\}$ and three row clusters by letting ξ_i randomly take a value from $\{-2, 0, 2\}$, following the DGP defined in (12).

To validate Theorem 2, we consider the first two DGPs and adopt the following criteria:

$$\begin{aligned} \text{CWRI} &= \text{RI}(\widehat{\mathscr{C}}^{\text{Col}}, \mathscr{C}_0), \\ \text{ADK} &= \frac{1}{D}\sum_{d=1}^{D}\left|\left|\widehat{\mathscr{C}}^{\text{Row};d}\right| - K_0^{(c(d))}\right|, \\ \text{ADW} &= 1(\text{RI}(\widehat{\mathscr{C}}^{\text{Col}}, \mathscr{C}_0)=1)\times\frac{1}{|\mathscr{C}_0|}\sum_{c\in\mathscr{C}_0}\frac{1}{K_0^{(c)}}\min_{\sigma^{(c)}}\sum_{i=1}^{K_0^{(c)}}\left|\hat{w}^{(c)}_{\sigma^{(c)}(i)} - w^{(c)}_{0;i}\right| + \\ &\quad 1(\text{RI}(\widehat{\mathscr{C}}^{\text{Col}}, \mathscr{C}_0)\neq 1)\times 2, \\ \text{ADP} &= 1(\widehat{K}^{(c)} \geq K_0^{(c)}, c\in\mathscr{C}_0)\times\frac{1}{|\mathscr{C}_0|}\sum_{c\in\mathscr{C}_0}\sqrt{\frac{1}{K_0^{(c)}}\min_{\sigma^{(c)}}\sum_{i=1}^{K_0^{(c)}}\left\|\hat{\theta}^{(c)}_{\sigma^{(c)}(i)} - \theta^{(c)}_{0;i}\right\|_2^2} + \\ &\quad \left(1(\widehat{K}^{(c)} < K_0^{(c)}) \vee 1(c\notin\mathscr{C}_0)\right)\times 1, \end{aligned} \tag{13}$$

where CWRI, ADK, ADW and ADP denote the columnwise Rand index, averaged absolute difference in the rowwise number of component, averaged absolute difference in the rowwise weights, and averaged ℓ_2 -difference in the rowwise accuracy. $\text{RI}(\mathscr{C}, \mathscr{C}')$ denotes the Rand index (Rand, 1971) between $\mathscr{C}$ and $\mathscr{C}'$, $c(d)$ represents the column cluster c to which the dth column is assigned, $\sigma^{(c)}(\cdot)$ refers to the permutation operator, $w^{(c)}_{0;i}$ and $\theta^{(c)}_{0;i}$ denote the ith true mixing weight and the ith true componentwise accuracy under the column (question) cluster c respectively. $\hat{w}^{(c)}_i$ and $\hat{\theta}^{(c)}_i$ represent the estimated values of $w^{(c)}_{0;i}$ and $\theta^{(c)}_{0;i}$ using the posterior mean. Note that the penalty for misidentifying the true column (question) partition is added to

ADW, which matches the maximum difference between the estimated mixing weights and the true mixing weights. A similar penalty is also attached to ADP. Ideally, we expect CWRI to converge to 1 and the other three criteria to shrink towards 0 if Theorem 2 is true. The correct limiting values and the decreasing standard error successfully manifest our theoretical results, suggested by Table 1.

In the last two DGPs, the assumptions of the Rasch model are satisfied. We propose to study the performance of our model in identifying the true column (question) and row (student) partitions, defined as the labelling of ψ_j and ξ_i respectively, and compare the performance of estimating $\theta_{0;i}^{(d)} \equiv \frac{\exp\{\xi_i - \psi_j\}}{1+\exp\{\xi_i - \psi_j\}}$ using the following two criteria in addition to CWRI:

$$\begin{aligned} \text{ARWRI} &= \frac{1}{D}\sum_{d=1}^{D} \text{RI}(\hat{\mathscr{C}}^{\text{Row};d}, \mathscr{C}_0^{\text{Row};(c(d))}), \\ D_1 &= \frac{1}{nD}\sum_{i=1}^{n}\sum_{d=1}^{D}\left|\hat{\theta}_i^{(d)} - \theta_{0;i}^{(d)}\right| \end{aligned} \tag{14}$$

where ARWRI is the abbreviation of averaged rowwise Rand index and $\hat{\theta}_i^{(d)}$ can be directly provided by the Rasch model or using the posterior mean for our model. The results are presented in Table 2.

It is interesting to note that when n increases, CWRI increases towards 1 and ARWRI stays at a high value. Though ARWRI is not guaranteed to converge towards 1, our model is able to identify most of the correct labels for examinees when the latent accuracy parameters are sufficiently well separated. In addition, our model achieves a higher ARWRI when more questions are available under each question cluster, which matches our intuition. That is, more questions are more helpful in correctly distinguishing different types of students by comparing the results of DGP4 with those of DGP3. By comparing the D_1 (distance definition) values of our proposed model and the Rasch model, our model provides a more efficient estimate of the accuracy parameter $\theta_{0;i}^{(c)}$, especially when n and D are large (e.g., DGP4).

TABLE 1 Median (standard error) of the four criteria over 100 Monte Carlo replications for each of the first two DGPs given different sample sizes.

DGP	n	CWRI	ADK	ADW	ADP
1	100	1.000 (0.005)	0.000 (0.181)	0.082 (0.395)	0.064 (0.192)
	300	1.000 (0.000)	0.000 (0.158)	0.034 (0.015)	0.033 (0.014)
	1000	1.000 (0.000)	0.000 (0.097)	0.014 (0.006)	0.013 (0.007)
2	100	1.000 (0.001)	0.625 (0.273)	0.811 (0.338)	0.417 (0.166)
	300	1.000 (0.000)	0.000 (0.144)	0.029 (0.136)	0.023 (0.068)
	1000	1.000 (0.000)	0.000 (0.075)	0.016 (0.005)	0.013 (0.003)

TABLE 2 Median (standard error) of the three criteria over 100 Monte Carlo replications for each of the last two DGPs given different sample sizes.

		ACBM			Rasch
DGP	n	CWRI	ARWRI	D_1	D_1
3	100	0.474 (0.193)	0.922 (0.088)	0.093 (0.018)	0.073 (0.005)
	300	1.000 (0.048)	0.814 (0.020)	0.077 (0.015)	0.068 (0.003)
	1000	1.000 (0.000)	0.818 (0.009)	0.066 (0.005)	0.066 (0.002)
4	100	0.919 (0.174)	0.978 (0.016)	0.026 (0.022)	0.050 (0.003)
	300	1.000 (0.007)	0.979 (0.007)	0.012 (0.003)	0.043 (0.002)
	1000	1.000 (0.000)	0.977 (0.004)	0.009 (0.001)	0.040 (0.001)

6 | TEST DATA ANALYSIS

6.1 | Descriptive analysis

The data consist of the English exam results for the 2020–2021 academic year from No. 11 Middle School of Wuhan, Bingjiang Campus, which is a state middle school in Jiang'an district of Wuhan, China. This exam is a final English exam for Grade 8 students in the autumn semester of the 2020–2021 academic year. There are 16 classes with 858 students taking this exam. The data set consists of 858 examinees ($n = 858$) and 70 questions ($D = 70$), where the questions are from a single exam. The 70 questions fall into four major types (listening comprehension, multiple choice, Cloze test, and reading comprehension). We proceed by carrying out an exploratory data analysis. By looking at the estimated accuracy marginalized for each question (column) or each row (examinee), visualized on the left-hand side of Figure 1, it is obvious that the questions are designed hierarchically in terms of their difficulty, indicated by estimated accuracies ranging from .247 to .981. In addition, the proficiency of examinees is fairly heterogeneous, as the displayed histogram demonstrates a left-skewed feature with a long tail. To be more specific, the histogram implies that most examinees can solve more than 70% of the questions, while a small proportion of the examinees, whose estimated accuracy is below .4, may probably have failed the test. Such heterogeneity can also be viewed from the boxplots of the Rasch parameters, as shown on the right-hand side of Figure 1, where $\boldsymbol{\xi}$ and $\boldsymbol{\psi}$ are defined similarly to those in (12). As the primary goal of analysing this data set is to explain the heterogeneity, we next present the results by applying our proposed model.

6.2 | ACBM analysis

To apply our proposed model, we set the number of iterations to $n_{\text{iter}} = 400$, $n_{\text{rep}} = 400$, which are sufficient to thoroughly explore the posterior high-density region based on our simulation analyses. The hyperparameters are chosen as $a_0 = .01$ and $b_0 = .01$ to ensure non-informative prior knowledge, while new column and row clusters can still be generated. Given such settings, our model is implemented repeatedly 100 times with different initial values. The reported column (question) partition is believed to be representative as the median Rand index between it and the other column partitions is 0.91 with a standard deviation of 0.04 over the 100 Monte Carlo replications. The estimated accuracy parameters using posterior mean under each column cluster and the number of entries corresponding to each accuracy parameter are summarized in Table 3.

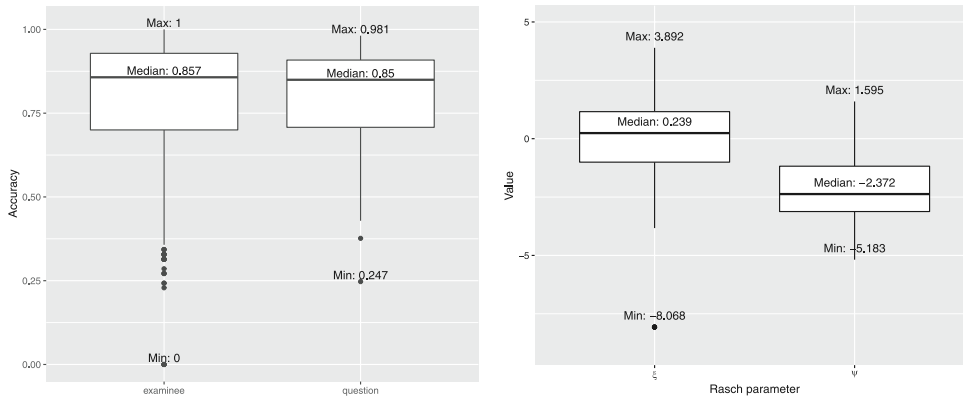

FIGURE 1 Left: boxplots of the estimated accuracy marginalized over each question or examinee; Right: boxplots of the Rasch parameters.

Following Table 3, a question cluster that contains more questions tends to possess more components. The estimation of most componentwise accuracy parameters is precise, since most estimated standard deviation values are one order of magnitude smaller than the corresponding estimated accuracy parameters. We further conjecture that the questions that are assigned to the clusters below the middle line in Table 3 are not effective in distinguishing different types of examinees, suggested by our model. Recall that the number of examinees' mixtures is bounded above by $(|\mathcal{C}|+1)/2$ to ensure model identifiability. It is hence impossible to identify more than one examinees' mixture when a question cluster has less than three questions. In other words, an ideal question cluster should consist of at least three questions to be able to detect the heterogeneity among examinees (referring to Lemma 1). Based on the test questions, it can be observed that the questions grouped under clusters 7, 8, and 9 require students to possess strong contextual comprehension skills. These questions are highly demanding and challenging as they assess students' ability to comprehend the entire article in an abstract manner. This is also supported by the low average accuracy suggested by the lower part of Table 3. On the other hand, the questions assigned to clusters 1, 2, and 6 comprise questions that are intended to assess the examinees' foundational knowledge, such as their proficiency in using various tenses, pronouns, and basic listening skills. The relatively lighter shade of the blocks corresponding to these clusters in Figure 2 also indicates this.

The estimated componentwise accuracy parameters can further be visualized using Figure 2, after rearranging the columns (questions) into a consecutive layout according to the estimated column (question) partition given by ACBM, for both ACBM and the Rasch estimations simultaneously.

TABLE 3 The number of components (K) and the estimated componentwise accuracy parameters under each question cluster and the corresponding cluster size ($|\mathcal{C}|$).

Cluster	Size ($|\mathcal{C}|$)	$K \leq (|\mathcal{C}|+1)/2$	Estimated accuracy
1	4	2	.443, .872
2	20	5	.001, .344, .645, .924, .999
3	5	2	.218, .671
4	17	4	.179, .419, .726, .937
5	12	4	.001, .417, .752, .989
6	8	3	.164, .853, .999
7	2	1	.411
8	1	1	.791
9	1	1	.247

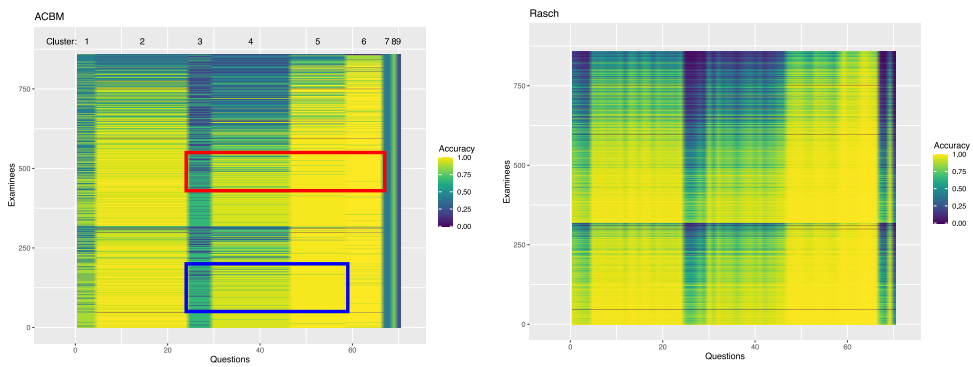

FIGURE 2 The estimated accuracy parameters aligned in a matrix after permuting the questions based on the estimated column (question) partition. Left: ACBM; Right: Rasch model.

Intuitively, the ACBM gradient plot looks like a discretized version of the Rasch gradient plot, which implies that our proposed model can recover the Rasch model's result to some extent. As an advantage over the Rasch model, our proposed model can automatically identify possible question clusters and the mixing structures on the examinees thereof. Note that the Guttman pattern is revealed locally if we look into the accuracy parameters of the corresponding examinees in question clusters 3 and 5. The questions in cluster 5 are listening comprehension and multiple-choice questions, which are in general easier compared to the questions assigned to question cluster 3, the majority of which are difficult reading comprehension questions. To provide more insights, we present a contingency table for clusters 3 and 5 in Table 4. For example, among the examinees who correctly answered questions from cluster 3 with a higher accuracy (.671), only one of them answered questions from cluster 5 with an accuracy being less than or equal to .417. In contrast, 469 (94.4%) of them answer correctly to the questions in cluster 5 with an accuracy of .988. This finding agrees with the prior belief that questions in cluster 5 are easier than those in cluster 3 and further indicates that our method can effectively capture the Guttman pattern locally based on specific question clusters.

We further discuss the heterogeneity as indicated by the red rectangle in Figure 2. Based on Table 5, for these examinees who are less proficient in question cluster 3 (accuracy = .218), 54.3% (25/46) of them did well in question cluster 4 with a .937 accuracy. On the other hand, for those who do well in question cluster 4 (accuracy $\geq$ .726), 37.5% (42/112) of them did not perform well in question cluster 3 (accuracy = .218). Such heterogeneity is not solely explained by randomness as we have a sufficiently large number of samples in estimating each accuracy parameter. Similar findings can also be discovered in the region formed by the blue rectangle in the same figure. It is gratifying to see that our method can capture such heterogeneity, whereas the Rasch model is unable to do so by using a single parameter to model the ability of examinees over all questions.

7 | DISCUSSION

In this paper we propose a novel IRT model using an averaged mixture of binomial distributions with constraints, the novelty of which basically comes from the modelling of heterogeneity and the justification of the identifiability issue. Our model is shown to be effective in both theoretical and practical aspects. Namely, the identifiability conclusion and posterior contraction results indicate that the latent accuracy parameters of interest to us can be estimated at a $\sqrt{n}$ (up to a log term) rate asymptotically. In addition, the posterior samples of these parameters can be obtained using a tractable sampling algorithm that satisfactorily approaches the stationary distributions according to the simulation results. Compared to the existing methods, including the Rasch model and multi-dimensional models, our

TABLE 4 Contingency table of examinees' count in terms of the accuracy parameters for clusters 3 and 5.

	C5 (Easy)			
C3 (Difficult)	**Acc = .001**	**Acc = .417**	**Acc = .753**	**Acc = .988**
Acc = .218	5	34	73	239
Acc = .671	0	1	27	469

TABLE 5 Contingency table of the count of the examinees in the red rectangle in terms of the accuracy parameters for cluster 3 and 4.

	C4		
C3	**Acc = .419**	**Acc = .726**	**Acc = .937**
Acc = .218	4	17	25
Acc = .671	3	14	56

model manages to reveal the multi-dimensionality of the examinees' proficiency level in handling different types of questions parsimoniously due to its discrete nature, thanks for the nested clustering structure. In fact, our proposed model is closely related to many existing multidimensional IRT models. Inspired by the statement in Reckase (2009, p. 79) – 'There are two major types of multidimensional item response models … One type of model is based on a linear combination of ψ-coordinates … The second type of model separates the cognitive tasks in a test item into parts and uses a unidimensional model for each part' – we reformulate our model and investigate its connection with these two majorities. Note that the conditional probability of $X_{i,j} = 1$ given by our model is

$$\begin{aligned}\Pr(X_{i,j} = 1 | \mathscr{C}, \theta_k^{(c(j))}, K^{(c(j))}, Z_i^{(c(j))}) &= \prod_{k=1}^{K^{(c(j))}} \left[\theta_k^{(c(j))}\right]^{1(Z_i^{(c(j))}=k)} \\ &= \prod_{k=1}^{K^{(c(j))}} \left[\frac{\exp\{\psi_k^{(c(j))}\}}{1+\exp\{\psi_k^{(c(j))}\}}\right]^{1(Z_i^{(c(j))}=k)},\end{aligned} \tag{15}$$

where $\psi_k^{(c(j))}$ is the natural parameter of $\theta_k^{(c(j))}$ and $\Pr(Z_i^{(c(j))} = k) = w_k^{(c)}$. The summation inside the product of the last display of (15) can be obtained by degenerating equation (4.5) of Reckase (2009),

$$\Pr(X_{i,j} = 1 | \boldsymbol{\psi}_i, \mathbf{a}, d_j) = \frac{\exp\{\sum_{\ell=1}^m a_\ell \psi_\ell + d_j\}}{1+\exp\{\sum_{\ell=1}^m a_\ell \psi_\ell + d_j\}}, \tag{16}$$

by letting $d_j = 0$ and $\sum_{\ell=1}^m a_\ell \psi_\ell$ be $\sum_{c \in \mathscr{C}} a_{c(j)} \psi_k^{(c(j))}$, where $a_{c(j)} = 1$ if $j \in c$ and 0 otherwise. Furthermore, if we take the product as a whole and revisit equation (4.20) of Reckase (2009),

$$\Pr(X_{i,j} = 1 | \boldsymbol{\psi}_i, \mathbf{d}_j) = \prod_{k=1}^m \frac{\exp\{\psi_{i,\ell} - d_{j,\ell}\}}{1+\exp\{\psi_{i,\ell} - d_{j,\ell}\}}, \tag{17}$$

we can also do the degeneration by specifying the parameters. The key difference between our method and the two alternatives above is that our method emphasizes the estimation of $\mathscr{C}$, $\psi_k^{(c(j))}$ and $w_k^{(c)}$ for all $c \in \mathscr{C}$ rather than the latent factors.

The main limitation of the current method is the lack of within-item dimensionality specification. Despite the ability to automatically identify the clustering structure at the question and examinee level, it is incapable of differentiating the types of proficiency exhibited by examinees while tackling certain question types. For example, our model cannot account for the linear combination of ψ-coordinates as defined in (17), which is also of research interest. To conduct such an analysis, one can certainly resort to a secondary multi-dimensional IRT analysis based on the clustering analysis results given by our model at the cost of losing the one-step integrity. Another limitation of our method is that it does not provide a direct inference of either the ability of examinees or the difficulty of questions, as is commonly achieved in most existing Rasch's models and multi-dimensional models. In contrast, our model centers on modelling the proficiency level of specific examinee subgroups correspoding to certain question clusters. Therefore, the choice of the model depends on the goal of the study, so that if the main objective is to make inferences about the abilities of examinees and the difficulties of questions, our model may not be the most suitable option.

For future study, one possible generalization is to consider the product of Bernoulli densities in place of the binomial density, such that the accuracy parameters of the questions assigned to a question cluster can be arranged in ascending order after a permutation. In other words, without loss of generality, suppose there exists a permutation $\sigma(\cdot)$ given a question cluster indexed by $1, 2, \ldots, D'$; we define the product of Bernoulli densities as

$$X_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}),\ p_{i,\sigma(1)} \leq p_{i,\sigma(2)} \leq \ldots \leq p_{i,\sigma(D')}, \quad \text{for } j = 1, \ldots, D', \tag{18}$$

where $\sigma(\cdot)$ is shared within the question cluster. We may call them ordered Bernoulli densities. We expect that such a gradient of the accuracy parameters can better explain the Guttman pattern than the kernel function currently used. The only concern of this structure is the identifiability of using this kernel density, which requires further investigation. One can directly apply our theoretical results if the product of Bernoulli densities is shown to be first-order identifiable under certain conditions. Another possible way of improving is to consider a more advanced sampling algorithm than ours, which is a typical application of Algorithm 1 proposed by Neal (2000). The method of split-merge sampling (Jain & Neal, 2004) or slice sampling (Neal, 2003) can be used to accelerate the procedure to approach the stationary distribution. Future simulation studies might also comprehensively examine the difference between our model and the alternatives that accommodate heterogeneity by introducing mixing structure at both the item and subject levels.

ACKNOWLEDGEMENTS

The authors would like to thank the editor, the associate editor, and the two reviewers for their valuable comments, which helped improve the presentation of this paper. We would also like to thank Shan Jiang for providing the English test data. Hu's research was partially supported by US NSF grants DMS-2210371 and SES-2243058.

CONFLICT OF INTEREST STATEMENT

All authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Guanyu Hu https://orcid.org/0000-0003-1410-1665

REFERENCES

Alexeev, N., Templin, J., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*, *48*(3), 313–332.

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory. In D. Andrich & I. Marais (Coords.), *Measuring in the educational, social and health sciences* (pp. 41–53). Springer.

Bartolucci, F., Farcomeni, A., & Scaccia, L. (2017). A nonparametric multidimensional latent class IRT model in a Bayesian framework. *Psychometrika*, *82*(4), 952–978.

Bell, R. C., Pattison, P. E., & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement*, *12*(1), 15–26.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, *4*, 201–218.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications [The calculus of probabilities and its applications]* (pp. 23–27). CNRS International Colloquia 13. Centre National de la Recherche Scientifique.

Ghosal, S., & Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference* (Vol. 44). Cambridge University Press.

Green, P. J., & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, *28*(2), 355–375.

Guha, A., Ho, N., & Nguyen, X. (2019). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *arXiv preprint*. arXiv:1901.05078.

Ho, N., & Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, *10*(1), 271–307.

Hu, G., Ma, Z., & Paek, I. (2020). A nonparametric Bayesian item response modeling approach for clustering items and individuals simultaneously. *arXiv preprint*. arXiv:2006.00105.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.
Jain, S., & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, *13*(1), 158–182.
Jang, Y., Kim, S.-H., & Cohen, A. S. (2018). The impact of multidimensionality on extraction of latent classes in mixture Rasch models. *Journal of Educational Measurement*, *55*(3), 403–420.
Keith, R. A. (1987). The functional independence measure: a new tool for rehabilitation. *Advances in Clinical Rehabilitation*, *2*, 6–18.
Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In *Multivariate and mixture distribution Rasch models* (pp. 329–346). Springer.
Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.
Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, *113*(521), 340–356.
Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.
Miyazaki, K., & Hoshino, T. (2009). A bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*(3), 375–393.
Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, *1992*(1), 1–30.
Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.
Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, *31*(3), 705–767.
Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. Carnegie Mellon University.
Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002*. Springer.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.
Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. ERIC.
Reckase, M. D. (2009). *Multidimensional item response theory models*. Springer.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.
Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the mmpi-2 psy-5 scales. *Journal of Personality Assessment*, *72*(2), 282–307.
Sen, S., Cohen, A. S., & Kim, S.-H. (2019). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement*, *43*(4), 272–289.
Shen, W., Tokdar, S. T., & Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, *100*(3), 623–640.
Teicher, H. (1963). Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *34*(4), 1265–1269.
Thomas, M. L. (2011). The value of item response theory in clinical assessment: a review. *Assessment*, *18*(3), 291–307.
Zumbo, B. D., Pope, G. A., Watson, J. E., & Hubley, A. M. (1997). An empirical test of Roskam's conjecture about the interpretation of an ICC parameter in personality inventories. *Educational and Psychological Measurement*, *57*(6), 963–969.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pan, T., Shen, W., Davis-Stober, C. P., & Hu, G. (2023). A Bayesian nonparametric approach for handling item and examinee heterogeneity in assessment data. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–16. https://doi.org/10.1111/bmsp.12322