

APPLICATION AND CASE STUDIES - ORIGINAL

# Bayesian Nonparametric Models for Multiple Raters: A General Statistical Framework

Giuseppe Mignemi<sup>1</sup> and Ioanna Manolopoulou<sup>2</sup>

<sup>1</sup>Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy; <sup>2</sup>Statistics Department, University College London, Bloomsbury, UK

**Corresponding author:** Giuseppe Mignemi; Email: [giuseppe.mignemi@unibocconi.it](mailto:giuseppe.mignemi@unibocconi.it)

(Received 28 October 2024; revised 30 June 2025; accepted 2 July 2025)

## Abstract

Rating procedure is crucial in many applied fields (e.g., educational, clinical, emergency). In these contexts, a rater (e.g., teacher, doctor) scores a subject (e.g., student, doctor) on a rating scale. Given raters' variability, several statistical methods have been proposed for assessing and improving the quality of ratings. The analysis and the estimate of the Intraclass Correlation Coefficient (ICC) are major concerns in such cases. As evidenced by the literature, ICC might differ across different subgroups of raters and might be affected by contextual factors and subject heterogeneity. Model estimation in the presence of heterogeneity has been one of the recent challenges in this research line. Consequently, several methods have been proposed to address this issue under a parametric multilevel modelling framework, in which strong distributional assumptions are made. We propose a more flexible model under the Bayesian nonparametric (BNP) framework, in which most of those assumptions are relaxed. By eliciting hierarchical discrete nonparametric priors, the model accommodates clusters among raters and subjects, naturally accounts for heterogeneity, and improves estimates' accuracy. We propose a general BNP heteroscedastic framework to analyze continuous and coarse rating data and possible latent differences among subjects and raters. The estimated densities are used to make inferences about the rating process and the quality of the ratings. By exploiting a stick-breaking representation of the discrete nonparametric priors, a general class of ICC indices might be derived for these models. Our method allows us to independently identify latent similarities between subjects and raters and can be applied in *precise education* to improve personalized teaching programs or interventions. Theoretical results about the ICC are provided together with computational strategies. Simulations and a real-world application are presented, and possible future directions are discussed.

**Keywords:** Bayesian hierarchical models; Bayesian mixture models; Bayesian nonparametric models; intraclass correlation coefficient; rating models

## 1. Introduction

Rating procedure is crucial in several applied scientific fields, such as educational assessment (Childs & Wooten, 2023; Chin et al., 2020), psychological and medical diagnoses (D'lima et al., 2024; Królikowska et al., 2023; Li et al., 2022), emergency rescue (Albrecht et al., 2024; Lo et al., 2021) or grant review process (Cao et al., 2010; Sattler et al., 2015). It implies that an observer, commonly called a rater (e.g., teacher, doctor), assesses some subject attribute or latent ability (e.g., student proficiency, patient severity) on a rating scale. Raters' variability might pose reliability concerns and uncertainty about the

quality of ratings (Bartoš & Martinková, 2024; Mignemi et al., 2024; Ten Hove et al., 2021). Several statistical methods have been proposed to address these issues, they aim to assess or improve the accuracy of ratings (Casabianca et al., 2015; Gwet, 2008; Martinková et al., 2023; McGraw & Wong, 1996; Nelson & Edwards, 2015). Multilevel modelling serves as a natural statistical framework for rating data since subjects are either nested within raters or crossed with them (Ten Hove et al., 2021). These models (e.g., one-way or two-way ANOVA, hierarchical linear or generalized linear models) decompose the total variance of observed ratings according to different sources of variability, i.e., subjects and raters (see Martinková & Hladká, 2023, chapter 4, for an overview). The observed rating is commonly broken down into different effects, for instance, the effect of the subject (i.e., true score, latent ability; Lord & Novick, 1968), the effect of the rater (i.e., rater's systematic bias) and a residual part (McGraw & Wong, 1996; Shrout & Fleiss, 1979). This allows us to jointly estimate the subject true score and the reliability of ratings, which is generally referred to as the proportion of total variance due to the subjects' variability (McGraw & Wong, 1996; Werts et al., 1974).

Several methods have been proposed to analyze rating data under the Item Response Theory (IRT) framework, such as the Generalized Many Facet Rasch Models (GMFRMs; Linacre, 1989; Uto et al., 2024; Uto & Ueno, 2020), the Hierarchical Raters Models (HRMs; DeCarlo et al., 2011; Molenaar et al., 2021; Nieto & Casabianca, 2019; Patz et al., 2002) or the Generalized Hierarchical Raters Models (GHRMs; Muckle & Karabatsos, 2009). These models jointly estimate the subject's latent ability, rater effects (e.g., systematic bias and reliability), and item features (i.e., difficulty, discrimination). They typically rely on the assumption that subjects' latent abilities are independent and identically distributed (i.i.d.) from a normal distribution. Other recent research lines concentrate on modeling and estimation issues in the presence of subjects' and raters' heterogeneity (Martinková et al., 2023; Mutz et al., 2012; Sattler et al., 2015; Ten Hove et al., 2022). These works model systematic differences among subjects or raters are to allow more accurate estimates and detailed information about the rating procedure. Individual subjects' or raters' characteristics may affect rating reliability, so that more flexible models result in separate reliability estimates (Martinková et al., 2023). Recent models have been proposed to address this issue under a parametric multilevel modeling framework (Erosheva et al., 2021; Martinková et al., 2023; Martinkova et al., 2018; Mutz et al., 2012) in which heterogeneity is addressed as a covariate-dependent difference among subjects and subject- and rater-specific effects are assumed to be i.i.d from a normal distribution. The normality assumption made under all the aforementioned models might be unrealistic under a highly heterogeneous scenario in which possible clusters among subjects or raters might be reasonably expected and the conditional density of the respective effects might be multimodal (Paganin et al., 2023; Verbeke & Lesaffre, 1996; Yang & Dunson, 2010). Such patterns have emerged from real data, showing that both the conditional densities of subjects' latent ability (e.g., Uto et al., 2024) and raters' systematic bias (e.g., Muckle & Karabatsos, 2009) might be multimodal and the normality assumption violated. In these cases, the data exhibit two levels of heterogeneity. The first, known as *individual* heterogeneity, captures the differences between individuals; the second, referred to as *population* heterogeneity, pertains to the differences between clusters. Although parametric mixture models might represent a suitable solution, the number of mixture components needs to be fixed. Models with different numbers of components have to be fitted and model selection techniques are required to identify the optimal number of clusters (Bartholomew et al., 2011).

### 1.1. Our contributions

Our proposal aims to overcome these restrictions under a Bayesian nonparametric (BNP) model, which naturally accommodates subgroups among students and raters and allows less restrictive distributional assumptions on the respective effects (Ferguson, 1973; Ghosal & van der Vaart, 2017; Hjort et al., 2010). BNP inference has led to new developments and advances during the last decades in psychometrics (Cremaschi et al., 2021; Karabatsos & Walker, 2009; Paganin et al., 2023; Roy et al., 2024; San Martín et al., 2011; Tang et al., 2017; Wang & Kingston, 2020; Yang & Dunson, 2010), but few contributions have been proposed for the analysis of rating data (DeYoreo & Kottas, 2018; Kottas et al., 2005; Mignemi et al.,

2024; Savitsky & Dalal, 2014). We provide a flexible statistical framework for rating models in which latent heterogeneity among subjects and raters is captured with the stochastic clustering induced by the Dirichlet Process Mixture (DPM) placed over their respective effects. Modelling subjects' and raters' effect parameters as an infinite mixture of some distribution family (e.g., Normal, Gamma) enables the model to account for possible multimodality without specifying the number of mixture components (De Iorio et al., 2023; Yang & Dunson, 2010). Although previous works have raised questions about the identifiability of the parameters in BNP IRT models San Martín et al. (2011), theoretical results by (Pan et al., 2024) have recently shown that BNP IRT models (e.g., 1PL) are identifiable.

Under the general case of a two-way design (McGraw & Wong, 1996), we specify a measurement model for the subject latent ability (e.g., student proficiency) in which the rater's systematic bias (i.e., severity) and reliability are consistently estimated. This makes our method more relevant for subject scoring purposes than the other BNP models proposed for the analysis of rating data (DeYoreo & Kottas, 2018; Kottas et al., 2005; Savitsky & Dalal, 2014). Our proposal may be suitable both for balanced (i.e., when all raters score each subject; Nelson & Edwards, 2010, 2015) and unbalanced designs (i.e., when a subset of raters scores each subject; Martinková et al., 2023; Ten Hove et al., 2022). Furthermore, we propose a Semiparametric model as a nested version of the BNP in which raters' effects are i.i.d. from a unimodal distribution. Very small rater sample sizes may not reasonably be considered representative of the overall rater population, making the semiparametric specification a potentially more suitable choice.

The advantages of the proposed method are manifold. First, it relies on more relaxed distributional assumptions for the subjects' and raters' effects, allowing for density estimation using mixtures (Escobar & West, 1994; Ghosal et al., 1999) and preventing model misspecification issues (Antonelli et al., 2016; Walker & Gutiérrez-Peña, 2007). As recently argued by Tang et al. (2017), BNP priors might be helpful in assessing the appropriateness of common parametric assumptions for psychometrics models and represent a solution under their violation (Antoniak, 1974; Ferguson, 1973). Second, it naturally enables independent clustering of subjects and raters, bringing more detailed information about their latent differences (De Iorio et al., 2023; Mignemi et al., 2024). This allows the joint analysis of *individual* and *population* heterogeneity of both subjects and raters. This aspect might be beneficial in the context of *precise education* (Coates, 2025; Cook et al., 2018), where information about individual and cluster differences might be used for implementing more personalized educational programs or interventions (Hart, 2016; Henderson et al., 2020). Third, exploiting a stick-breaking representation of the Dirichlet Process (DP) (Ghosal & van der Vaart, 2017; Ishwaran & James, 2001), a general class of ICC indices might be derived, and different indices might be computed according to distinct clusters of subjects or raters. Fourth, it is readily extended to account for coarse or ordinal ratings (Goel & Thakor, 2015; Lockwood et al., 2018). Fifth, the general hierarchical formulation of our model allows comparisons with other methods and further extensions under unifying modelling frameworks (e.g., generalized linear latent and mixed model, GLLMM Rabe-Hesketh & Skrondal, 2016). This facilitates a straightforward communication between different statistical fields and a wider application of the BNP method.

Model parameters are learned through full posterior sampling. Since most of the parameters in the model have conjugate prior distributions, full conditional Gibbs sampling is possible for most of the parameters (Ishwaran & James, 2001). Nonetheless, few parameters do not have conjugate priors and a derivatives matching technique is involved to approximate the full conditional (Miller, 2019).

## 1.2. Outline of the article

The outline of the article is as follows: we present the general framework and introduce the model in Sections 2.1–2.3, respectively; different approximate ICC indices are derived in Section 3 and a reduced model for one-way designs is detailed in Section 4; prior elicitation and posterior sampling are discussed and presented in Section 5; simulations and real-world applications are illustrated, respectively, in Section 6 and Section 7; the model extension for coarse ratings and is presented in Section 8, along with some numerical results from real and generated data. Advantages and limitations of the proposal are discussed in Section 9. Further BNP extensions, proofs for ICCs indices, and additional plots are

given in the Appendices. Additional results on balanced design in small sample sizes, technical details on out-of-sample predictive performance assessment and posterior computation for this class of models are presented in the Supplementary Material. We provide an R package *RatersBNP* to facilitate direct usage by researchers and practitioners of our method. Code and Supplementary Material are available online through the link: [https://osf.io/3yx4j/?view\\_only=98c600198a6b4807878989765118f97e](https://osf.io/3yx4j/?view_only=98c600198a6b4807878989765118f97e).

## 2. BNP rating model

### 2.1. General framework

Several model specifications have been proposed for different data structures and designs (Gwet, 2008; Shrout & Fleiss, 1979; Ten Hove et al., 2022). One-way designs are preferred when rater differences are typically considered as noise (Martinková et al., 2023), whereas two-way designs are usually involved if the rater's effect needs to be identified (Casabianca et al., 2015; Mignemi et al., 2024). Balanced designs require each subject to be rated by all the raters, while in an unbalanced design each subject is only rated by a generally small subset of them (Ten Hove et al., 2021). Raters might be considered either fixed or random (i.e., drawn from the population) depending on the inference the researcher might be interested in (Koo & Li, 2016).

The unbalanced two-way design with random raters is considered a general case to present our model. The reasons for this choice are both theoretical and practical. We aim to provide a comprehensive statistical framework for modeling the dependency of ratings on different categorical predictors (i.e., subjects' and raters' identities). This setting is a neat compromise between the one-way design, which implies only one categorical predictor (i.e., subject identity), and more complex dependency structures that involve more than two identities (i.e., several categorical predictors). However, an example of these extensions is given in Appendix A.2. Indeed, our proposal might be alternatively reduced or extended to be suitable for these different levels of complexity. The unbalanced design implies some sparsity in the co-occurrence between subjects and raters and each subject is rated only by a small subset of raters (Papaspiliopoulos et al., 2019; Papaspiliopoulos et al., 2023), as a consequence each rater might score a different number of subjects. This makes the framework general and flexible, it might be seen as an extension of cross-classified models in which uncertainty is modeled also hierarchically. From a practical perspective, our choice is reasonable since many large studies and applications use unbalanced designs to distribute the workload across different raters (Ten Hove et al., 2022).

### 2.2. Preliminaries on Bayesian nonparametric inference

In this subsection, we briefly review some basic preliminaries on BNP inference providing here a very general framework which is detailed in Sections below (refer to Ghosal & van der Vaart, 2017 and Hjort et al., 2010 for exhaustive treatments).

Suppose  $Y_1, \dots, Y_n$ , are observations (e.g., ratings), with each  $Y_i$  taking values in a complete and separable metric space  $\mathbb{Y}$ . Let  $\Pi$  denote a prior probability distribution on the set of all probability measures  $\mathbf{P}_{\mathbb{Y}}$  such that

$$Y_i | p \stackrel{\text{iid}}{\sim} p, \quad p \sim \Pi, \quad (1)$$

for  $i = 1, \dots, n$ . Here  $p$  is a random probability measure on  $\mathbb{Y}$  and  $\Pi$  is its probability distribution and might be interpreted as the prior distribution for Bayesian inference (De Blasi et al., 2015). The inferential problem is called parametric when  $\Pi$  degenerates on a finite-dimensional subspace of  $\mathbf{P}_{\mathbb{Y}}$ , and nonparametric when the support of  $\Pi$  is infinite-dimensional (Hjort et al., 2010, chapter 3). To the best of our knowledge, the vast majority of the contributions present in rating models literature (Bartoš & Martinková, 2024; Casabianca et al., 2015; Martinková et al., 2023; Martinkova et al., 2018; Nelson & Edwards, 2010, 2015; Ten Hove et al., 2021, 2022; Zupanc & Štrumbelj, 2018) are developed within a parametric framework making use of a prior that assigns probability one to a small subset



of  $\mathbf{P}_Y$ . Although Mignemi et al. (2024) recently proposed a Bayesian semi-parametric (BSP) model for analyzing rating data. Even if they relax the normality assumption for the rater effect (i.e., the systematic bias), normality is still assumed for the subject true score distribution. This strong prior assumption is overcome through a BNP approach (Ghosal & van der Vaart, 2017) in the present work.

*Dirichlet processes.* For the present proposal, we assume  $\Pi$  to be a discrete nonparametric prior and correspond to a DP which has been widely used in BNP psychometric research (Cremaschi et al., 2021; Karabatsos & Walker, 2009; Paganin et al., 2023; Yang & Dunson, 2010). Given  $\Pi = DP(\alpha P_0)$ ,  $p$  is a random measure on  $\mathbb{Y}$  following a DP with concentration parameter  $\alpha > 0$  and base measure  $P_0$ . This implies that for every finite measurable partition  $\{B_1, \dots, B_k\}$  of  $\mathbb{Y}$ , the joint distribution  $(p(B_1), \dots, p(B_k))$  follows a  $k$ -variate Dirichlet distribution with parameters  $\alpha P_0(B_1), \dots, \alpha P_0(B_k)$ :

$$(p(B_1), \dots, p(B_k)) \sim \text{Dir}(\alpha P_0(B_1), \dots, \alpha P_0(B_k)). \quad (2)$$

The base measure  $P_0$  is our *prior guess* at  $p$  as it is the prior expectation of the DP, i.e.,  $E[p] = P_0$ . The parameter  $\alpha$  (also termed precision parameter) controls the concentration of the prior for  $p$  about  $P_0$ . In the limit of  $\alpha \rightarrow \infty$ , the probability mass is spread out and  $p$  gets closer to  $P_0$ ; on the contrary, as  $\alpha \rightarrow 0$ ,  $p$  is less close to  $P_0$  and concentrates at a point mass.

*Dirichlet process mixtures.* Given the discrete nature of the DP, whenever  $\mathbb{Y} = \mathbb{R}$  it is not a reasonable prior for the real-valued random variable  $Y$ . Nonetheless, it might be involved in density estimation through hierarchical mixture modeling (Ghosal & van der Vaart, 2017). Let  $f(\cdot; \tilde{\theta})$  denote a probability density function for  $\tilde{\theta} \in \Theta \subseteq \mathbb{R}$ , we modify (1) such that for  $i = 1, \dots, n$ :

$$Y_i | \tilde{\theta}_i \stackrel{\text{iid}}{\sim} f(\cdot; \tilde{\theta}_i), \quad \tilde{\theta}_i | p \stackrel{\text{iid}}{\sim} p, \quad p \sim DP(\alpha P_0). \quad (3)$$

The realizations of the DP are almost surely (a.s.) discrete which implies a positive probability that  $\tilde{\theta}_i = \tilde{\theta}_{i'}$ , for  $i \neq i'$ . Indeed, a random sample  $(\tilde{\theta}_1, \dots, \tilde{\theta}_n)$  from  $p$  features  $1 \leq K_n \leq n$  different unique values  $(\tilde{\theta}_1^*, \dots, \tilde{\theta}_{K_n}^*)$  and leads to a random partition of  $\{1, \dots, n\}$  into  $K_n$  blocks such that  $\tilde{\theta}_i \in (\tilde{\theta}_1^*, \dots, \tilde{\theta}_{K_n}^*)$  for  $i = 1, \dots, n$ . This naturally induces a mixture distribution for the observations  $Y_1, \dots, Y_n$  with probability density:

$$f(Y) = \int f(Y; \tilde{\theta}) p(d\tilde{\theta}). \quad (4)$$

To provide some intuition, by using a DP as a prior for an unknown mixture distribution we mix parametric families nonparametrically (Gelman et al., 2014). This model specification introduced by Lo (1984) and termed DPM provides a BNP framework to model rating data.

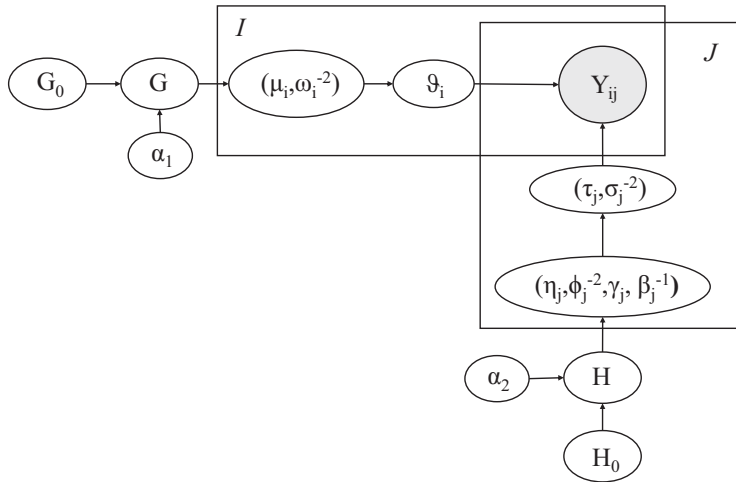
### 2.3. Proposed model

Consider a subject  $i = 1, \dots, I$ , whose attribute is independently scored by a finite random subset of raters  $\mathcal{R}_i \subseteq \{1, \dots, J\}$  on a continuous rating scale. We assume that the observed rating  $Y_{ij} \in \mathbb{R}$  depends independently on subject  $i$  and rater  $j \in \mathcal{R}_i$ . The effect of the former is interpreted as  $i$ 's true score and is the rating procedure's focus. We let the residual part, that is the difference between the true and the observed score, depend on rater  $j$ 's effects, i.e., systematic bias and reliability.

*Modelling rating  $Y_{ij}$ .* We specify the following decomposition of rating  $Y_{ij}$ :

$$Y_{ij} = \theta_i + \tau_j + \varepsilon_{ij}, \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i. \quad (5)$$

Here  $\theta_i$  captures the subject  $i$ 's latent "true" score and  $\tau_j + \varepsilon_{ij}$  is the difference between the observed and the true score, representing the error of rater  $j$ . The true score  $\theta_i$  might be regarded as the latent consensus across the raters  $\mathcal{R}_i$  on the attribute of subject  $i$ . We assume these terms to be mutually independent.



**Figure 1.** Graphical representation of the dependencies implied by the model. The boxes indicate replicates, the four outer plates represent, respectively, subjects and raters, and the inner grey plate indicates the observed rating.

*Modelling subject's true score.* For each subject  $i = 1, \dots, I$  we assume that the true score  $\theta_i$  is independently distributed following a normal distribution with mean  $\mu_i$  and variance  $\omega_i^2$ :

$$\theta_i | \mu_i, \omega_i^2 \stackrel{\text{ind}}{\sim} N(\mu_i, \omega_i^2). \quad (6)$$

Here  $\mu_i$  is the mean of subject  $i$ 's true score,  $\omega_i^2$  is its variability and we assume them to be independent. Conditional on the rater's error, higher values of  $\theta_i$  imply higher levels of the subjects' attribute (e.g., higher student proficiency); on the contrary lower values indicate poor levels of their attribute (e.g., poor student proficiency).

We specify a DP prior with precision parameter  $\alpha_1$  and base measure  $G_0$  for the pair  $(\mu_i, \omega_i^2)$ ,  $i = 1, \dots, I$ :

$$(\mu_i, 1/\omega_i^2) | G \stackrel{\text{iid}}{\sim} G, \quad G \sim DP(\alpha_1 G_0). \quad (7)$$

We choose  $G_0 = N(\mu_0, S_0) \times Ga(w_0, w_0/W_0)$ , where  $\mu_0$  and  $S_0$  are the mean and variance of the normal distribution and  $w_0$  and  $W_0$  are, respectively, the shape and the mean parameters of the gamma. We note that  $G$  is a.s. discrete with a non-zero probability of ties, such that different subjects will share the same values of  $(\mu_i, 1/\omega_i^2)$  with a probability greater than zero, that is  $P[(\mu_i, 1/\omega_i^2) = (\mu_{i'}, 1/\omega_{i'}^2)] > 0$ , for  $i \neq i'$ . This discreteness property naturally induces clustering across subjects and leads to a location-scale DPM prior for  $\theta_i$ . That is, this formulation can capture clusters of subject abilities. Figure 1 shows the hierarchical dependence of subjects' true scores.

*Modelling rater's bias and reliability.* For each rater  $j = 1, \dots, J$ , who scores a subset of subjects  $S_j \subseteq \{1, \dots, I\} : j \in \mathcal{R}_i$ , the difference between the observed rating  $Y_{ij}$  and the subject's true score  $\theta_i$ ,  $i \in S_j$ , is decomposed into the rater effects  $\tau_j$  and  $\varepsilon_{ij}$  (5), assuming  $\tau_j \perp \varepsilon_{ij}$ . We model  $\tau_j$  to be normally distributed with mean  $\eta_j$  and variance  $\omega_j^2$ :

$$\tau_j | \eta_j, \phi_j^2 \stackrel{\text{iid}}{\sim} N(\eta_j, \phi_j^2), \quad j = 1, \dots, J. \quad (8)$$

Here  $\eta_j$  and  $\phi_j^2$  are the mean and the variance of the rater  $j$ 's effect  $\tau_j$ . It captures  $j$ 's specific systematic bias, i.e., the mean difference between the observed rating  $Y_{ij}$  and the subject's true score  $\theta_i$ ,  $i \in S_j$ . Given two raters such that  $\tau_j < \tau_{j'}$ ,  $j$  is said to be more strict and expected to give systematically smaller ratings than  $j$  on average.

The residual term  $\varepsilon_{ij}$  is assumed to be i.i.d. for  $i \in \mathcal{S}_j$  following a normal distribution with zero mean and variance  $\sigma_j^2$ . We let this parameter vary across raters and assume  $1/\sigma_j^2$  follows a gamma distribution with shape and rate parameters  $\gamma_j, \gamma_j/\beta_j$ , respectively:

$$\varepsilon_{ij} | \sigma_j^2 \stackrel{\text{iid}}{\sim} N(0, \sigma_j^2), \quad i \in \mathcal{S}_j, \quad (9)$$

$$1/\sigma_j^2 | \gamma_j, \beta_j \stackrel{\text{ind}}{\sim} Ga(\gamma_j, \gamma_j/\beta_j), \quad j = 1, \dots, J. \quad (10)$$

Under this parametrization,  $1/\sigma_j^2$  is the rater  $j$ 's specific reliability with mean  $\beta_j$  and  $\gamma_j$  is the shape parameter. We prefer this parametrization for interpretability purposes, which implies a simpler notation below. Conditional on subjects' true score  $\theta_i$ ,  $i \in \mathcal{S}_j$ , larger values of  $\sigma_j^2$  imply more variability across the ratings given by  $j$  and might be interpreted as a poorly consistent rating behaviour. On the contrary, smaller values of  $\sigma_j^2$  indicate less variability and higher consistency for  $j$  across subjects. As a consequence, the parameter  $1/\sigma_j^2$  might be equivalently referred to as the rater-specific precision.

We specify a DP prior with concentration parameter  $\alpha_2$  and base measure  $H_0$  for the four-dimensional vector  $(\eta_j, 1/\phi_j^2, \gamma_j, 1/\beta_j)$ ,  $j = 1, \dots, J$ :

$$(\eta_j, 1/\phi_j^2, \gamma_j, 1/\beta_j) | H \stackrel{\text{iid}}{\sim} H, \quad H \sim DP(\alpha_2 H_0). \quad (11)$$

We assume mutual independence for the elements of the vector and choose  $H_0 = N(\eta_0, D_0) \times Ga(a_0, a_0/A_0) \times Ga(b_0, b_0/B_0) \times Ga(m_0, m_0/M_0)$ , where  $\eta_0$  and  $D_0$  are mean and scale parameters, respectively;  $a_0, b_0, m_0$  are shape parameters and  $A_0, B_0, M_0$  are mean parameters. This formulation induces a DPM prior for raters' bias and reliability  $\tau_j$  and  $1/\sigma_j^2$ . Figure 1 gives a graphical representation of the model. The independence assumption might be relaxed by employing a suitable multivariate base measure accounting for possible dependencies among the four elements of the vector. However, this implies a more complex specification, which is beyond the purpose of this work. Further constraints on raters' systematic bias  $\tau$  are needed for identifiability purposes which are discussed in Section 2.5, after presenting the stick-breaking representation.

#### 2.4. Stick-breaking representation

The random probability measures  $G$  and  $H$  are assigned discrete priors, as a consequence they might be represented as a weighted sum of point masses:

$$G = \sum_{n \geq 1} \pi_{1n} \delta_{\xi_n}, \quad (12)$$

$$H = \sum_{k \geq 1} \pi_{2k} \delta_{\zeta_k}, \quad (13)$$

where the weights  $\{\pi_{1n}\}_{n=1}^\infty$  and  $\{\pi_{2k}\}_{k=1}^\infty$  take values on the infinite probability simplex and  $\delta_x(\cdot)$  stands for the Dirac measure and denotes a point mass at  $x$ . Note that, we index the components of the infinite mixture (12) corresponding to the subjects with  $n = 1, \dots, \infty$ , whereas  $k = 1, \dots, \infty$  is used for that corresponding to the raters (13). The random vectors  $\xi_n = (\mu_n, \omega_n^2)$ ,  $n = 1, \dots, \infty$  are i.i.d. from the base measure  $G_0$ ,  $\zeta_k = (\eta_k, \phi_k^2, \gamma_k, \beta_k)$ ,  $k = 1, \dots, \infty$  are i.i.d. from the base measure  $H_0$ , and both vectors are assumed to be independent of the corresponding weights. This makes clear why the expectations of the DPs are  $G_0$  and  $H_0$ , respectively, and are said to be our *prior guess* at  $G$  and  $H$  (see Section 2.2).

This discreteness property of the DP allows us to define  $G$  and  $H$  through the stick-breaking representation introduced by Sethuraman (1994):

$$G = \sum_{n \geq 1} \pi_{1n} \delta_{\xi_n}, \quad \pi_{1n} = V_{1n} \prod_{l < n} (1 - V_{1l}), \quad V_{1n} \stackrel{\text{iid}}{\sim} \text{Beta}(1, \alpha_1), \quad \xi_n \stackrel{\text{iid}}{\sim} G_0, \quad (14)$$

and

$$H = \sum_{k \geq 1} \pi_{2k} \delta_{\zeta_k}, \quad \pi_{2k} = V_{2k} \prod_{l < k} (1 - V_{2l}), \quad V_{2k} \stackrel{\text{iid}}{\sim} \text{Beta}(1, \alpha_2), \quad \zeta_k \stackrel{\text{iid}}{\sim} H_0. \quad (15)$$

This construction of the *DP* implies that, for each subject  $i = 1, \dots, I$ ,  $(\mu_i, \omega_i^2) = \xi_n$  with probability  $\pi_{1n} = V_{1n} \prod_{l < n} (1 - V_{1l})$ . Equivalently, for each rater  $j = 1, \dots, J$ , the probability that  $(\eta_j, \phi_j^2, \gamma_j, \beta_j) = \zeta_k$  is given by  $\pi_{2k} = V_{2k} \prod_{l < k} (1 - V_{2l})$ .

*Moments of student latent true score  $\theta_i$ .* The mean and the variance of the subject's true score  $\theta_i$ ,  $i = 1, \dots, I$ , under a  $DP(\alpha_1 G_0)$  prior are:

$$\mathbf{E}[\theta_i|G] = \mu_G = \sum_{n \geq 1} \pi_{1n} \mu_n, \quad \mathbf{Var}[\theta_i|G] = \omega_G^2 = \sum_{n \geq 1} \pi_{1n} (\mu_n^2 + \omega_n^2) - \mu_G^2, \quad (16)$$

where  $\mu_n$  and  $\omega_n^2$  are the mean and the variance of  $\theta_i$  for the  $n$ th component of the mixture. Here  $\mu_G$  is the weighted average across components and captures the mean true score across subjects. The parameter  $\omega_G^2$  is the conditional variance of the infinite mixture and indicates the variability of true scores across subjects.

*Moments of raters' bias  $\tau_j$ .* The mean and the variance of the rater's bias  $\tau_j$ ,  $j = 1, \dots, J$ , under a  $DP(\alpha_2 H_0)$  prior are:

$$\mathbf{E}[\tau_j|H] = \eta_H = \sum_{k \geq 1} \pi_{2k} \eta_k, \quad \mathbf{Var}[\tau_j|H] = \phi_H^2 = \sum_{k \geq 1} \pi_{2k} (\eta_k^2 + \phi_k^2) - \eta_H^2, \quad (17)$$

where  $\eta_k$  and  $\phi_k^2$  are the mean and the variance of  $\tau_j$  for the  $k$ th component of the mixture. Here  $\eta_H$  and  $\phi_H^2$  capture the mean and the variance of the systematic bias within the general population of raters.

*Moments of raters' reliability  $1/\sigma_j^2$ .* Raters' residual mean is fixed to zero by the model (9), that is  $\mathbf{E}[\varepsilon] = 0$ ; mean and variance of raters reliability  $1/\sigma_j^2$  under a  $DP(\alpha_2 H_0)$  prior are:

$$\mathbf{E}[1/\sigma_j^2|H] = \beta_H = \sum_{k \geq 1} \pi_{2k} \beta_k, \quad \mathbf{Var}[1/\sigma_j^2|H] = \psi_H^2 = \sum_{k \geq 1} \pi_{2k} (\beta_k^2 + \psi_k) - \beta_H^2, \quad (18)$$

where  $\beta_H$  captures raters' weighted average reliability and  $\psi_H^2$  indicates the total reliability variance across them. Here  $\beta_k$  and  $\psi_k = \beta_k^2/\gamma_k$  are, respectively, the mean and the variance of  $1/\sigma_j^2$  for the  $k$ th component of the mixture.

Note that we model the independent rater's features, i.e., bias and reliability, by placing the same  $DP(\alpha_2 H_0)$  prior. In other terms,  $\tau_j$  and  $1/\sigma_j$  are two independent elements of the same vector drawn from  $H$ .

*Finite stick-breaking approximation.* The recursive generation defined in (14) and (15) implies a decreasing stochastic order of the weights  $\{\pi_{1n}\}_{n=1}^\infty$  and  $\{\pi_{2k}\}_{k=1}^\infty$  as the indices  $n$  and  $k$  grow. Considering the expectations  $\mathbf{E}[V_{1n}] = 1/(1 + \alpha_1)$  and  $\mathbf{E}[V_{2k}] = 1/(1 + \alpha_2)$  it is clear that the rates of decreasing depend on the concentration parameters  $\alpha_1$  and  $\alpha_2$ , respectively. Values of these parameters close to zero imply a mass concentration on the first couple of atoms, with the remaining atoms being assigned small probabilities; which is consistent with the general formulation of the *DP* discussed in Section 2.2. Given this property of the weights, in practical applications the infinite sequences (12) and (13), are truncated at enough large values of  $R \in \mathbb{N}$ :

$$G = \sum_{n=1}^R \pi_{1n} \delta_{\xi_n}, \quad H = \sum_{k=1}^R \pi_{2k} \delta_{\zeta_k}. \quad (19)$$

We use this finite stick-breaking approximation proposed by Ishwaran & James (2001) to let  $V_{1R} = V_{2R} = 1$ , and discard the terms  $R + 1, \dots, \infty$ , for  $G$  and  $H$ .

The moment formulas (16), (17), and (18) are readily modified accordingly to the truncation and computed as finite mixture moments.

*Nested versions.* Semiparametric nested versions of the BNP model might be specified in which alternatively  $G$  or  $H$  are degenerate on a single component and  $R = 1$  for one of them in the finite approximation. That is, subjects or raters are all clustered together. For instance, for very small values

of  $J$  (i.e., raters' sample size), raters might not be reasonably considered a representative sample of their population and limited information is available for drawing inference about it. Under these scenarios, raters' effects might be assumed to be i.i.d. from a normal distribution.

### 2.5. Semi-centered DPM

Hierarchical models (e.g., GLMM, Linear Latent Factor models), might suffer from identifiability issues, and constraints on the latent variable distributions are needed for consistently identify and interpret model parameters (Bartholomew et al., 2011; Gelman & Hill, 2006; Yang & Dunson, 2010). More specifically, under the linear random effects models a standard procedure to achieve model identifiability is to constrain the mean of the random effects to be zero (Agresti, 2015). We aim to consistently involve the same mean constraint for our proposal and allow straightforward and interpretable comparisons between the parametric and the nonparametric models. Similar to Yang & Dunson (2010), we encompass a DPM-centered prior such that the expected value of the rater systematic bias is fixed to zero,  $E[\tau_j] = 0$ , for  $j = 1, \dots, J$ .

Since the rating process focuses on the subjects' true scores, it might be more reasonable to centre the DPM for the raters' effects and let the model estimate the mean of the true scores  $\mu_G$ . Given that the mean of the raters' residual is fixed to zero in (9), the mean raters' bias needs to be fixed. We adapt the centering procedure based on a parameter-expanded approach proposed by Yang et al., 2010 and Yang & Dunson, 2010 to our proposal. We specify a semi-centered DPM (SC-DPM) involving an expansion in raters' systematic bias  $\{\tau_j^*\}_1^J$ , such that their mean  $\eta_H^* = 0$  a.s. The expanded-parameters (8) can be expressed as

$$\tau_j^* = \tau_j - \eta_H, \quad \tau_j | \eta_j, \phi_j^2 \stackrel{\text{ind}}{\sim} N(\eta_j, \phi_j^2), \quad j = 1, \dots, J, \quad (20)$$

and the decomposition of rating  $Y_{ij}$  (5) becomes

$$Y_{ij} = \theta_i + \tau_j^* + \varepsilon_{ij}, \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i. \quad (21)$$

Given the location transformation in (20) the expectation of the expanded parameters is zero:

$$E[\tau_j^* | H] = 0. \quad (22)$$

It is worth noting that the centering needs only to concern the location of the systematic bias and not its scale as it is in the centered-DPM introduced by Yang & Dunson (2010), which explains the term “semi-centring” adopted here to avoid confusion. Accordingly, under the semiparametric specifications, the only location of the parametric distribution needs to be fixed; a zero mean normal distribution might be a suitable solution.

### 3. BNP intra-class correlation coefficient

Intra-class correlation coefficient (ICC) is widely used in applied statistics to quantify the degree of association between nested observations (Agresti, 2015; Gelman et al., 2014) and to get relevant information about the level of heterogeneity across different groups (Mulder & Fox, 2019). Indeed, it is commonly applied in psychometrics to assess the consistency of ratings given by different raters to the same subject (Erosheva et al., 2021; Martinková et al., 2023; Nelson & Edwards, 2010, 2015; Ten Hove et al., 2021, 2022). We provide a within-subject correlation structure (for any subject and a given raters pair)  $ICC_{j,j'}$  based on the BNP model presented in Section 2.3. This formulation relates to those proposed in psychometric literature regarding the  $ICC_1$  (e.g., Bradlow et al., 1999; De Boeck, 2008; Erosheva et al., 2021; Fox & Glas, 2001; Shrout & Fleiss, 1979; Werts et al., 1974), but doesn't rely on strong distributional assumptions and naturally accommodates for both subjects and raters sub-populations. We also propose a lower bound  $ICC_A$  for the expected ICC which might be used for inference purposes about the general population of raters. An exact formula for the ICC suitable for

the reduced one-way designs is proposed in Section 4.1. We propose a BNP class of ICCs which are a generalization of those recently proposed by (Lin et al., 2025) and Ten Hove et al. (2021) under the generalizability theory.

The paragraphs below provide preliminary information on computing the ICC under a parametric framework necessary to detail the BNP extension.

*Parametric ICC.* Under a parametric standard framework, i.e., equipping the parameters with finite-dimensional priors, the ICC is defined as the proportion of variance of the ratings due to the subjects' true score:

$$ICC = \frac{\omega_i^2}{\omega_i^2 + \phi_j^2 + \sigma_j^2} = \frac{\omega^2}{\omega^2 + \phi^2 + \sigma^2}, \quad (23)$$

assuming  $\omega_i^2 = \omega^2$ , for  $i = 1, \dots, I$ ;  $\phi_j^2 = \phi^2$  and  $\sigma_j^2 = \sigma^2$ , for  $j = 1, \dots, J$ . Given two raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$  who rate the same subject  $i$ , the ICC is the correlation between the ratings  $Y_{ij}$  and  $Y_{ij'}$ . Note that under this formulation  $ICC \in [0, 1]$ , it can not capture any negative correlations. This index is also interpreted as the inter-rater reliability of a single rating and is also indicated by  $IRR_1$  (see Erosheva et al., 2021 for further details). The homoscedastic assumption may be relaxed and raters' residual variance might be let to vary across raters according to (9) and (10), given  $\gamma_j = \gamma$  and  $\beta_j = \beta$  for  $j = 1, \dots, J$ .

Given that  $\sigma_j^2 \neq \sigma_{j'}^2$  for  $j \neq j'$ , it is possible to compute as many ICCs indices as possible pairs of raters, i.e.,  $J(J-1)/2$ . In such cases the resulting  $ICC_{j,j'}$  is the conditional correlation between the ratings given to a random subject by raters  $j$  and  $j'$ , given the other parameters:

$$ICC_{j,j'} = \frac{\omega^2}{\sqrt{\omega^2 + \phi^2 + \sigma_j^2} \sqrt{\omega^2 + \phi^2 + \sigma_{j'}^2}}. \quad (24)$$

A more general index accounting for all raters' residual variance might be more useful in applications. Despite the expected ICC, i.e.,  $\mathbf{E}[ICC|\omega^2, \phi^2]$ , might represent a neat solution, it is not available in a close form and the posterior mean taken over the MCMC might be prohibitive in large scale assessments since there are  $J(J-1)/2$  ICCs indices to compute for each iteration. An alternative index that might be readily computed is the ICC between two raters with average reliability. That is, we replace  $\sigma_j^2$  with its expectation, i.e.,  $\mathbf{E}[\sigma^2]$ :

$$ICC_A = \frac{\omega^2}{\omega^2 + \phi^2 + \mathbf{E}[\sigma^2]}. \quad (25)$$

It gives the correlation between the ratings given to the same random subject  $i = 1, \dots, I$  by two random raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , satisfying  $\sigma_j^2 = \sigma_{j'}^2 = \mathbf{E}[\sigma^2]$ . That is the correlation between two ratings given to the same random student by two raters having an average reliability level. We note that they are different quantities: the expected pairwise ICC and the pairwise ICC between two mean reliable raters. Nonetheless, relying on a theoretical result that is given below, we can use the  $ICC_A$  to have information about the other.

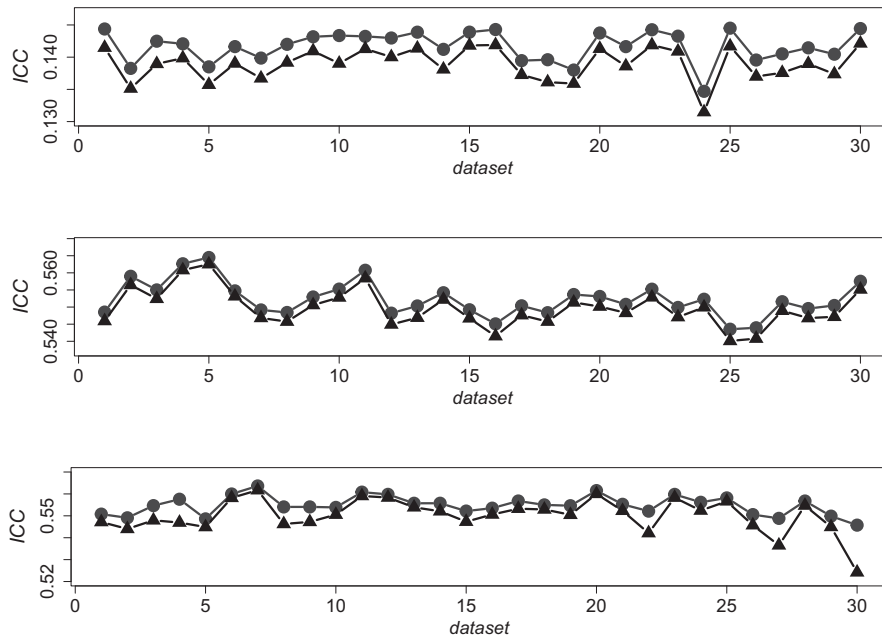
Given that the rater's reliability is assumed to follow a gamma distribution (9), the inverse follows an inverse gamma distribution  $\sigma_j^2|\gamma, \beta \stackrel{\text{ind}}{\sim} IGa(\gamma, \gamma/\beta)$  for  $j = 1, \dots, J$ , whose expected value is only defined for  $\gamma > 1$ . In such cases we reparametrize (9):

$$1/\sigma_j^2|\gamma, \beta \stackrel{\text{iid}}{\sim} Ga\left(1 + \gamma, \frac{1 + \gamma}{\beta}\right), \quad j = 1, \dots, J. \quad (26)$$

This specification ensures the expectation of raters' residual variance to be defined for any  $\gamma > 0$  and implies

$$\mathbf{E}[\sigma_j^2|\gamma, \beta] = \tilde{\sigma} = \frac{1 + \gamma}{\beta\gamma}. \quad (27)$$





**Figure 2.** Illustrative examples of empirical  $ICC_A$  and  $E[ICC]$  across independent datasets and under different reliability scenarios. The grey balls indicate the mean pairwise ICC between each rater and the others; the black triangles represent the computed  $ICC_A$ .

It is the mean raters' residual variance and its derivation is given in the Supplementary Material. The  $ICC_A$  under the new parametrization is

$$ICC_A = \frac{\omega^2}{\omega^2 + \phi^2 + \tilde{\sigma}}. \quad (28)$$

Figure 2 shows the difference between the empirical mean pairwise ICC between each rater (red solid line) and the others and the computed  $ICC_A$  (blue solid line) across independent datasets and different reliability scenarios. The mean difference between these two indices is consistently tight, and it seems to be narrower at increasing reliability levels.

**BNP ICC.** The moments defined in (16), (17), and (18) account for heterogeneous populations of subjects and raters and can be used to compute a flexible ICC.

**Proposition 1.** Given a random subject  $i = 1, \dots, I$ , independently scored by two random raters  $j, j' \in \mathcal{R}_i, j \neq j'$ , the conditional correlation between the scores  $Y_{ij}$  and  $Y_{ij'}$  is

$$ICC_{j,j'} = \text{Corr}(Y_{ij}, Y_{ij'} | G, H, \sigma_j^2, \sigma_{j'}^2) = \frac{\omega_G^2}{\sqrt{\omega_G^2 + \phi_H^2 + \sigma_j^2} \sqrt{\omega_G^2 + \phi_H^2 + \sigma_{j'}^2}}. \quad (29)$$

The proof is reported in Appendix B. However, a more general index, unconditioned on specific raters' parameters, might be more useful in practice. For this reason, we propose a  $ICC_A$  index for this BNP class of models. To this aim, the variance of subjects' true score  $\omega_G^2$  and the variance of raters' systematic bias  $\phi_H^2$  can be directly plugged into the ICC formula. Since we have heteroscedasticity across raters, we need to take the expectation of raters' residual variance  $E[\sigma^2 | H] = \tilde{\sigma}_H^*$ . Similarly to the above parametric case, we reparametrize (10) with:

$$1/\sigma_j^2 | \gamma, \beta \overset{\text{ind}}{\sim} \text{Ga}\left(1 + \gamma_j, \frac{1 + \gamma_j}{\beta_j}\right), \quad j = 1, \dots, J, \quad (30)$$

and define:

$$\mathbf{E}[\sigma_j^2|G, H] = \mathbf{E}[\sigma_j^2|H] = \tilde{\sigma}_H = \sum_{k \geq 1} \pi_{2k} \tilde{\sigma}_k, \quad (31)$$

where  $\tilde{\sigma}_k = (1 + \gamma_k)/(\beta_k \gamma_k)$  is the mean residual variance for the  $k$ th component of the infinite mixture. As a result, the  $ICC_A$  for the BNP models might be computed as reported below.

**Proposition 2.** *Given a random subject  $i = 1, \dots, I$ , independently scored by two random raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , satisfying  $\sigma_j^2 = \sigma_{j'}^2 = \tilde{\sigma}_H$ :*

(i) *the conditional correlation between the ratings  $Y_{ij}$  and  $Y_{ij'}$  is*

$$ICC_A = \text{Corr}(Y_{ij}, Y_{ij'}|G, H, \sigma_j^2 = \sigma_{j'}^2 = \tilde{\sigma}_H) = \frac{\omega_G^2}{\omega_G^2 + \phi_H^2 + \tilde{\sigma}_H}; \quad (32)$$

(ii) *the  $ICC_A$  is the lower bound of the conditional expectation of the correlation between the ratings  $Y_{ij}$  and  $Y_{ij'}$  (ICC):*

$$ICC_A \leq \mathbf{E}[\text{Corr}(Y_{ij}, Y_{ij'}|G, H)] = \mathbf{E}[ICC|G, H]. \quad (33)$$

The proofs are reported in Appendix B. The index therefore accounts for the heterogeneity of the two populations (subjects and raters). It reduces to the parametric  $ICC_A$  (23) whenever  $\omega_n^2 = \omega^2$ , for  $n = 1, \dots, \infty$ ;  $\phi_k^2 = \phi^2$  and  $\tilde{\sigma}_k = \tilde{\sigma}$ , for  $k = 1, \dots, \infty$ ;  $ICC_A$  (32) is a generalization of its parametric version (23). The  $ICC_A$  might reveal valuable information in inter-rater reliability or agreement analysis. For instance, when the ICC is used as an inter-rater reliability index (Erosheva et al., 2021; Martinková et al., 2023; Ten Hove et al., 2022), the  $ICC_A$  is the lower bound of the expected inter-rater reliability of a single rating.

In this work, we mainly focus on the population level  $ICC_A$ , but different ICC indices can be computed and compared under this framework by conditioning on different subjects or raters' clusters.

We note that this class of ICCs generalizes those proposed by Lin et al. (2025) for multilevel data. As a consequence, propositions (1) and (2) hold for standard parametric multilevel models (Ten Hove et al., 2021, 2022). The BNP ICCs we propose account for heteroskedasticity across raters and naturally accommodate multiple clusters of both subjects and raters, whereas the standard ICCs commonly do not encompass these heterogeneous aspects. More details on these aspects are given in Appendix A.2.

#### 4. Reduced model for one-way designs

One-way designs are common when raters' identity is unknown and the systematic biases  $\{\tau_j\}_1^J$  can not be identifiable. It might be seen as a limiting case in which each rater only scores one subject, i.e.,  $|\mathcal{S}_j| = 1$ .

Some blocks of the model in Section 2.3 reduce as briefly presented below. Note that we model subjects' true score  $\theta_i$  as in the main model (6) and (7).

*Modelling rating  $Y_{ij}$ .* We decompose the observed rating  $Y_i$  as

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i, \quad (34)$$

Here  $\varepsilon_{ij}$  is the error of rater  $j$  in rating the subject  $i$  and it is the difference between the observed score  $Y_{ij}$  and the subject true score  $\theta_i$ .

*Modelling raters' errors  $\varepsilon_n$*  For each rating  $Y_{ij}$  we assume that the rater's error  $\varepsilon_{ij}$  is drawn independently from a normal distribution with mean  $\eta_{ij}$  and variance  $\phi_{ij}^2$ :

$$\varepsilon_{ij}|\eta_{ij}, \phi_{ij}^2 \stackrel{\text{ind}}{\sim} N(\eta_{ij}, \phi_{ij}^2), \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i. \quad (35)$$

We specify a DP prior with concentration parameter  $\alpha_2$  and base measure  $H_0$  for the two-dimensional vector  $(\eta_{ij}, \phi_{ij}^2)$ , for  $i = 1, \dots, I$  and  $j \in \mathcal{R}_i$ :

$$(\eta_{ij}, \phi_{ij}^2) | H \stackrel{\text{iid}}{\sim} H, \quad H \sim DP(\alpha_2 H_0). \quad (36)$$

We assume  $\eta_{ij}, \phi_{ij}^2$  to be independent and choose  $H_0 = N(\eta_0, D_0) \times IGA(a_0, A_0)$ , where  $\eta_0$  and  $D_0$  are mean and scale parameters, respectively. This formulation induces a DPM prior for raters' error  $\varepsilon_{ij}$ .

#### 4.1. Identifiability and ICC

The moments of the error  $\varepsilon_{ij}$ ,  $i = 1, \dots, I$  and  $j \in \mathcal{R}_i$ , are, respectively:

$$\mathbf{E}[\varepsilon_{ij} | H] = \eta_H = \sum_{k \geq 1} \pi_{2k} \eta_k, \quad \mathbf{Var}[\varepsilon_{ij} | H] = \phi_H^2 = \sum_{k \geq 1} \pi_{2k} (\eta_k^2 + \phi_k^2) - \eta_H^2. \quad (37)$$

The centering strategy detailed in Section 2.5 is here used and a SC-DPM is here placed over  $\varepsilon_{ij}$ :

$$\varepsilon_{ij}^* = \varepsilon_{ij} - \eta_H, \quad \varepsilon_{ij} | \eta_{ij}, \phi_{ij}^2 \stackrel{\text{iid}}{\sim} N(\eta_{ij}, \phi_{ij}^2), \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i. \quad (38)$$

Under this parameter-expanded specification, the decomposition of rating  $Y_{ij}$  (34) becomes

$$Y_{ij} = \theta_i + \varepsilon_{ij}^*, \quad i = 1, \dots, I; \quad j \in \mathcal{R}_i. \quad (39)$$

Given the location transformation in (38), the expectation of the residuals is zero:

$$\mathbf{E}[\varepsilon_{ij}^* | H] = 0. \quad (40)$$

For the one-way designs, the exact general ICC might be consistently estimated.

**Proposition 3.** *Given a random subject  $i$ ,  $i = 1, \dots, I$ , independently scored by two random raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , the conditional correlation between the ratings  $Y_{ij}$  and  $Y_{ij'}$  is*

$$\text{Corr}(Y_{ij}, Y_{ij'} | G, H) = \text{ICC} = \frac{\omega_G^2}{\omega_G^2 + \phi_H^2}. \quad (41)$$

The proof is given in Appendix B. Conditioning on different clusters of subjects or raters and different ICC formulations lead to possible comparisons among clusters similar to the main model.

## 5. Posterior inference

The parameters of the DPs' base measures (i.e.,  $G_0, H_0$ ) and the respective concentration parameters  $\alpha_1$  and  $\alpha_2$  have to be assigned either a value or a hyperprior to complete the model specification and conduct posterior inference. This section outlines our choices about the hyperprior and the posterior computation. Several parameter specifications may be considered for the DP parameters (Ghosal & van der Vaart, 2017; Hjort et al., 2010) as they may be assigned a prior or fixed in advance. We placed a hyperprior on those parameters and let the data inform their parameters.

Under this model specification, the most natural choices to compute the posterior are conditional sampling schemes, such as Blocked Gibbs Sampling, which rely upon the approximate stick-breaking construction of the DP. They directly involve the prior in the sampling scheme avoiding its marginalization and accommodating hyperprior for the base measures (Ishwaran & James, 2001). They also come with further advantages, such as an improved mixing property, better interpretability of the mixture parameters (Gelman et al., 2014; Hjort et al., 2010) and the direct computation of the ICC. Indeed, avoiding the prior marginalization, the moments (16), (17), and (18) can be readily computed and plugged in the ICC formula (32).

However, tailored considerations have to be made in practical applications based on specific data features.

### 5.1. Hyperprior specification

Eliciting the concentrations ‘and base measures’ parameters has a role in controlling the posterior distribution over clustering (Gelman et al., 2014). Small values of the variance parameters of the base measures  $G_0$ , and  $H_0$  favor the clustering of subjects and raters, respectively, to different clusters. On the contrary, larger values of  $G_0$  and  $H_0$  variances favor the allocation of different subjects and raters, respectively, to the same cluster.

We improve model flexibility by placing a prior on the base measures  $G_0$  and  $H_0$ , and the concentration parameters  $\alpha_1$  and  $\alpha_2$  letting them be informed by the data. For the subjects’ true score base measure  $G_0 = N(\mu_0, S_0) \times Ga(w_0, w_0/W_0)$  the following hyperpriors are specified:

$$\mu_0 \sim N(\lambda_{\mu_0}, \kappa_{\mu_0}^2), \quad S_0 \sim IGa(q_{S_0}, Q_{S_0}), \quad w_0 \sim Ga(q_{w_0}, Q_{w_0}), \quad W_0 \sim IGa(q_{W_0}, Q_{W_0}).$$

We let  $\lambda_{\mu_0}$  be the rating scale’s center value (e.g.,  $\lambda_{\mu_0} = 50$  on a 1-100 rating scale),  $\kappa_{\mu_0}^2 = 100$  and the parameters  $q_{w_0}, Q_{w_0}, q_{W_0}, Q_{W_0}$  equal to 0.005. For the raters’ base measure  $H_0 = N(\eta_0, D_0) \times Ga(a_0, a_0/A_0) \times Ga(b_0, b_0/B_0) \times Ga(m_0, m_0/M_0)$ , the following hyperpriors are specified:

$$\eta_0 \sim N(\lambda_{\eta_0}, \kappa_{\eta_0}^2), \quad D_0 \sim IGa(q_{D_0}, Q_{D_0}), \quad a_0 \sim Ga(q_{a_0}, Q_{a_0}), \quad A_0 \sim IGa(q_{A_0}, Q_{A_0}),$$

$$b_0 \sim Ga(q_{b_0}, Q_{b_0}), \quad B_0 \sim IGa(q_{B_0}, Q_{B_0}), \quad m_0 \sim Ga(q_{m_0}, Q_{m_0}), \quad M_0 \sim IGa(q_{M_0}, Q_{M_0}).$$

Where  $\lambda_{\eta_0} = 0$ ,  $\kappa_{\eta_0}^2 = 100$ , and the other hyperparameters are fixed to 0.005. The concentration parameters  $\alpha_1$  and  $\alpha_2$  are assumed to follow, respectively, a gamma distribution:

$$\alpha_1 \sim Ga(a_1, A_1) \quad \alpha_2 \sim Ga(a_2, A_2).$$

where  $a_1, A_1, a_2, A_2$  are fixed to 1. The values we fix for the hyperprior’s parameters are very common in literature and they are consistent with those proposed by many other studies on BNP models (e.g., Gelman et al., 2014; Heinzl et al., 2012; Mignemi et al., 2024; Paganin et al., 2023; Yang & Dunson, 2010).

### 5.2. Posterior computation

Since most of the parameters in the model have conjugate prior distributions, a Blocked Gibbs sampling algorithm was used for the posterior sampling (Ishwaran & James, 2001). No conjugate priors are available for the gamma’s shape parameters (e.g.,  $\gamma_k$ ,  $k = 1, \dots, R$ ,  $a_0$ ,  $b_0$ ), thus we approximate the full conditionals using a derivatives-matching procedure (D-M) which is involved as an additional sampling step within the MCMC. This method has several advantages over other sampling schemes (e.g., adaptive rejection sampling or Metropolis-Hasting) in terms of efficiency, flexibility, and convergence property (Miller, 2019). We use the same D-M algorithm introduced by Miller, 2019 to approximate the posterior of the gamma shape parameters of the base measures, i.e.,  $w_0, a_0, b_0, m_0$  and a modified version for the parameters  $\gamma_k$ ,  $k = 1, \dots, R$ , since the parametrization (30) is adopted. We detail this adapted version of the D-M algorithm in the paragraph below and provide the complete Gibbs sampling in the Supplementary Material.

The notation on the independent allocation of subjects and rater to the corresponding clusters is introduced here. Let  $c_{1i}$  denote the cluster allocation of subject  $i = 1, \dots, I$ , with  $c_{1i} = n$  whenever  $\xi_i = \xi_n$ ,  $n = 1, \dots, R$ . Given the finite stick-breaking approximation detailed in Section 2.4,  $R$  is the maximum number of clusters. We indicate the set of all the subjects assigned to the  $n$ th cluster with  $\mathcal{C}_{1n}$  and with  $N_{1n} = |\mathcal{C}_{1n}|$  its cardinality. Accordingly, let  $c_{2j}$  denote the cluster allocation of rater  $j = 1, \dots, J$ , such that  $c_{2j} = k$  whenever  $\zeta_j^* = \zeta_k^*$ ,  $k = 1, \dots, R$ . The set of all the raters assigned to the  $k$ th cluster is denoted by  $\mathcal{C}_{2k}$  with  $N_{2k} = |\mathcal{C}_{2k}|$  being its cardinality.

*Derivatives-matching procedure.* Since no conjugate priors are available for the gamma’s shape parameters  $\{\gamma_k\}_1^R$ , we involve, for each of these parameters, a D-M procedure to find a gamma distribution that approximates the full conditional distribution of these parameters, when their prior is also a gamma distribution (Miller, 2019).

We aim to approximate  $p(\gamma_k|\cdot)$ , i.e., the true full conditional density of  $\gamma_k$ ,  $k = 1, \dots, R$ , by finding  $U_{1k}$  and  $U_{2k}$  such that

$$p(\gamma_k|\cdot) \approx g(\gamma_k|U_{1k}, U_{2k}), \quad k = 1, \dots, R, \quad (42)$$

where  $g(\cdot)$  is a gamma density,  $U_{1k}$  and  $U_{2k}$  are shape and rate parameters, respectively. The algorithm aims to find  $U_{1k}$  and  $U_{2k}$  such that the first and the second derivatives of the corresponding log densities of  $p(\gamma_j|\cdot)$  and  $g(\gamma_k|U_{1k}, U_{2k})$  match at a point  $\gamma_k$ . Miller (2019) suggest to choose  $\gamma_k$  to be near the mean of  $p(\gamma_k|\cdot)$  for computational convenience. The approximation is iteratively refined by matching derivatives at the current  $g(\cdot)$  mean as shown by Algorithm 1. We adapt the algorithm to our proposal, more specifically we consider the model involving the shape constraint introduced in equation (30). When this constraint is not imposed, the original algorithm by Miller (2019) may be directly used.

We denote with  $X_{1k}$  and  $X_{2k}$  the sufficient statistics for  $\gamma_k$  corresponding to the  $k$ th raters' mixture component. For the implementation of the Algorithm 1 we set the convergence tolerance  $\varepsilon_0 = 10^{-8}$  and the maximum number of iterations  $M = 10$ . Here  $\psi(\cdot)$  and  $\psi'(\cdot)$  are the digamma and trigamma functions, respectively.

The parameters  $U_{1k}$  and  $U_{2k}$ , returned by the algorithm, are used to update  $\gamma_k \sim Ga(U_{1k}, U_{2k})$ ,  $k = 1, \dots, R$ , through the MCMC sampling. The derivation of the algorithm is given in the Supplementary Material.

---

**Algorithm 1** D-M Algorithm
 

---

```

 $X_{1k} \leftarrow \sum_{j \in \mathcal{C}_{2k}} \log(1/\sigma_j^2)$ 
 $X_{2k} \leftarrow \sum_{j \in \mathcal{C}_{2k}} 1/\sigma_j^2$ 
 $T_k \leftarrow X_{2k}/\beta_k - X_{1k} + N_{2k} \log(\beta_k) - N_{2k}$ 
 $U_{1k} \leftarrow b_0 + N_{2k}/2$ 
 $U_{2k} \leftarrow B_0 + T_k$ 
for  $m = 1, \dots, M$  do
   $\gamma_k = U_{1k}/U_{2k}$ 
   $U_{1k} \leftarrow b_0 + N_{2k} \gamma_k^2 \psi'(1 + \gamma_k) - N_{2k} \gamma_k^2/(1 + \gamma_k)$ 
   $U_{2k} \leftarrow B_0 + (U_{1k} - b_0)/\gamma_k N_{2k} \log(1 + \gamma_k) + N_{2k} \psi(1 + \gamma_k) + T_k$ 
  if  $|\gamma_k/(U_{1k}/U_{2k})| < \varepsilon_0$  then
    return  $U_{1k}, U_{2k}$ 
  end if
end for

```

---

### 5.3. Post-processing procedures

*Semi-centered DPM processes.* The sampling scheme detailed in the Supplementary Material provides draws under the noncentered DPM model. However, as discussed in Section 2.5, it is not identifiable, and we need to post-process the MCMC samples to make inferences under the SC-DPM parameter-expanded model Yang & Dunson (2010). Since it is a semi-centered model that naturally constrains the raters' systematic bias  $\{\tau_j\}_1^J$  to have zero mean, a few location transformations are needed. After computing  $\eta_H$  according to 17 for each iteration, the samples of  $\mu_0, \mu_G, \{\theta_i\}_1^I$ , and  $\{\tau_j\}_1^J$  are computed:

$$\begin{aligned}
 \mu_0^* &= \mu_0 + \eta_H, \\
 \mu_G^* &= \mu_G + \eta_H, \\
 \theta_i^* &= \theta_i + \eta_H, & \text{for } i = 1, \dots, I; \\
 \tau_j^* &= \tau_j - \eta_H, & \text{for } j = 1, \dots, J.
 \end{aligned}$$

The first three are due to the location transformation of  $\tau_j$  and have to be considered for inference purposes under the SC-DPM model.

*Posterior densities and clusters point estimates.* Each density equipped with a BNP prior might be monitored along the MCMC by a dense grid of equally spaced points (Gelman et al., 2014; Mignemi et al., 2024; Yang & Dunson, 2010). Each point of the grid is evaluated according to the mixture resulting from the finite stick-breaking approximation at each iteration. At the end of the MCMC, for each point of the grid posterior mean and credible interval might be computed, and as a by-product, the pointwise posterior distribution of the density might be represented.

The BNP model provides a posterior over the entire space of subjects' and raters' partitions, respectively. However, we can summarize these posteriors and determine the point estimates of these clustering structures by minimizing the respective variation of information (VI) loss functions. We refer to Wade & Ghahramani (2018) and Meilä (2007) for further details on VI and point estimates of probabilistic clustering.

As for every parameter of the model, we use the posterior distribution of the subjects' specific parameters for inference purposes. Point estimates of the subjects' true scores  $\{\theta_i\}_1^I$ , such as the posterior mean (i.e., *expected a posteriori*, EAP) or the *maximum a posteriori* (i.e., MAP), might be used as official evaluations (i.e., final grades), and the posterior credible intervals as uncertainty quantification around those values. The  $ICC_A$  index (32) can be computed at each iteration of the MCMC to get its posterior distribution, which might be used for inference purposes.

*Computational details.* In the present work, both for the simulations and the real data analysis, similarly to previous works (e.g., Heinzl et al., 2012; Paganin et al., 2023), the number of iterations is fixed to 80,000 (with a thin factor of 60 due to memory constraints), discarding the first 20,000 as burn-in. We fix the maximum number of clusters to be  $R = 25$ , respectively, for subjects' and raters' DPM priors (Gelman et al., 2014). The package *mcclust.ext* (Wade, 2015) is used for the point estimate of the clustering structures based on the VI loss functions. We graphically check out trace plots for convergence and use the package *coda* for model diagnostics (De Iorio et al., 2023; Plummer et al., 2006). Convergence is also confirmed through multiple runs of the MCMC with different starting values<sup>1</sup>.

## 6. Simulation study

We perform a simulation study to compare the performance of the proposed models (BNP and a nested version) over the standard parametric one, highlighting the strength of our method. Concerning the *individual-specific level*, the three models are evaluated on the accuracy of the estimates of the individual-specific parameters they provide (i.e., how close  $\theta_i$ ,  $\tau_j$ ,  $\sigma_j^2$  are to the respective true values). Regarding the *population level*, we compare the estimated population distribution of the subjects' and raters' features and evaluate the predictive performance of the three methods across different scenarios.

*BP model.* The first model is the Bayesian parametric one (BP model), which can be considered a reduction of the BNP model in which all the subjects and the raters are allocated to the same cluster, respectively, such that  $\mu_i = \mu$  and  $\omega_i^2 = \omega^2$ , for  $i = 1, \dots, I$ , and  $\eta_j = \eta$ ,  $\phi_i^2 = \phi^2$ ,  $\gamma_j = \gamma$  and  $\beta_j = \beta$  for  $j = 1, \dots, J$ . This model might be obtained by fixing the maximum number of clusters  $R = 1$ .

*BSP model.* The second model is the BSP one (BSP model), in which the normality assumption is relaxed for the subjects' true score such that we model  $\{\theta_i\}_1^I$  as detailed in Section 2.3, but we model the raters' effects  $\{\tau_j, 1/\sigma_j^2\}_1^J$  as in the parametric model (i.e., they are all assigned to the same cluster). This implies  $R = 1$  only for the rater-related DPM. Since in this model both  $G$  and  $H$  are degenerate on a mixture of only one component, we refer to the structural parameter as  $\mu_G = \mu$ ,  $\omega_G^2 = \omega^2$  and  $\phi_H^2 = \phi^2$ .

<sup>1</sup>CPU configuration: 12th Gen Intel(R) Core(TM) i9 12900H.



*BNP model.* The third model is the BNP model presented in Section 2.3 in which the normality assumption is relaxed both for subjects and raters. Under this model, subjects and raters are allowed to be, respectively, assigned to different clusters.

Three data-generative processes are set up with different clustering structures for subjects and raters. The densities of the subject's true score and the rater's effects are either unimodal, bimodal or multimodal. This allows us to assess the extent to which BNP priors might mitigate model misspecification and the BNP model reduces to the parametric one when the latter is properly specified; this setup is consistent with other works on BNP modeling in psychometrics (Paganin et al., 2023).

We keep some features of the generated data similar to the real data set analyzed in Section 7 (e.g., sample size, rating scale, ratings per subject), they are also comparable with those of other works on rating models (Bartoš & Martinková, 2024; Martinková et al., 2023). Additional simulation results on small sample size applications of our proposal are presented in the Supplementary Material.

### 6.1. Setting

We generate subjects' ratings on a continuous scale,  $Y_{ij} \in (1, 100)$ , the number of subjects  $I = 500$  and raters  $J = 100$  are fixed, whereas the number of ratings per subject and the true generative model vary across scenarios.

*Generative scenarios.* We manipulate the number of ratings per subject to be  $|\mathcal{R}_i| \in \{2, 4\}$  for  $i = 1, \dots, I$ , since in many real contexts (e.g., education, peer review) it is common for the subjects to be rated only by two or few more independent raters (Zupanc & Štrumbelj, 2018).

Data are generated as specified by equations 5, 9, and one of the schemes below, according to the three different scenarios:

**Unimodal:** Under this scenario, subjects' true score and raters' effects densities are unimodal:

$$\theta_i \stackrel{\text{iid}}{\sim} N(50, 50), \quad (\tau_j, 1/\sigma_j^2) \stackrel{\text{iid}}{\sim} N(0, 25) \text{Ga}(10, 10/0.15),$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . This corresponds to the standard BP model in which subjects' true scores are assumed to be i.i.d across subjects and raters' effects are drawn jointly i.i.d. across raters.

**Bimodal:** In this scenario, both subjects' and raters' populations are composed, respectively, of two different clusters:

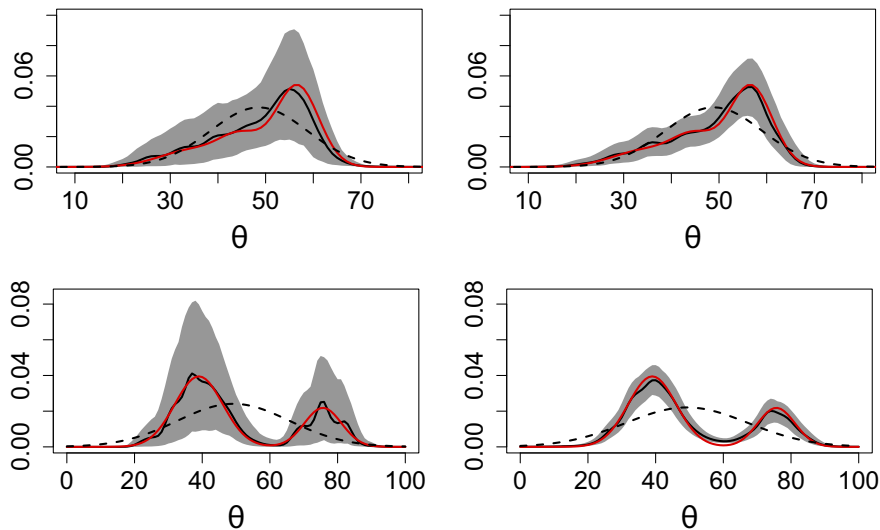
$$\begin{aligned} \theta_i &\stackrel{\text{iid}}{\sim} 0.7 \cdot N(39, 50) + 0.3 \cdot N(75.6, 30), \\ (\tau_j, 1/\sigma_j^2) &\stackrel{\text{iid}}{\sim} 0.5 \cdot N(-5, 10) \text{Ga}(10, 10/0.1) + 0.5 \cdot N(5, 5) \text{Ga}(10, 10/0.2), \end{aligned}$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ .

**Multimodal:** Under this scenario, both subjects and raters are assigned, respectively, to three clusters:

$$\begin{aligned} \theta_i &\stackrel{\text{iid}}{\sim} 0.2 \cdot N(35, 50) + 0.2 \cdot N(45, 20) + 0.6 \cdot N(56.6, 20), \\ (\tau_j, 1/\sigma_j^2) &\stackrel{\text{iid}}{\sim} 0.4 \cdot \text{SN}(-5, 3.162, -5) \text{Ga}(10, 10/0.15) \\ &\quad + 0.4 \cdot N(0, 10) \text{Ga}(10, 10/0.10) \\ &\quad + 0.2 \cdot N(10, 10) \text{Ga}(10, 10/0.20), \end{aligned}$$

for  $i = 1, \dots, I$  and  $j = 1, \dots, J$ . Here  $\text{SN}(\xi, \omega, \alpha)$  stands for the skew-normal distribution with location, scale and slant parameters,  $\xi$ ,  $\omega$  and  $\alpha$ , respectively.



**Figure 3.** Average estimated density across 10 independent datasets under different scenarios. The columns indicate the cardinality of  $|\mathcal{R}_i| = \{2, 4\}$ : left and right, respectively; the rows indicate *bimodal* or *multimodal* scenario: first and second row, respectively. The solid red lines indicate the true densities; the solid black line and the shaded grey area indicate, respectively, the point-wise mean and 95% quantile-based credible intervals; the density implied by the BP model (black dotted lines).

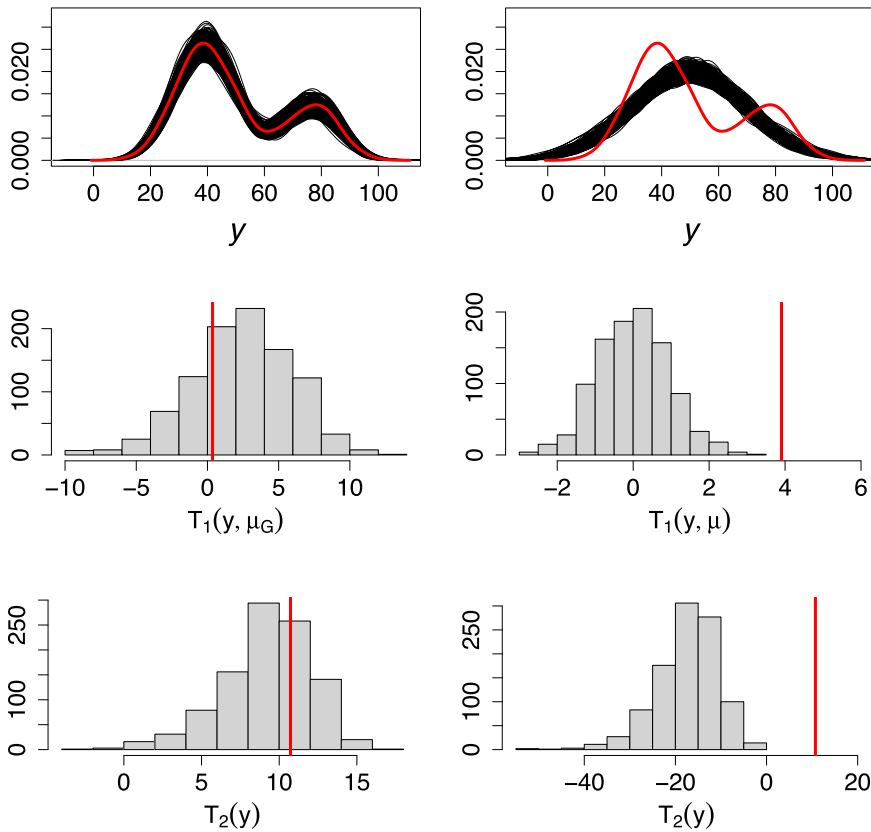
These scenarios mimic three different levels of heterogeneity. From an interpretative point of view, in the first scenario, all the subjects' true scores are concentrated around the center of the rating scale, and the raters are quite homogeneous in their severity and reliability. The heterogeneity of the subjects and the raters is only at the individual level since they are not nested with clusters. Under the second scenario, we introduce heterogeneity at the population level as both subjects and raters are assigned to different clusters, respectively. Here, we mimic the case in which subjects are clustered within two different levels of true score (e.g., low versus high proficiency level), and raters are either systematically slightly more lenient and reliable or more severe and less reliable. Under the third scenario, subjects and raters are assigned, respectively, to three poorly separated clusters. This results in a highly negatively skewed distribution for the subjects' true score and a multimodal distribution for the raters' systematic bias. Figures 3 and 4, Figure C.1 in Appendix C, and Figures 1, 2, and 3 in the Supplementary Material show the respective true densities and the empirical distributions of the generated ratings.

Ten independent data sets are generated under the six scenarios resulting from the  $2 \times 3$  design, for each data set, the standard parametric (BP), the semi-parametric (BSP), and the nonparametric (BNP) models are fitted.

**Model recovery assessment.** Parameter recovery performance is assessed through the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE) computed, respectively, as the root mean square difference and the mean absolute difference between the posterior mean and the true value of the parameters across data sets. For the subject and raters specific parameters, i.e.,  $\{\theta_i\}_1^I$ ,  $\{\tau_j, 1/\sigma_j^2\}_1^J$ , RMSE, and MAE are average both across individuals and data sets.

For the sake of comparison across different scenarios, we report the standardized version of both indices (S-RMSE, S-MAE) for the structural parameters. More precisely, those related to  $\mu$  and  $\mu_G$  are divided by the mean value of the rating scale, i.e., 50; those regarding  $\omega^2$ ,  $\omega_G^2$ ,  $\phi^2$ ,  $\phi^2$ ,  $\tilde{\sigma}$ ,  $\tilde{\sigma}_H$ , and the  $ICC_A$  are divided by their true value.

The models' performance in recovering the density distributions of individuals' specific parameters is evaluated through visual inspection. We give an example of how different densities might lead to very different conclusions on the data generative process (Gelman et al., 2013; Paganin & de Valpine, 2024; Steinbakk & Storvik, 2009). Specifically, we draw new replications from the respective



**Figure 4.** Top row: empirical distribution of the data (red solid line) and empirical distribution of replicated data (black solid lines) from the respective BNP and BP posterior distributions (left and right columns, respectively). Middle and bottom row: Test statistics computed on the data (red solid line) and histograms of those computed on replicated data.

posterior predictive distributions and compare these samples to the original data. If the models capture relevant aspects of the data, they should look similar, and replications should not deviate systematically from the data. We measure discrepancy in central asymmetry through the statistic  $T_1(y, \mu_G) = |y_{.25} - \mu_G| - |y_{.75} - \mu_G|$ , where  $y_{.25}$  and  $y_{.75}$  are the first and the third quartile, and in the left tail weight by the statistics  $T_2(y) = \min(y)$ .

## 6.2. Results

Results from the simulation study suggest that our proposals (i.e., BSP and BNP) systematically improve the estimates of the individual-specific parameters across scenarios. However, the accuracy of these estimates is comparable under the *unimodal* scenarios across the three models. Meanwhile, the BSP and BNP models overcome, on average, the BP under the *bimodal* and *multimodal* scenarios in both conditions  $|\mathcal{R}_i| = 2$  and  $|\mathcal{R}_i| = 4$ . As expected, the accuracy of subjects' and raters' specific parameters is higher in the conditions with a larger number of raters per subject  $|\mathcal{R}_i| = 4$  (Table 1). As indicated by *RMSE* and *MAE* indices, on average, the estimates of subjects' and raters' specific parameters provided by all the models degrade from the *unimodal* to the *multimodal* scenario.

Regarding the population parameter estimates, all the models provide overall similar estimates (Table 2). We observe the largest improvement of the BSP and BNP over the parametric model under the *bimodal* scenarios concerning subjects' true score variance  $\omega_G^2$  and raters' systematic bias variance  $\phi_H^2$ . However, in these cases, the BP model provides better estimates of the expected residual variance  $\hat{\sigma}_H$ .

**Table 1.** RMSE and MAE of individuals parameters corresponding to BP, BSP and BNP models.

			Generative model					
			Unimodal		Bimodal		Multimodal	
			RMSE	MAE	RMSE	MAE	RMSE	MAE
$ \mathcal{R}_i  = 2$	$\theta$	BP	<b>2.123</b>	1.686	2.346	1.846	2.497	2.009
		BSP	2.127	1.689	<b>2.308</b>	<b>1.822</b>	<b>2.347</b>	<b>1.889</b>
		BNP	<b>2.123</b>	<b>1.683</b>	2.327	1.841	2.439	1.961
	$\tau$	BP	1.404	1.102	1.575	1.208	1.892	1.566
		BSP	1.407	1.104	1.554	<b>1.206</b>	<b>1.700</b>	<b>1.387</b>
		BNP	<b>1.401</b>	<b>1.101</b>	<b>1.553</b>	1.212	1.774	1.460
	$1/\sigma^2$	BP	0.070	0.060	0.092	0.076	0.085	0.070
		BSP	<b>0.069</b>	<b>0.059</b>	<b>0.071</b>	0.054	<b>0.066</b>	<b>0.050</b>
		BNP	0.071	<b>0.059</b>	<b>0.071</b>	<b>0.052</b>	<b>0.066</b>	<b>0.050</b>
$ \mathcal{R}_i  = 4$	$\theta$	BP	1.442	1.154	1.512	1.192	1.817	1.471
		BSP	1.441	1.155	1.474	1.164	1.593	1.275
		BNP	<b>1.439</b>	<b>1.151</b>	<b>1.466</b>	<b>1.157</b>	<b>1.527</b>	<b>1.217</b>
	$\tau$	BP	0.860	0.688	0.920	0.726	1.384	1.157
		BSP	0.860	0.686	0.886	0.711	1.088	0.885
		BNP	<b>0.849</b>	<b>0.680</b>	<b>0.878</b>	<b>0.707</b>	<b>0.996</b>	<b>0.798</b>
	$1/\sigma^2$	BP	0.037	0.029	0.054	0.042	<b>0.046</b>	0.036
		BSP	0.037	0.029	0.054	0.041	0.048	0.037
		BNP	0.037	0.029	<b>0.047</b>	<b>0.035</b>	0.047	<b>0.035</b>

Note: The bold text indicates the average, most accurate estimates.

As a result, these differences are not detectable in the  $ICC_A$  estimates and we observe equal accuracy for this index across the three models.

Figure 3 gives some examples of the estimated true score densities under the *bimodal* and *multimodal* scenarios; those under the *unimodal* scenario are reported in the Appendix C. The raters’ features density plots are shown in the Supplementary Material. The BNP model consistently estimates the respective densities under all the considered scenarios. The most prominent improvement of our proposals over the parametric model is observed under the heterogeneous scenarios. Accurate estimates of the densities are also provided under the extreme case of  $|\mathcal{R}_i| = 2$ , that is, when each subject is rated by only two independent raters. Nonetheless, we note that the uncertainty about the densities is reduced when subjects are rated by a larger number of raters (i.e.,  $|\mathcal{R}_i| = 4$ ). This reduction mostly regards the subjects’ true score densities across all the scenarios. Our proposals capture the latent clustering structures of both subjects and raters as displayed by the posterior similarity matrices in Figure C.2 in Appendix C. The entries of these matrices are the pairwise probability that two entries (e.g., subjects or raters) are clustered together. The clustering structure implied by the generative process under the *bimodal* scenario is readily recognized by the graphical inspection.

The BNP model effectively captures relevant latent aspects of the data, such as deviations from normality both in the center and in the tails of the distributions across all the scenarios. As a by-product, the replications drawn from the posterior predictive distribution of the BNP model are remarkably more plausible than those generated under the BP model. As shown in Figure 4, the normality assumptions made in the latter model restrict the shapes of the distributions for subjects’ and raters’ features. As a

**Table 2.** Standardized RMSE and standardized MAE of structural parameters corresponding to BP, BSP and BNP models.

			Generative model					
			Unimodal		Bimodal		Multimodal	
			S-RMSE	S-MAE	S-RMSE	S-MAE	S-RMSE	S-MAE
$ \mathcal{R}_i  = 2$	$\mu$	BP	0.015	0.014	0.020	0.017	0.029	0.028
	$\mu_G$	BSP	0.015	<b>0.012</b>	<b>0.014</b>	<b>0.010</b>	<b>0.022</b>	<b>0.021</b>
	$\mu_G$	BNP	0.015	0.013	0.016	<b>0.010</b>	0.025	0.026
	$\omega^2$	BP	<b>0.080</b>	<b>0.066</b>	0.284	0.284	<b>0.064</b>	<b>0.052</b>
	$\omega_G^2$	BSP	0.113	0.102	<b>0.040</b>	<b>0.036</b>	0.080	0.072
	$\omega_G^2$	BNP	0.094	0.080	0.065	0.045	0.110	0.094
	$\phi^2$	BP	0.979	0.110	2.343	2.341	0.273	0.239
	$\phi^2$	BSP	0.152	0.111	0.161	0.142	0.261	0.229
	$\phi_H^2$	BNP	<b>0.134</b>	<b>0.103</b>	<b>0.112</b>	<b>0.084</b>	<b>0.195</b>	<b>0.173</b>
	$\tilde{\sigma}$	BP	0.244	0.242	<b>0.169</b>	<b>0.154</b>	0.225	0.221
	$\tilde{\sigma}$	BSP	0.226	0.213	0.206	0.186	<b>0.097</b>	<b>0.081</b>
	$\tilde{\sigma}_H$	BNP	<b>0.223</b>	<b>0.209</b>	0.253	0.228	0.108	0.091
	$ICC_A$	BP	0.002	0.002	0.001	0.001	0.023	0.023
		BSP	0.002	0.002	0.001	0.001	0.023	0.023
		BNP	0.002	0.002	0.001	0.001	0.023	0.23
$ \mathcal{R}_i  = 4$	$\mu$	BP	0.013	0.011	0.022	0.019	0.027	0.022
	$\mu_G$	BSP	<b>0.012</b>	0.011	<b>0.017</b>	<b>0.014</b>	<b>0.018</b>	<b>0.015</b>
	$\mu_G$	BNP	<b>0.012</b>	0.011	0.018	<b>0.014</b>	0.019	<b>0.015</b>
	$\omega^2$	BP	<b>0.055</b>	<b>0.046</b>	0.281	0.281	<b>0.049</b>	<b>0.042</b>
	$\omega_G^2$	BSP	0.108	0.092	<b>0.046</b>	<b>0.043</b>	0.066	0.052
	$\omega_G^2$	BNP	0.088	0.073	0.054	0.051	0.119	0.101
	$\phi^2$	BP	0.994	0.110	2.319	2.317	0.275	0.258
	$\phi^2$	BSP	0.124	0.109	0.180	0.146	0.279	0.262
	$\phi_H^2$	BNP	<b>0.119</b>	<b>0.105</b>	<b>0.132</b>	<b>0.095</b>	<b>0.209</b>	<b>0.188</b>
	$\tilde{\sigma}$	BP	<b>0.042</b>	<b>0.034</b>	<b>0.140</b>	0.130	<b>0.053</b>	<b>0.041</b>
	$\tilde{\sigma}$	BSP	0.054	0.036	0.146	0.123	0.076	0.066
	$\tilde{\sigma}_H$	BNP	0.043	0.038	0.141	<b>0.114</b>	0.074	0.063
	$ICC_A$	BP	0.002	0.002	0.001	0.001	0.023	0.023
		BSP	0.002	0.002	0.001	0.001	0.023	0.023
		BNP	0.002	0.002	0.001	0.001	0.023	<b>0.022</b>

Note: The bold text indicates the average, most accurate estimates.

result, when these assumptions are violated, any inferences about the data-generating process might be misleading and unreliable. Replications under the BP model are far from the data both in the centre and on the tails of the distribution, as suggested by the statistics  $T_1(y, \mu)$  and  $T_2(y)$  in Figure 4.

The improvement of our method over the parametric one is more prominent when the design is balanced (e.g., fully crossed designs) and the samples of subjects and raters are smaller. We present these results in the Supplementary Material.

7. Application on large-scale essay assessment

We analyze the *Matura* data set from Zupanc & Štrumbelj (2018) as an illustrative example. The data come from a large-scale essay assessment conducted by the National Examination Centre in upper secondary schools in Slovenia during the nationwide external examination. Each student received a holistic grade on a 1-50 rating scale by two independent teachers. We considered a random sample of  $I = 700$  students out of the 6995 who were examined during the spring term argumentative essays for the year 2014. A sample of  $J = 152$  teachers were involved who graded, on average, 9.21 students, with a minimum of 2 and a maximum of 21 (see Figure 5). The observed ratings ranged from 0 to 50, with a mean of 29.35, a skewness of  $-0.051$ , and a kurtosis of 3.148 (see Figure 5). More details about the assessment procedure might be found in Zupanc & Štrumbelj (2018).

*Model comparison.* The three different models detailed in Section 6, i.e., the parametric (BP) model, the semiparametric (BSP) model, and the nonparametric (BNP) model, were fitted to these data and compared on their out-of-sample prediction accuracy. The Watanabe–Akaike information criterion (WAIC) was used for this purpose.

This is a fully Bayesian approach for estimating the out-of-sample expectation, which relies on the computed log pointwise posterior predictive density and on a penalty term correction for the effective number of parameters to prevent overfitting (Gelman et al., 2014). The respective WAIC formulas are provided in the Supplementary Material.

7.1. Results

The total computational elapsed time for the BP, BSP, and BNP models was 180, 300, and 355 minutes, respectively. No convergence or mixing issues emerged from the graphical inspections of the MCMCs and diagnostics from CODA package (Plummer et al., 2006); further details and examples of trace plots are given in the Supplementary Material. Table 3 shows the WAIC indices for each fitted model and shows that the selection procedure indicates that the BNP model best fits the data and overcomes the others in predicting out-of-sample ratings. These results are consistent with the additional hold-out validation procedure presented in the Supplementary Material. Based on the model comparison procedure, we focus on the results from the BNP model.

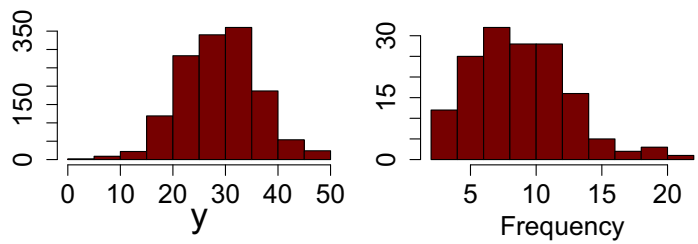


Figure 5. The empirical distribution of ratings and the frequency of students per teacher are reported at left and right, respectively.

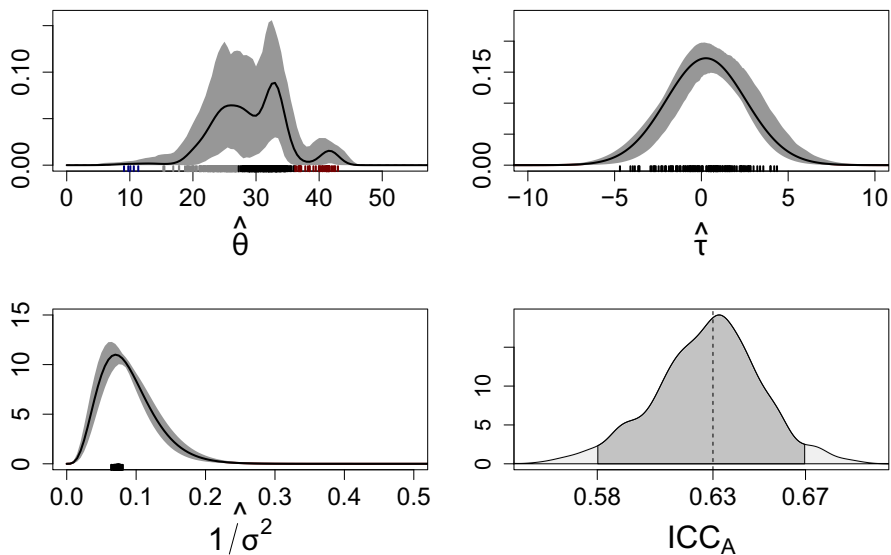
Table 3. The WAIC is reported for each of the fitted models: BNP, BP, and BSP; the pairwise WAIC difference ( $\Delta WAIC$ ) between the model with the best fit and each other is reported

Fitted model	WAIC	$\Delta WAIC$
BNP model	56267.43	–
BSP model	67159.21	–10891.78
BP model	168701.8	–112434.4



**Table 4.** Posterior mean and 95% quantile-based credible intervals of the estimated structural parameters of the BNP model are reported

		Posterior mean	95% Credible interval
Subjects' parameters	$\mu_G$	29.126	(27.886, 29.837)
	$\omega_G^2$	32.702	(28.198, 37.702)
	$\alpha_1$	4.053	(0.915, 9.218)
Raters' parameters	$\phi_H^2$	5.465	(4.085, 7.476)
	$\tilde{\sigma}_H$	13.913	(12.424, 15.583)
	$\alpha_2$	1.839	(0.194, 5.206)
	$ICC_A$	0.627	(0.577, 0.672)



**Figure 6.** The estimated densities of the subject's true score  $\theta$ , rater's systematic bias  $\tau$  and the residual term  $\varepsilon$  are reported; the black solid lines and the shade grey areas indicate the pointwise posterior mean and 95% quantile-based credible intervals of the respective densities. Bottom-right figure shows the posterior distribution of the  $ICC_A$ , the black solid and dotted lines indicate, respectively, the 95% credible interval and the posterior mean. The rugs at the margins of the first three figures indicate the clustering of individuals.

The posterior expectation of student ability mean  $\mu_G$  and variance  $\omega_G^2$  population parameters are 29.126, and 32.702, respectively (Table 4). The respective narrow credible intervals suggest low uncertainty about these values. As expected from Antoniak (1974), the posterior values of the concentration parameters  $\alpha_1$  and  $\alpha_2$  are proportional to the respective sample sizes and larger for the former. Details of the posterior values of base measures' parameters are reported in the Supplementary Material. The posterior expectation of raters' systematic bias variance  $\phi_H^2$  and reliability  $\tilde{\sigma}_H$  are, respectively, 5.465 and 13.913. The corresponding credible intervals suggest low uncertainty around these values (Table 4).

Figure 6 gives the graphical representation of the respective estimated densities. The multimodal distribution of student ability  $\theta$  implies heterogeneity among student abilities and points to the presence of multiple sub-populations. The variance in ratings is broadly due to students' ability, despite the variability of raters' systematic bias and reliability. Regarding the clustering structure of subjects and raters, the posterior similarity matrix, reported in Figure C.2 in Appendix C, suggests the presence of some latent partition of subjects, whereas no evidence of raters' clusters emerged from the posterior. This is coherent with the clusters' point estimate based on the variation of information (VI) loss function,

which indicates four clusters for the subjects and one cluster for the raters. We render this result in Figure 6 through rugs of different colors at the margin of the density plots; these values indicate the posterior mean of each subject and rater specific parameter. It is worth noting that we observe a cluster of subjects whose proficiency level is remarkably lower than the others, and another cluster in which subjects' performance is slightly superior than the others (Figure 6, upper-left; blue and brown rugs, respectively). These subjects might benefit from more personalized and specialized educational pathways. The posterior distribution of the  $ICC_A$  with mean and credible intervals, respectively, equal to 0.627 and (0.577, 0.672), suggests a moderate inter-rater reliability; Figure 6 shows the posterior distribution of this index. Since  $ICC_A$  might be interpreted as the lower bound of the expected inter-rater reliability of a single rating, poor levels of reliability can be excluded (Koo & Li, 2016). However, this result is coherent with the findings of the original study by Zupanc & Štrumbelj (2018), where raters' variability and reliability have a substantial effect on ratings. Aggregate or average ratings over different teachers might mitigate inter-rater reliability issues (Erosheva et al., 2021).

## 8. Coarsened ratings extension

Ratings data might be arbitrarily coarsened into a small number of ordered categories (Goel & Thakor, 2015; Harbaugh & Rasmusen, 2018; Peeters, 2015; van Praag et al., 2025). As a result, continuous ratings that fall between two consecutive cut-offs are collapsed into the same ordered category, and fine-grained distinctions between individual scores are missing (Ho & Reardon, 2012; Reardon et al., 2017). The available ratings are ordinal in these cases, and the rating model proposed in Section 2.3 has to be modified accordingly.

We leverage the underlying response variable formulation to extend the model to the ordinal case and consider the data coarsening mechanism (Agresti, 2015; Albert & Chib, 1993; Bartholomew et al., 2011; Cao et al., 2010; Nelson & Edwards, 2015). Our proposal might be seen as a BNP extension of the heteroscedastic ordered probit (HETOP; Lockwood et al., 2018). We specify the cumulative density function of the standard normal  $\Phi(\cdot)$  as a link function, which implies that we only need to modify the equation (5). This extension might readily adapt to the One-Way designs presented in Section 4.

We note that coarse and ordinal ratings might be rather different. In the first case, the categories are consecutive intervals of a continuous rating scale, which is not the case for ordinal ratings. Here, we propose the HETOP specification as a possible straightforward extension of the main model for coarsened ratings and leave more advantageous formulations for ordinal data for future investigations.

### 8.1. Categorical modeling

*Modeling rating  $Y_{ij}$*  We assume that the observed ordinal rating  $Y_{ij} \in \{1, \dots, K\} \subset \mathbb{N}$  is generated by an underlying unobserved normally distributed variable  $Y_{ij}^*$  (Jöreskog & Moustaki, 2001) and that we observe  $Y_{ij} = k$  if  $\delta_{k-1} < Y_{ij}^* \leq \delta_k$ ;  $\delta_0 = -\infty < \delta_1 < \dots < \delta_K = +\infty$  are ordered thresholds over the underlying response variable distribution and are equal across raters. The underlying variable  $Y_{ij}^*$  might be interpreted as a latent rating or the original continuous rating before the coarsening procedure. The conditional probability that  $Y_{ij} = k$  is

$$\mathbb{P}[Y_{ij} = k | \theta_i, \tau_j, \sigma_j, \delta_k, \delta_{k+1}] = \Phi\left(\frac{\delta_{k+1} - \theta_i - \tau_j}{\sigma_j}\right) - \Phi\left(\frac{\delta_k - \theta_i - \tau_j}{\sigma_j}\right), \quad (43)$$

for  $i = 1, \dots, I$ ;  $j \in \mathcal{R}_i$ . Additional considerations on the interpretation of  $\sigma_j$  under this formulation are given in the Supplementary Material.

*Identifiability issues.* Under this parametrization, we need to put additional constraints for identifiability purposes since the underlying response variables' mean and variance are freely estimated (DeYoreo & Kottas, 2018; Kottas et al., 2005). Two thresholds (e.g.,  $\delta_1, \delta_{K-1}$  as proposed by Song et al., 2013) have to

be fixed in advance, as it is common in multi-group analysis (Lockwood et al., 2018). From a statistical perspective, we note that each rater might be seen as a group of observations (Papaspiliopoulos et al., 2023). Moreover, an SC-DPM prior has to be placed on the subject's true score  $\{\theta_i\}_1^I$  to fix their mean and resolve identifiability issues (Gelman et al., 2014), as a by-product under the parameter-expanded specification, equation (43) becomes

$$\mathbb{P}[Y_{ij} = k | \theta_i^*, \tau_j^*, \sigma_j, \delta_k, \delta_{k+1}] = \Phi\left(\frac{\delta_{k+1} - \theta_i^* - \tau_j^*}{\sigma_j}\right) - \Phi\left(\frac{\delta_k - \theta_i^* - \tau_j^*}{\sigma_j}\right), \quad (44)$$

for  $i = 1, \dots, I$ ;  $j \in \mathcal{R}_i$ . Whenever  $K = 2$ , i.e., dichotomous rating scale,  $\{\sigma_j\}_1^I$  can not be identified and need to be fixed in advance, e.g.,  $\sigma_j = 1, j = 1, \dots, J$ , which implies assuming raters to be equally reliable (Lockwood et al., 2018).

*Generalized ICCs.* Under this model specification, the ICCs computed according to propositions 1 and 2 are generalized ICCs that indicate the polychoric correlation between two latent ratings (Jöreskog, 1994; Uebersax, 1993). For instance, proposition 1 implies here:

$$ICC_{j,j'}^* = \text{Corr}(Y_{ij}^*, Y_{ij'}^* | G, H, \sigma_j^2, \sigma_{j'}^2) = \frac{\omega_G^2}{\sqrt{\omega_G^2 + \phi_H^2 + \sigma_j^2} \sqrt{\omega_G^2 + \phi_H^2 + \sigma_{j'}^2}}, \quad (45)$$

where  $ICC_{j,j'}^*$  indicates the conditional pairwise polychoric correlation between the latent ratings given by raters  $j \neq j'$  to subject  $i$ . Similar considerations might be extended to propositions 2 and 3. As a by-product, the  $ICC_A^*$  is the lower bound of the expected polychoric correlation between the latent ratings  $Y_{ij}^*$  and  $Y_{ij'}^*$ , with  $j \neq j'$ :

$$ICC_A^* \leq \mathbb{E}[\text{Corr}(Y_{ij}^*, Y_{ij'}^* | G, H)] = \mathbb{E}[ICC^* | G, H]. \quad (46)$$

## 8.2. Posterior computation

A data augmentation procedure may simulate the underlying response variables (Albert & Chib, 1993). The underlying continuous ratings  $Y_{ij}^*$ ,  $i = 1, \dots, I, j \in \mathcal{R}_i$  are sampled:

$$Y_{ij}^* | \cdot \stackrel{\text{ind}}{\sim} N(\theta_i^* - \tau_j, \sigma_j^2) \times I(\delta_{k-1} < Y_{ij}^* \leq \delta_k), \quad k = 1, \dots, K.$$

Here  $I(\cdot)$  is an indicator function. Following Albert & Chib (1993) the conditional posterior distribution of the  $K - 3$  freely estimated thresholds, e.g.,  $\delta_2, \dots, \delta_{K-2}$  might be seen to be uniform on the respective intervals:

$$\delta_k | \cdot \stackrel{\text{ind}}{\sim} U(\max\{\max\{Y_{ij}^* : Y_{ij} = k\}, \delta_{k-1}\}, \min\{\min\{Y_{ij}^* : Y_{ij} = k + 1\}, \delta_{k+1}\}),$$

here  $U(\cdot)$  stands for uniform distribution.

All the other parameters are updated according to the posterior sampling scheme detailed in Section 3.1 of Supplementary Material and the post-process transformation outlined in Section 5.3 needs to take into account the double-centering. After computing  $\mu_G$  and  $\eta_H$  according to 16 and 17 for each iteration, the samples of  $\mu_0, \mu_G, \{\theta_i\}_1^I$ , and  $\{\tau_j\}_1^J$  are computed as follows:

$$\begin{aligned} \mu_0^* &= \mu_0 - \mu_G + \eta_H, \\ \theta_i^* &= \theta_i - \mu_G + \eta_H, & \text{for } i = 1, \dots, I; \\ \tau_j^* &= \tau_j - \eta_H + \mu_G, & \text{for } j = 1, \dots, J. \end{aligned}$$

## 8.3. Generated and real coarsened ratings analysis

In this section, we present the analysis of real and generated coarsened ratings and compare the results with those presented in Sections 6 and 7. For the real data, we deliberately coarsened the original

**Table 5.** RMSE and MAE of individuals parameters across *bimodal* scenarios with coarsened ratings

		$ \mathcal{R}_i  = 2$		$ \mathcal{R}_i  = 4$	
		RMSE	MAE	RMSE	MAE
$\theta$	BP	6.333	5.151	4.995	4.016
	BNP	<b>4.846</b>	<b>3.802</b>	<b>3.677</b>	<b>2.883</b>
$\tau$	BP	3.197	2.532	2.002	1.586
	BNP	<b>2.896</b>	<b>2.278</b>	<b>1.832</b>	<b>1.437</b>
$1/\sigma^2$	BP	0.195	0.184	<b>0.065</b>	<b>0.054</b>
	BNP	<b>0.104</b>	<b>0.080</b>	0.097	0.074

Note: The bold text indicates the average, most accurate estimates.

**Table 6.** The WAIC is reported for each of the fitted models: BNP, BP, and BSP; the pairwise WAIC difference ( $\Delta WAIC$ ) between the model with the best fit and each other is reported

Fitted model	WAIC	$\Delta WAIC$
BNP model	3798.11	–
BP model	3815.65	–17.54
BSP model	3897.22	–99.11

continuous ratings analyzed in Section 7 into  $K = 4$  ordered categories according to the following cutoffs:  $\delta_1 = 20$ ,  $\delta_2 = 30$ ,  $\delta_3 = 40$ . The fit of the BP, BSP, and BNP models to the data are compared according to the WAIC for ordered data discussed in the Supplementary Material.

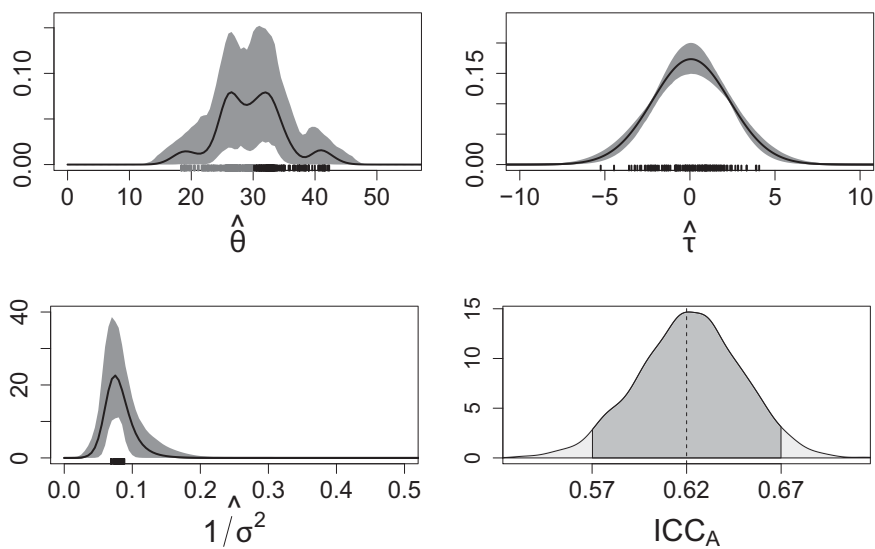
We performed a simulation study to assess the accuracy of the BNP and the BP versions for ordered ratings. More specifically, the same data sets generated under the *bimodal* scenarios in Section 6 are coarsened and considered for this study. We coarse these ratings into  $K = 4$  ordered categories according to three consecutive cutoffs:  $\delta_1 = 35$ ,  $\delta_2 = 50$ ,  $\delta_3 = 75$ . The same parameter recovery assessment procedure detailed in Section 6 is consistently used here.

In real context, the cutoffs of the coarsening procedure are generally known since the continuous rating scale is deliberately broken down into a small number of consecutive intervals and raters are explicitly asked to coarse their ratings accordingly (Peeters, 2015; van Praag et al., 2025). For example, on a 1–100 continuous scale, they might be asked to indicate which of the following intervals each subject’s score falls into: (1–25), (25–50), (50,75) or (75,100). On the contrary, when ratings are directly given on an ordinal scale, the categories’ labels are not necessarily associated with any continuous scale intervals (e.g., “poor”, “acceptable”, “good”, “very good”). In these scenarios, we consider the observed ordered ratings as coarsened representations of underlying continuous values according to some unknown consecutive cutoffs. In the first case, this coarsening process is factual; in the second, it is merely assumed. However, since in both cases at least two cutoffs need to be fixed for identification purposes, we decide to fix  $\delta_1$  and  $\delta_3$  to the true values and let the model estimate  $\delta_2$ , both for real and generated data.

**Results.** The total computational elapsed times for the BP, BSP, and BNP models were roughly similar to those of previous Sections. Upon graphical inspections of the MCMC chains and diagnostics, no convergence or mixing issues emerged for both generated and real data. Table 6 gives the WAIC indices for each fitted model and suggests that the BNP model provides the best fit to the data. Based on this model comparison procedure, we focus on the results from the BNP model. As shown in Table 7, the

**Table 7.** Posterior mean and 95% quantile-based credible intervals of the estimated structural parameters of the BNP model are reported

		Posterior mean	95% credible interval
Subjects' parameters	$\delta_2$	29.671	(28.932,30.291)
	$\mu_G$	29.678	(28.970,30.384)
	$\omega_G^2$	30.513	(25.577,36.228)
	$\alpha_1$	4.174	(1.148,8.847)
Raters' parameters	$\phi_H^2$	5.958	(4.133,9.395)
	$\tilde{\sigma}_H$	13.1080	(11.191,15.351)
	$\alpha_2$	1.911	(0.237,5.249)
	$ICC_A$	0.627	(0.577,0.672)



**Figure 7.** The estimated densities of the subject's true score  $\theta$ , rater's systematic bias  $\tau$  and the residual term  $\varepsilon$  are reported; the black solid lines and the shade grey areas indicate the pointwise posterior mean and 95% quantile-based credible intervals of the respective densities. Bottom-right figure shows the posterior distribution of the  $ICC_A$ , the black solid and dotted lines indicate, respectively, the 95% credible interval and the posterior mean. The rugs at the margins of the first three figures indicate the clustering of individuals.

estimates are equivalent to those obtained under the continuous BNP model presented in Section 7. We note that the only notable difference concerns the point estimate of the subjects' clustering structure (Figure 7). In this case, they are clustered into two (instead of four) subjects' groups.

Results from generated data suggest that the BNP model provides more accurate estimates of subjects' and raters' specific parameters and overcomes the BP model. The only exception is observed for the rater-specific reliability parameter  $1/\sigma^2$  under the scenario  $|\mathcal{R}_i| = 4$ ; here, the BP model overcomes our proposal. Under the standard parametric model, we only have two population parameters  $\gamma$  and  $\beta$  (i.e.,  $(\gamma_j, \beta_j) = (\gamma, \beta)$ , for  $j = 1, \dots, J$ ) and, as a consequence, more information is available for their estimation. This might result in a faster accuracy improvement of this model for this set of parameters as the ratio of students per rater increases. The comparison between the *RMSE* and the *MAE* of Tables 1 and 5 suggests that the estimates of both the BP and BNP models degrade with coarse data. The same trend emerged regarding the structural parameters and the densities; we report these results in the Supplementary Material.

## 9. Concluding remarks

A flexible BNP framework is proposed for the analysis of holistic rating data. We adopt the two-way unbalanced design as a general setting (McGraw & Wong, 1996) which allows us to relate our proposal to other existing models (e.g., cross-classified or crossed random effects models, multilevel models, IRT-based rating models). We specify a measurement model to jointly estimate the subject's latent quality (e.g., student's proficiency) and the rater's features (i.e., severity and consistency). Our proposal may be suitable both for balanced (i.e., when all raters score each subject; Nelson & Edwards, 2010, 2015) and unbalanced designs (i.e., when a subset of raters scores each subject; Martinková et al., 2023; Ten Hove et al., 2022). This method aims to capture latent heterogeneity among subjects and raters with the stochastic clustering induced by the DPM placed over their effects. This allows us to relax the common distributional assumptions on the respective parameters, preventing model misspecification issues (Antonelli et al., 2016; Walker & Gutiérrez-Peña, 2007).

Results from the simulation study highlight the flexibility of our proposal, which provides accurate estimates across different scenarios. Exploiting the DPM prior, the respective densities of the students' and raters' effects are consistently estimated both when the normality assumption holds and when it is violated. Our method provides a more prominent improvement in small sample sizes and with coarse data. Our proposal provides the best fit to the real data, both for continuous and coarse ratings, compared to the parametric competitor. Nonetheless, the accuracy of the estimates with coarse ratings might be a concern when subjects are only rated by a very small number of raters and the estimated true scores are used, for instance, for selection purposes or as official grades. The theoretical results presented in Section 3 are employed to make inferences about the inter-rater reliability of the single ratings.

The relatively long computational times of the MCMC chains might be prohibitive if used for repeated or massive scoring procedures. In such cases, if one is interested in capturing systematic heterogeneity among subjects or raters, any formulation of a mixture model (parametric or nonparametric) might be computationally cumbersome. In contrast, if this is not the focus of the analysis, the parametric model might be a computationally faster solution.

Under our model, rater's systematic bias and reliability are assumed to be independent conditional on the parameters of the cluster; additionally, the reliability of the raters is assumed to be independent of their specific workload  $|S_j|$  (i.e., the cardinality of the subset of subjects the rater has to evaluate). These assumptions might be unrealistic in some real contexts, and they might be relaxed under more general model specifications. For example, a multivariate distribution might be specified as a base measure  $H_0$  to account for the correlation between the rater's features, and the rater-specific workload  $|S_j|$  might be modeled as a random variable correlated to the rater's features. Furthermore, because the measurement model includes raters' effects only as an additive component, all raters are assumed to have the same ability to discriminate between subjects with different latent true scores. This assumption might be relaxed by specifying an additional rater-specific multiplicative effect for the subject's true score, similar to the GMFRMs (Uto & Ueno, 2016).

The model detailed in Section 2.3 might be further extended to account for multidimensional ratings, i.e., when subjects are rated on multiple items. Under this three-way design, item parameters might be identified under some general conditions, and the model might extend Paganin et al., 2023, or Karabatsos & Walker, 2009 to account for raters' characteristics. Further BNP generalizations of the existing rating models, e.g., GMFRMs (e.g., Uto & Ueno, 2016; Uto et al., 2024) HRMs (e.g., Casabianca et al., 2015; DeCarlo et al., 2011; Molenaar et al., 2021) or Trifactor Models (e.g., Shin et al., 2019; Soland & Kuhfeld, 2022) are left for future investigations.

The effect of covariate and contextual factors might be incorporated in the structural models 6, 8, or 9 if additional information on subjects or raters is available. This extension might relate our model to Explanatory Response Models (Kim & Wilson, 2020; Wilson & De Boeck, 2004) and be a BNP generalization of those methods. According to the data structure, more complex hierarchical priors might be placed over the subjects' true scores, such as hierarchical (Paisley et al., 2014; Teh et al., 2006), nested DPMs (Gelman et al., 2014; Hjort et al., 2010; Rodriguez et al., 2008) or hidden hierarchical DPMs recently introduced by (Lijoi et al., 2023) which overcomes some flaws of the previous ones.



Stochastic Approximations of the DPM might be further considered for the stick-breaking constructions avoiding a maximum number of clusters (Arbel et al., 2019).

Our method might provide practitioners with valuable insights about the subjects' and raters' specific features along with the respective clustering structures. This information might be used to great advantage of individualized teaching programs (Coates, 2025) and might improve the matching procedure between subjects in peer teaching activities (Stigmar, 2016). Our theoretical finding and computational solution might enhance the analysis of rating data and contribute novel knowledge about the rating process.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/psy.2025.10035>.

**Acknowledgments.** We gratefully acknowledge the anonymous reviewers for their insightful suggestions regarding the extension to the coarse rating case. We are sincerely grateful to Professors Patricia Martinkova and Elena Erosheva for the precious insights on the first draft of the article, and to Professors Antonio Lijoi and Igor Prünster for fruitful discussions on these classes of priors.

**Competing interests.** The authors declare none.

## References

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. Wiley.
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.
- Albrecht, R., Espejo, T., Riedel, H. B., Nissen, S. K., Banerjee, J., Conroy, S. P., Nickel, C. H. (2024). Clinical frailty scale at presentation to the emergency department: interrater reliability and use of algorithm-assisted assessment. *European Geriatric Medicine*, 15(1), 105–113.
- Antonelli, J., Trippa, L., & Haneuse, S. (2016). Mitigating bias in generalized linear mixed models: The case for Bayesian nonparametrics. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 31(1), 80.
- Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6), 1152–1174.
- Arbel, J., Blasi, P. D., & Prünster, I. (2019). Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*, 14(4), 1201–1219.
- Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3<sup>rd</sup> ed.) Wiley.
- Bartoš, F., & Martinková, P. (2024). Assessing quality of selection procedures: Lower bound of false positive rate as a function of inter-rater reliability. *British Journal of Mathematical and Statistical Psychology*, 77(3), 651–671.
- Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64, 153–168.
- Brix, A. (1999). Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability*, 31(4), 929–953.
- Cao, J., Stokes, S. L., & Zhang, S. (2010). A Bayesian approach to ranking and rater evaluation: An application to grant reviews. *Journal of Educational and Behavioral Statistics*, 35(2), 194–214.
- Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, 75(2), 311–337.
- Childs, T. M., & Wooten, N. R. (2023). Teacher bias matters: An integrative review of correlates, mechanisms, and consequences. *Race Ethnicity and Education*, 26(3), 368–397.
- Chin, M. J., Quinn, D. M., Dhaliwal, T. K., & Lovison, V. S. (2020). Bias in the air: A nationwide exploration of teachers' implicit racial attitudes, aggregate bias, and student outcomes. *Educational Researcher*, 49(8), 566–578.
- Coates, W. C. (2025). Precision education—A call to action to transform medical education. *International Journal of Emergency Medicine*, 18(1), 21.
- Cook, C. R., Kilgus, S. P., & Burns, M. K. (2018). Advancing the science and practice of precision education to enhance student outcomes. *Journal of School Psychology*, 66, 4–10.
- Cremaschi, A., De Iorio, M., Seng Chong, Y., Broekman, B., Meaney, M. J., & Kee, M. Z. (2021). A Bayesian nonparametric approach to dynamic item-response modeling: An application to the gusto cohort study. *Statistics in Medicine*, 40(27), 6021–6037.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., & Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 212–229.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533–559.
- DeCarlo, L. T., Kim, Y., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48(3), 333–356.

- De Iorio, M., Favaro, S., Guglielmi, A., & Ye, L. (2023). Bayesian nonparametric mixture modeling for temporal dynamics of gender stereotypes. *The Annals of Applied Statistics*, 17(3), 2256–2278.
- DeYoreo, M., & Kottas, A. (2018). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, 27(1), 71–84.
- D'lima, J., Taylor, S. E., Mitri, E., Harding, A., Lai, J., & Manias, E. (2024). Assessment of inter-rater reliability of screening tools to identify patients at risk of medication-related problems across the emergency department continuum of care. *Australasian Emergency Care*, 27(2), 136–141.
- Eroshva, E. A., Martinková, P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), 904–919.
- Escobar, M., & West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577–588.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2), 209–230.
- Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271–288.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. (2<sup>nd</sup> ed.) Taylor & Francis.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.
- Ghosal, S., Ghosh, J. K., & Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1), 143–158.
- Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.
- Goel, A. M., & Thakor, A. V. (2015). Information reliability and welfare: A theory of coarse credit ratings. *Journal of Financial Economics*, 115(3), 541–557.
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29–48.
- Harbaugh, R., & Rasmusen, E. (2018). Coarse grades: Informing the public by withholding information. *American Economic Journal: Microeconomics*, 10(1), 210–235.
- Hart, S. A. (2016). Precision education initiative: Moving toward personalized education. *Mind, Brain, and Education*, 10(4), 209–211.
- Heinzel, F., Fahrmeir, L., & Kneib, T. (2012). Additive mixed models with Dirichlet process mixture and p-spline priors. *Asta Advances in Statistical Analysis*, 96, 47–68.
- Henderson, N. C., Louis, T. A., Rosner, G. L., & Varadhan, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics*, 21(1), 50–68.
- Hjort, N., Holmes, C., Müller, P., & Walker, S. (2010). *Bayesian nonparametrics*. Cambridge University Press.
- Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal “proficiency” categories. *Journal of Educational and Behavioral Statistics*, 37(4), 489–517.
- Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, 54(3), 360–381.
- Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161–173.
- Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59(3), 381–389.
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36(3), 347–387.
- Karabatsos, G., & Walker, S. (2009). A Bayesian nonparametric approach to test equating. *Psychometrika*, 74, 211–232.
- Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement*, 80(4), 726–755.
- Koo, T., & Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155–163.
- Kottas, A., Müller, P., & Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3), 610–625.
- Królikowska, A., Reichert, P., Karlsson, J., Mouton, C., Becker, R., & Prill, R. (2023). Improving the reliability of measurements in orthopaedics and sports medicine. *Knee Surgery, Sports Traumatology, Arthroscopy*, 31(12), 5277–5285.
- Li, F., Cui, Y., Li, Y., Guo, L., Ke, X., Liu, J., Luo, X., Zheng, Y., & Leckman, J. F. (2022). Prevalence of mental disorders in school children and adolescents in China: Diagnostic data from detailed clinical assessments of 17,524 individuals. *Journal of Child Psychology and Psychiatry*, 63(1), 34–46.
- Lijoi, A., Mena, R. H., & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4), 715–740.

- Lijoi, A., Prünster, I., & Rebaudo, G. (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics*, 50(1), 213–234.
- Lin, T.-Y., Tuerlinckx, F., & Vanbelle, S. (2025). Reliability for multilevel data: A correlation approach. *Psychological Methods*, Advance online publication.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement* [Unpublished doctoral dissertation]. The University of Chicago.
- Lo, A. X., Heinemann, A. W., Gray, E., Lindquist, L. A., Kocherginsky, M., Post, L. A., & Dresden, S. M. (2021). Inter-rater reliability of clinical frailty scores for older patients in the emergency department. *Academic Emergency Medicine*, 28(1), 110–113.
- Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1), 351–357.
- Lockwood, J. R., Castellano, K. E., & Shear, B. R. (2018). Flexible Bayesian models for inferences from coarsened, group-level achievement data. *Journal of Educational and Behavioral Statistics*, 43, 663–692.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company, Inc.
- Martinková, P., Bartoš, F., & Brabec, M. (2023). Assessing inter-rater reliability with heterogeneous variance components models: Flexible approach accounting for contextual variables. *Journal of Educational and Behavioral Statistics*, 48(3), 349–383.
- Martinkova, P., Goldhaber, D., & Erosheva, E. (2018). Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PloS One*, 13, e0203002.
- Martinková, P., & Hladká, A. (2023). *Computational aspects of psychometric methods: With R*. Chapman and Hall/CRC.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30.
- Meilă, M. (2007). Comparing clusterings—An information based distance. *Journal of Multivariate Analysis*, 98(5), 873–895.
- Mignemi, G., Calcagni, A., Spoto, A., & Manolopoulou, I. (2024). Mixture polarization in inter-rater agreement analysis: A Bayesian nonparametric index. *Statistical Methods & Applications*, 33, 325–355.
- Miller, J. W. (2019). Fast and accurate approximation of the full conditional for gamma shape parameters. *Journal of Computational and Graphical Statistics*, 28(2), 476–480.
- Molenaar, D., Ulman, M., Tavşanlı, E., & De Boeck, P. (2021). The hierarchical rater thresholds model for multiple raters and multiple items. *Open Education Studies*, 3, 33–48.
- Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, 46(2), 198–219.
- Mulder, J., & Fox, J.-P. (2019). Bayes factor testing of multiple intraclass correlations. *Bayesian Analysis*, 14(2), 521–552.
- Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS One*, 7(10), e48509.
- Nelson, K., & Edwards, D. (2010). Improving the reliability of diagnostic tests in population-based agreement studies. *Statistics in Medicine*, 29(6), 617–626.
- Nelson, K., & Edwards, D. (2015). Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine*, 34, 3116–3132.
- Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement*, 56(3), 547–581.
- Paganin, S., & de Valpine, P. (2024). Computational methods for fast Bayesian model assessment via calibrated posterior p-values. *Journal of Computational and Graphical Statistics*, 34(2), 1–12.
- Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodríguez, A., Rabe-Hesketh, S., & de Valpine, P. (2023). Computational strategies and estimation performance with Bayesian semiparametric item response theory models. *Journal of Educational and Behavioral Statistics*, 48(2), 147–188.
- Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I. (2014). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2), 256–270.
- Pan, T., Shen, W., Davis-Stober, C. P., & Hu, G. (2024). A Bayesian nonparametric approach for handling item and examinee heterogeneity in assessment data. *British Journal of Mathematical and Statistical Psychology*, 77(1), 196–211.
- Papaspiliopoulos, O., Roberts, G. O., & Zanella, G. (2019). Scalable inference for crossed random effects models. *Biometrika*, 107(1), 25–40.
- Papaspiliopoulos, O., Stumpf-Fétizon, T., & Zanella, G. (2023). Scalable Bayesian computation for crossed and nested hierarchical models. *Electronic Journal of Statistics*, 17(2), 3575–3612.
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341–384.
- Peeters, M. J. (2015). Measuring rater judgment within learning assessments—Part 1: Why the number of categories matters in a rating scale. *Currents in Pharmacy Teaching and Learning*, 7(5), 656–661.
- Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11.
- Rabe-Hesketh, S., & Skrondal, A. (2016). Generalized linear latent and mixed modeling. In *Handbook of item response theory*, Wim J. van der Linden, (pp. 531–554).

- Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, 42(1), 3–45.
- Rodriguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483), 1131–1154.
- Roy, S., Daniels, M. J., & Roy, J. (2024). A Bayesian nonparametric approach for multiple mediators with applications in mental health studies. *Biostatistics*, 25(3), 919–932.
- San Martín, E., Jara, A., Rolin, J. M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, 76, 385–409.
- Sattler, D. N., McKnight, P. E., Naney, L., & Mathis, R. (2015). Grant peer review: Improving inter-rater reliability with training. *PLoS One*, 10(6), e0130450.
- Savitsky, T. D., & Dalal, S. R. (2014). Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(4), 539–557.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2), 639–650.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420.
- Soland, J., & Kuhfeld, M. (2022). Examining the performance of the trifactor model for multiple raters. *Applied Psychological Measurement*, 46(1), 53–67.
- Song, X.-Y., Lu, Z.-H., Cai, J.-H., & Ip, E. (2013). A Bayesian modeling approach for generalized semiparametric structural equation models. *Psychometrika*, 78(4), 624–647.
- Steinbakk, G. H., & Storvik, G. O. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, 36(2), 320–336.
- Stigmar, M. (2016). Peer-to-peer teaching in higher education: A critical literature review. *Mentoring & Tutoring: Partnership in Learning*, 24(2), 124–136.
- Tang, N., Chow, S.-M., Ibrahim, J. G., & Zhu, H. (2017). Bayesian sensitivity analysis of a nonlinear dynamic factor analysis model with nonparametric prior and possible nonignorable missingness. *Psychometrika*, 82(4), 875–903.
- Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.
- Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2021). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, 27(4), 650–666.
- Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, 29(5), 967–979.
- Uebbersax, J. S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association*, 88(422), 421–427.
- Uto, M., Tsuruta, J., Araki, K., & Ueno, M. (2024). Item response theory model highlighting rating scale of a rubric and rater-rubric interaction in objective structured clinical examination. *PLoS One*, 19(9), 1–23.
- Uto, M., & Ueno, M. (2016). Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, 9, 157–170.
- Uto, M., & Ueno, M. (2020). A generalized many-facet rasch model and its Bayesian estimation using hamiltonian Monte Carlo. *Behaviormetrika*, 47, 1–28.
- van Praag, B. M., Hop, J. P., & Greene, W. H. (2025). Estimation of linear models from coarsened observations: A method of moments approach. *Psychometrika*, published online 2025:1–34
- Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91, 217–221.
- Wade, S. (2015). Point estimation and credible balls for Bayesian cluster analysis [Computer software manual]. Retrieved from <https://www.maths.ed.ac.uk/~swade/docs/mcclust.ext-manual.pdf> (R package version 1.0)
- Wade, S., & Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2), 559–626.
- Walker, S. G., & Gutiérrez-Peña, E. (2007). Bayesian parametric inference in a nonparametric framework. *Test*, 16, 188–197.
- Wang, W., & Kingston, N. (2020). Using Bayesian nonparametric item response function estimation to check parametric model fit. *Applied Psychological Measurement*, 44(5), 331–345.
- Werts, C., Linn, R., & Jöreskog, K. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34(1), 25–33.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). Springer.
- Yang, M., & Dunson, D. B. (2010). Bayesian semiparametric structural equation models with latent variables. *Psychometrika*, 75, 675–693.
- Yang, M., Dunson, D. B., & Baird, D. (2010). Semiparametric Bayes hierarchical models with mean and variance constraints. *Computational Statistics & Data Analysis*, 54(9), 2172–2186.
- Zupanc, K., & Štrumbelj, E. (2018). A Bayesian hierarchical latent trait model for estimating rater bias and reliability in large-scale performance assessment. *PLoS One*, 13(4), 1–16.

## Appendices

### A. Further extensions

In this section, we present some model extensions for more flexible clustering and complex hierarchical structures. We briefly detail alternative discrete priors that generalize the Dirichlet Process, and provide a more general framework accounting for multiple source of variability.

#### A.1. DP generalizations

Following the notation in Section 2.2, given  $\Pi = DP(\alpha P_0)$  the number of different unique values  $K_n$  generated by  $p$  increase asymptotically at a logarithmic rate, with  $K_n \sim \alpha \log(n)$  a.s. for  $n \rightarrow \infty$ . Alternative priors might be specified over  $p$  which overcome this issue and allow for a more flexible prior specification on the number of clusters. More general specifications of  $\Pi$  are briefly presented below.

Our proposal might readily encompass these priors, and since they all share the stick-breaking representation presented in Section 14, the ICCs estimation and the Semi-centered identifiability procedure still hold for these cases.

##### A.1.1. Mixture of Pitman–Yor process

One of the most common generalizations of the Dirichlet Process is the Pitman–Yor Process  $PY(d, \alpha, P_0)$ , indexed by a discount parameter  $0 < d < 1$ , a concentration parameter  $\alpha > -d$ , and a base measure  $P_0$ . This is also termed the two-parameter Poisson DP. For instance, we can place the PY as a prior over the subject random measure  $G \sim PY(d, \alpha, H_0)$ , which might be represented as

$$G = \sum_{n \geq 1} \pi_{1n} \delta_{\xi_n}, \quad \pi_{1n} = V_{1n} \prod_{l < n} (1 - V_{1l}), \quad V_{1n} \stackrel{\text{iid}}{\sim} \text{Beta}(1 - d, \alpha_1 + nd), \quad \xi_n \stackrel{\text{iid}}{\sim} G_0,$$

Under this specification, the number of observed clusters  $K_I$  out of a sample of  $I$  subjects increase asymptotically at a rate  $I^d$ , with  $K_I \sim S_{d,\alpha} I^d$  as  $I \rightarrow \infty$ . Here  $S_{d,\alpha}$  is a limiting random variable with a probability distribution depending on  $d$  and  $\alpha$  and a positive density on  $\mathbb{R}^+$ . For  $d \rightarrow 0$ , we recover the  $DP(\alpha, G_0)$ , whereas for larger values of  $d$ , the rate of increase of  $K_I$  is faster. The discount parameter  $d$  might be interpreted as the proportion of small clusters that will be observed out of a sample of  $I$  subjects. Indeed, this parameter plays a double role in the clustering behavior of the model. The higher values of  $d$  imply a *reinforcement mechanism* that favors the allocation of a subject to the larger clusters (the “rich-get-richer” property) and, at the same time, a higher probability of being assigned to a new cluster. This is clear from  $\mathbb{E}[\pi_{1n}] = O(n^{-1/d})$ , for  $0 < d < 1$ , which suggests that the decay of the cluster sizes is governed by a power law.

##### A.1.2. Mixture of Normalized Generalized Gamma process

We can alternatively specify a Normalized Generalized Gamma (NGG) process as a prior for  $p \sim NGG(\alpha, d, P_0)$  (Brix, 1999; Lijoi et al., 2007). This distribution is characterized by  $\tau > 0$ ,  $d \in (0, 1)$ , and a base measure  $P_0$ . Following the previous example, we can consider the subject random measure to be distributed according to an NGG,  $G \sim NGG(\alpha, d, G_0)$ . It might be represented as

$$G = \sum_{n \geq 1} \pi_{1n} \delta_{\xi_n}, \quad \pi_{1n} = T_n / \sum_{i \geq 1} T_i, \quad \xi_n \stackrel{\text{iid}}{\sim} G_0,$$

where  $T_n$  are points of a generalized gamma process with parameters  $\alpha > 0$ ,  $d \in (0, 1)$ , and  $\sum_{i \geq 1} T_i < \infty$  (Brix, 1999). For  $d \rightarrow 0$  we recover the DP. See Ghosal & van der Vaart (2017) for the correspondence between the PY and NGG processes. The interpretation of the parameters  $\alpha$  and  $d$ , and the comments on the power law tails behavior of the PY process might be readily applied to the NGG process.

In educational rating contexts, the PY and the NGG processes might be preferred to the DP when the interest is to identify a few large clusters of subjects with similar proficiency levels and subjects who might need more *one-on-one* or personalized teaching. We refer to De Blasi et al. (2015); Hjort et al. (2010) and Ishwaran & James (2001) for a broader treatment of this class of priors.

### A.2. Multiple ways unbalanced designs

We present here an extension of our proposal for multiple-way unbalanced designs. Without loss of generality, we consider a three-way design in which ratings are additionally nested within an independent factor. It is the case where subjects are rated over different tasks (Ten Hove et al., 2022) or across different occasions (Lin et al., 2025).

Consider a subject  $i = 1, \dots, I$ , whose attribute is independently scored by a random subset of raters  $\mathcal{R}_i \subseteq \{1, \dots, J\}$  on a continuous rating scale across different tasks  $d \in \{1, \dots, D\}$ . We assume that the observed rating  $Y_{ijd} \in \mathbb{R}$  depends independently on subject  $i$ , rater  $j \in \mathcal{R}_i$  and task  $d \in \{1, \dots, D\}$ . We let the residual part, that is the difference between the true

and the observed score, depend on rater  $j$ 's effects, i.e. systematic bias and reliability. We specify the following decomposition of rating  $Y_{ijd}$ :

$$Y_{ijd} = \theta_i + \tau_j + \xi_d + SR_{ij} + SD_{id} + RD_{jd} + \varepsilon_{ijd}, \quad (47)$$

for  $i = 1, \dots, I$ ,  $j \in \mathcal{R}_i$  and  $d \in \{1, \dots, D\}$ . Here  $\theta_i$  is the true score of subject  $i$ ,  $\tau_j$  is the systematic bias of rater  $j$ ,  $\xi_d$  is the average easiness of task  $d$ ,  $\tau_j + SR_{ij}$  is the average bias of rater  $j$  in scoring subject  $i$ ,  $\xi_d + SD_i$  is the easiness of task  $d$  for subject  $i$ ,  $\tau_j + SR_{ij} + SD_{jd}$  is the bias of rater  $j$  in scoring subject  $i$  in task  $d$  and  $\varepsilon_{ijd}$  is a residual term following a zero-mean normal distribution with variance  $\sigma_j^2$  as in the two-way unbalanced design. All the parameters in (47) are assumed to be mutually independent.

We extend the model presented in Section 2.3 by placing a prior over the new set of parameters. They might be equivalently equipped with parametric or nonparametric priors. In the latter case, the same modelling strategy as described for  $\{\tau_j\}_1^J$  in Section 2.3 and 2.4 might be freely adopted for any of the new parameters.

The ICCs class for these multiple-way models might be derived straightforwardly as in the two-way design setting and following the rationale detailed in Lin et al. (2025). If nonparametric priors are placed over the new set of parameters, the variance of their respective mixture distribution might be estimated as indicated in Section 2.4. As for the two-way case, the BNP ICCs class for these models is a generalization of most of those proposed for multilevel data (e.g., Lin et al., 2025; Ten Hove et al., 2021, 2022), since the residuals' heteroschedasticity and the flexible mixture distribution of the parameters. As a consequence, when parametric priors are placed over the parameters in (47) and given  $\sigma_j^2 = \sigma^2$ , for  $j = 1, \dots, J$ , the standard ICCs in Lin et al. (2025) are recovered.

## B. Proofs

### B.1. Proof of Proposition 1

*Proof.* Let  $Y_{ij}$  and  $Y_{ij'}$  be the ratings given by two random raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , to a random subject  $i$ , for  $i = 1, \dots, I$ :

$$Y_{ij} = \theta_i + \tau_j + \varepsilon_{ij}, \quad Y_{ij'} = \theta_i + \tau_{j'} + \varepsilon_{ij'}.$$

Assuming mutual independence between the terms of the decomposition:

$$\text{Var}[Y_{ij}|G, H] = \omega_G^2 + \phi_H^2 + \sigma_j^2, \quad \text{Var}[Y_{ij'}|G, H] = \omega_G^2 + \phi_H^2 + \sigma_{j'}^2$$

and the conditional covariance between the two ratings is

$$\begin{aligned} \text{Cov}[Y_{ij}, Y_{ij'}|G, H] &= \text{Cov}[\theta_i + \tau_j + \varepsilon_{ij}, \theta_i + \tau_{j'} + \varepsilon_{ij'}|G, H] \\ &= \text{Cov}[\theta_i, \theta_i|G, H] + \text{Cov}[\theta_i, \tau_{j'}|G, H] + \text{Cov}[\theta_i, \varepsilon_{ij'}|G, H] \\ &\quad + \text{Cov}[\tau_j, \theta_i|G, H] + \text{Cov}[\tau_j, \tau_{j'}|G, H] + \text{Cov}[\tau_j, \varepsilon_{ij'}|G, H] \\ &\quad + \text{Cov}[\varepsilon_{ij}, \theta_i|G, H] + \text{Cov}[\varepsilon_{ij}, \tau_{j'}|G, H] + \text{Cov}[\varepsilon_{ij}, \varepsilon_{ij'}|G, H] \\ &= \text{Cov}[\theta_i, \theta_i|G, H] \\ &= \omega_G^2. \end{aligned}$$

The correlation between the ratings is

$$\begin{aligned} ICC_{j,j'} &= \text{Cor}[Y_{ij}, Y_{ij'}|G, H, \sigma_j^2, \sigma_{j'}^2] = \frac{\text{Cov}[Y_{ij}, Y_{ij'}|G, H]}{\sqrt{(\text{Var}[Y_{ij}|G, H])\sqrt{\text{Var}[Y_{ij'}|G, H])}}} \\ &= \frac{\omega_G^2}{\sqrt{\omega_G^2 + \phi_H^2 + \sigma_j^2} \sqrt{\omega_G^2 + \phi_H^2 + \sigma_{j'}^2}}. \end{aligned} \quad \square$$

### B.2. Proof of statement (i) of Proposition 2

*Proof.* Let  $Y_{ij}$  and  $Y_{ij'}$  be the ratings given by two random raters  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , satisfying  $\sigma_j^2 = \sigma_{j'}^2 = \tilde{\sigma}_H$  to a random subject  $i$ ,  $i = 1, \dots, I$ :

$$Y_{ij} = \theta_i + \tau_j + \varepsilon_{ij}, \quad Y_{ij'} = \theta_i + \tau_{j'} + \varepsilon_{ij'}.$$

Assuming mutual independence between the terms of the decomposition:

$$\text{Var}[Y_{ij}|G, H] = \omega_G^2 + \phi_H^2 + \tilde{\sigma}_H, \quad \text{Var}[Y_{ij'}|G, H] = \omega_G^2 + \phi_H^2 + \tilde{\sigma}_H$$

and the conditional covariance between the two ratings is

$$\begin{aligned} \text{Cov}[Y_{ij}, Y_{ij'}|G, H] &= \text{Cov}[\theta_i + \tau_j + \varepsilon_{ij}, \theta_i + \tau_{j'} + \varepsilon_{ij'}|G, H] \\ &= \text{Cov}[\theta_i, \theta_i|G, H] + \text{Cov}[\theta_i, \tau_{j'}|G, H] + \text{Cov}[\theta_i, \varepsilon_{ij'}|G, H] \end{aligned}$$



$$\begin{aligned}
 & + \mathbf{Cov}[\tau_j, \theta_i | G, H] + \mathbf{Cov}[\tau_j, \tau_{j'} | G, H] + \mathbf{Cov}[\tau_j, \varepsilon_{ij'} | G, H] \\
 & + \mathbf{Cov}[\varepsilon_{ij}, \theta_i | G, H] + \mathbf{Cov}[\varepsilon_{ij}, \tau_{j'} | G, H] + \mathbf{Cov}[\varepsilon_{ij}, \varepsilon_{ij'} | G, H] \\
 & = \mathbf{Cov}[\theta_i, \theta_i | G, H] \\
 & = \omega_G^2.
 \end{aligned}$$

The correlation between the ratings is

$$\begin{aligned}
 ICC_A = \mathbf{Cor}[Y_{ij}, Y_{ij'} | G, H] &= \frac{\mathbf{Cov}[Y_{ij}, Y_{ij'} | G, H]}{\sqrt{(\mathbf{Var}[Y_{ij} | G, H])\mathbf{Var}[Y_{ij'} | G, H]}} \\
 &= \frac{\omega_G^2}{\omega_G^2 + \phi_H^2 + \tilde{\sigma}_H}. \quad \square
 \end{aligned}$$

### B.3. Proof of statement (ii) of Proposition 2

*Proof.* Let us consider the function  $ICC$  which, conditional on  $G$  and  $H$ , is a convex function of the random variable  $\sigma_j^2$ :

$$ICC(\sigma_j^2 | G, H) = \frac{\omega_G^2}{\omega_G^2 + \phi_H^2 + \sigma_j^2}, \quad j = 1, \dots, J. \quad (48)$$

. Let  $ICC_A$  be the  $ICC$  function of the expected value of  $\sigma_j$ :

$$ICC_A = \frac{\omega_G^2}{\omega_G^2 + \phi_H^2 + \mathbf{E}[\sigma_j^2 | G, H]}, \quad j = 1, \dots, J. \quad (49)$$

Note that  $\mathbf{E}[\sigma_j^2 | G, H] = \mathbf{E}[\sigma_{j'}^2 | G, H]$  for  $j, j' = 1, \dots, J, j \neq j'$ . It readily follows from the conditional Jensen's Inequality that

$$ICC(\mathbf{E}[\sigma_j | G, H]) \leq \mathbf{E}[ICC(\sigma_j^2 | G, H)]. \quad (50)$$

Since for brevity we define  $ICC_A = ICC(\mathbf{E}[\sigma_j | G, H])$  and  $ICC = ICC(\sigma_j^2 | G, H)$ :

$$ICC_A \leq \mathbf{E}[ICC | G, H]. \quad (51)$$

where  $\mathbf{E}[ICC | G, H] = \mathbf{E}[\text{Corr}(Y_{ij}, Y_{i,j'} | G, H)]$ ,  $i = 1, \dots, I$  and  $j, j' \in \mathcal{R}_i$ . That is the expected correlation between two independent ratings given to a random subject.  $\square$

### B.4. Proof of Proposition 3

*Proof.* Let  $Y_{ij}$  and  $Y_{ij'}$  be the ratings given by  $j, j' \in \mathcal{R}_i$ ,  $j \neq j'$ , to a random subject  $i$ ,  $i = 1, \dots, I$ :

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad Y_{ij'} = \theta_i + \varepsilon_{ij'}.$$

Assuming mutual independence between the terms of the decomposition:

$$\mathbf{Var}[Y_{ij} | G, H] = \omega_G^2 + \phi_H^2, \quad \mathbf{Var}[Y_{ij'} | G, H] = \omega_G^2 + \phi_H^2$$

and the conditional covariance between the two ratings is

$$\begin{aligned}
 \mathbf{Cov}[Y_{ij}, Y_{ij'} | G, H] &= \mathbf{Cov}[\theta_i + \varepsilon_{ij}, \theta_i + \varepsilon_{ij'} | G, H] \\
 &= \mathbf{Cov}[\theta_i, \theta_i | G, H] + \mathbf{Cov}[\theta_i, \varepsilon_{ij'} | G, H] + \mathbf{Cov}[\varepsilon_{ij}, \theta_i | G, H] + \mathbf{Cov}[\varepsilon_{ij}, \varepsilon_{ij'} | G, H] \\
 &= \mathbf{Cov}[\theta_i, \theta_i | G, H] \\
 &= \omega_G^2.
 \end{aligned}$$

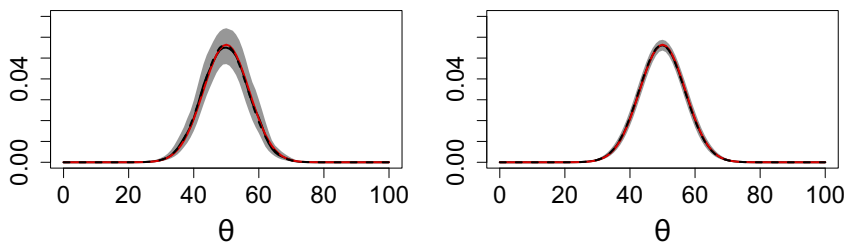
The correlation between the ratings is

$$\begin{aligned}
 ICC = \text{Cor}[Y_{ij}, Y_{ij'} | G, H] &= \frac{\mathbf{Cov}[Y_{ij}, Y_{ij'} | G, H]}{\sqrt{(\mathbf{Var}[Y_{ij} | G, H])\mathbf{Var}[Y_{ij'} | G, H]}} \\
 &= \frac{\omega_G^2}{\omega_G^2 + \phi_H^2}.
 \end{aligned}$$

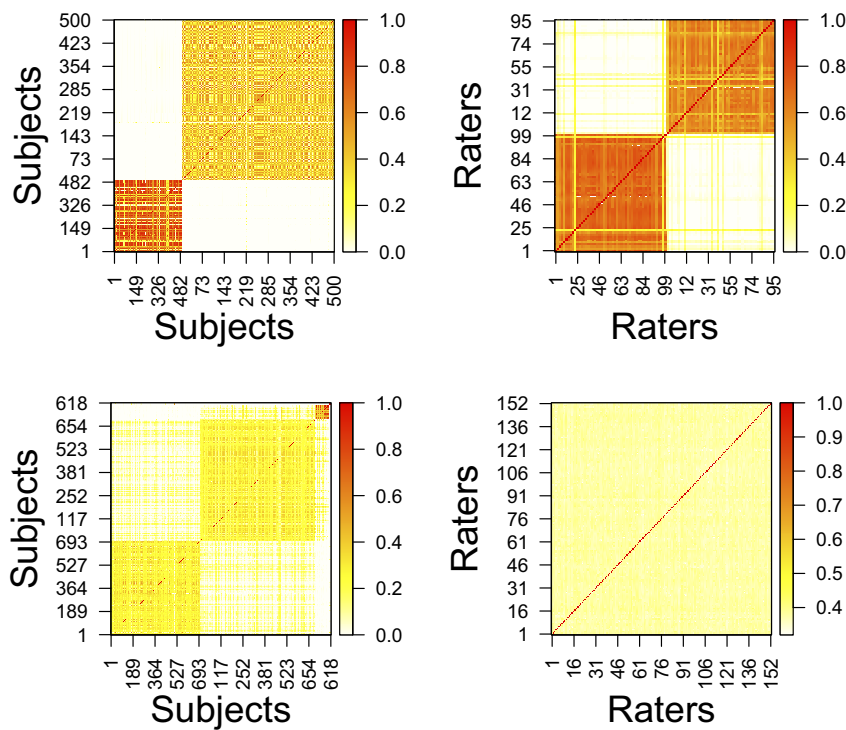
Since the conditional variance of ratings is equal across subjects  $\mathbf{Var}[Y_{ij} | G, H] = \omega_G^2 + \phi_H^2$  for  $i = 1, \dots, I$ , the  $ICC$  is unique for all the subjects.  $\square$



C. Plots



**Figure C.1.** Average estimated density across 10 independent datasets under the *unimodal* scenario. The columns indicate the cardinality of  $|\mathcal{R}_i| = \{2, 4\}$ : left and right, respectively. The solid red lines indicate the true densities; the solid black line and the shaded grey area indicate, respectively, the point-wise mean and 95% quantile-based credible intervals; the density implied by the BP model (black dotted lines).



**Figure C.2.** First row: examples of posterior similarity matrices for pairwise subject and raters allocation (left and right column, respectively). Second row: posterior similarity matrices for pairwise subject and raters allocation in real data analyzed in Section 7.