Journal of Data Science 0 (0), 1–14
??? 2025
DOI: 10.6339/25-JDS1213
Statistical Data Science

A Bayesian Approach to Pre-Post Comparison of Inter-Rater Agreement in Ordinal Ratings

Aiden Berry[1], Jennifer Cao[2], and Song Zhang[3,*]

[1]*Department of Statistics and Data Science, Southern Methodist University, Dallas, TX, USA*
[2]*Department of Ophthalmology, University of Texas Southwestern Medical Center, Dallas, TX, USA*
[3]*Department of Health Data Science and Biostatistics, University of Texas Southwestern Medical Center, Dallas, TX, USA*

Abstract

Inter-rater agreement is fundamental to decision making in medicine, psychology, and the social sciences, as it reflects the quality and reliability of rating systems. ICC (intraclass correlation) has been widely used as a measure of inter-rater agreement. To date, there has been no methodological development that properly assesses improvement in ICC for pre–post studies with ordinal ratings. It remain uninvestigated whether/how correlations between pre- and post-intervention scores impact the estimation and comparison of ICC. We present a Bayesian hierarchical probit framework for evaluating changes in ICCs in such settings. The model incorporates rater- and item-level correlations and compares two parameterizations: an "individual components" prior that separately models variances and correlations, and an inverse Wishart prior. Simulation studies show that accounting for pre–post correlation substantially improves estimation accuracy and power to detect changes in agreement, while ignoring it reduces efficiency. Application to a multicenter study on conjunctival inflammation demonstrates that a novel grading scale markedly increased inter-rater agreement. This framework underscores the importance of modeling ordinal outcomes appropriately and provides a flexible Bayesian tool for evaluating the effectiveness of interventions on inter-rater agreement in pre-post studies.

Keywords *Bayesian; inter-rater agreement; intraclass correlation; ordinal; pre-post design*

1 Introduction

In areas such as medicine, psychology, education, and business, a routine task is to score items (e.g., images, bio-samples, responses, products) into mutually exclusive (often ordinal) categories, which are used to support decision making (e.g., whether to start chemotherapy on a cancer patient or whether to flag an email as spam). The validity of the resulting decisions hinges upon the quality of the scoring. As a measure of reliability of a scoring system, the agreement among two or more raters is of high interest to researchers and practitioners alike in a wide range of areas (Hallgren, 2012). High inter-rater agreement indicates that different raters have a high tendency of scoring the same items to the same categories, hence are less prone to the influence of variation in human judgment (Shrout and Fleiss, 1979). Such assessments are particularly relevant in domains where subjective or semi-quantitative evaluations are common, including radiology, biomarker research, and survey research.

*Corresponding author. Email: song.zhang@utsouthwestern.edu.

Received September 2, 2025; Accepted December 8, 2025

The most widely used statistics to assess inter-rater agreement include Kappa-like measures and intraclass correlation coefficient (ICC). Kappa-like measures are computed as chance-corrected agreement, which apply to categorical variables and in most cases with 2 raters (Fleiss et al., 2013; Cohen, 1960). For situations involving ordered categories, weighted kappa has been proposed. It employs weighting schemes that give partial credit when raters' scores are close but not identical (Cohen, 1968). The concept of ICC was first introduced by R.A Fisher to examine the familial resemblance between siblings (Fisher, 1921). ICC is a measure that compares the variance that is attributable to the difference between what is being rated to the total variance of all scores, which further includes variability among raters and random error (Müller and Büttner, 1994). In other words, ICC quantifies how much of the observed variability is due to true differences across items, as opposed to inconsistency between raters and measurement error. For continuous variables, ICC is often derived from linear mixed models with a clear interpretation in terms of variance components (Shrout and Fleiss, 1979).

In many applications, however, the ratings are ordinal categories instead of continuous scores, e.g., tumor grades and Parkinson's severity scales. To facilitate data collection and communication, the ordinal categories are often represented by numbers. In data analysis, including the assessment of inter-rater agreement by ICC, a common practice has been to treat ordinal scores as if they were continuous and normally distributed (Gajewski et al., 2007). It has been established that ICC is sensitive to the assumption of normality and its maximum likelihood estimator is negatively biased (Fisher, 1921; Konishi, 1985; Wang et al., 1991; Atenafu et al., 2012). Treating ordinal scores as continuous in the estimation of ICC exacerbates the methodological issues, resulting in biases as well as misleading p-values and confidence intervals. For remedy, researchers have turned to latent variable models, in particular, ordinal probit models, for the assessment of inter-rater agreement in ordinal rating systems. Specifically, the observed ordinal scores are assumed to arise from a continuous latent variable upon which variance components can be extracted for the calculation of ICC (Gajewski et al., 2007; Yue et al., 2015). The observed categories are determined by the value of the latent variable in relation to a set of thresholds. For example, let x_{ij} be the latent variable and suppose that there are K ordinal categories. Then the observed ratings are

$$y_{ij} = \begin{cases} 1, & \text{if } x_{ij} < c_1 \\ 2, & \text{if } c_1 \leqslant x_{ij} < c_2 \\ \vdots, & \vdots \\ K, & \text{if } x_{ij} \geqslant c_{K-1} \end{cases}$$

where $c_1 < c_2 < \cdots < c_{K-1}$ are the ordinal cutoffs. Defining ICCs for ordinal scores on the latent variable provides a statistically sound solution, as it removes bias caused by violations of the normality assumption. However, the ICC on the latent scale is somewhat abstract. It reflects agreement on an underlying continuous construct rather than on the observed ordinal ratings, which introduces conceptual challenges for interpretation.

Under the frequentist paradigm, inference about the ICC (such as p-values and confidence intervals) has traditionally relied on normal approximations and asymptotic theory (Giraudeau and Mary, 2001). In addition to deviation from the normality assumption, estimation of ICC can be further complicated by experimental design. For example, in wine competition (Olkin et al., 2015), each rater may score only a random subset of wines, and each wine may be evaluated by only a random subset of raters, giving rise to incomplete data. To assess rater reliability, identical (blinded) samples of the same wine may be presented to a rater multiple times, introducing

a hierarchical data structure. As the model and design become more complicated, frequentist inference for ICC becomes increasingly intractable. The Bayesian framework provides a practical alternative: Bayesian hierarchical models offer flexibility in handling incomplete data and complex data structures, while prior distributions allow incorporating evidence from previous studies or expert opinion. Moreover, because inference is drawn from random samples of the posterior distribution, Bayesian inference does not rely on large-sample theory for its validity (Gelman et al., 1995; Albert and Chib, 1993).

In this study, we present a Bayesian approach to evaluating the effectiveness of an intervention on improving inter-rater agreement under a pre-post study design. It is motivated by a multicenter prospective validation study that assessed the inter-rater agreement of a novel grading scale evaluating conjunctival inflammation in cicatrizing conjunctivitis associated with Pemphigoid (Eziama et al., 2025). Ocular cicatricial pemphigoid (OCP) is an autoimmune disease characterized by chronic inflammation resulting in subepithelial fibrosis and conjunctival scarring. Close monitoring of disease progression is critical but existing grading systems utilize irreversible features of ocular damage that occur over months to years, resulting in loss of valuable intervention time. Researchers at the University of Texas Southwestern Medical Center proposed a visual analog grading scale that utilizes typified images of conjunctival inflammation of OCP patients. It offers increased sensitivity to subtle changes, thereby allowing ample opportunity for treatment before the onset of irreversible ocular damages. The validation study recruited 106 raters, including trainees and attendings. The clinical images of palpebral conjunctiva of 14 patients with OCP were first presented. The raters were asked to independently score the degree of conjunctival inflammation by 5 grades: 0 (no inflammation) to 4+ (severe inflammation). Afterwards, the visual grading scale was presented for reference, and the participants were asked to re-score the same 14 conjunctival images in the presence of the grading scale. To minimize potential bias from recall or anchoring effect—where earlier assessments influence later ones—participants were restricted from revisiting prior responses or changing prior responses. The primary research question is whether the new grading scale improves inter-rater agreement.

There has been limited development of Bayesian methods for the assessment of inter-rater agreement in ordinal scores. For the cases of two raters, Tran et al. (2021) presented Bayesian approaches to the weighted kappa-like measures. Van Oest and Girard (2022) presented a generalized Bayesian framework for chance-corrected inter-rater agreement that accommodates various weighting schemes and incomplete data. Calle-Alonso and Perez Sanchez (2015) proposed a unified Monte Carlo–based framework to estimate different types of measures of agreement in qualitative responses. Fanshawe et al. (2008) proposed an intuitive measure that reflects the extent to which the distribution of ratings provided by individual raters agrees with those provided by all raters, and estimated it using a Bayesian probit model. Gajewski et al. (2007) proposed a Bayesian approach for interrater agreement with ordinal data, incorporating a hierarchical ordinal probit model with prior distributions for the error variance and ICC. To the best of our knowledge, there has been no development of Bayesian methods that formally compare inter-rater agreement in the setting of pre-post studies with ordinal ratings. The correlation between pre- and post-intervention ratings cannot be ignored in statistical inference. In particular, whether/how this correlation impacts the comparison of pre- and post-intervention ICCs warrants investigation.

We present a Bayesian hierarchical probit model framework that appropriately accounts for the complex correlation structure, where each image is graded by each rater twice (pre- and post-intervention). Inference about ICC (estimation and comparison) is based on posterior samples of variance components on the latent scale. The rest of the paper is organized as follows.

In Section 2 we present Bayesian probit models with and without accounting for pre-post correlation. Extensive simulation results are presented in Section 3, where we show that properly accounting for pre-post correlation markedly improves the power in comparing ICCs. We apply the proposed method to the motivating example in Section 4. Finally, we briefly review the proposed Bayesian method and discuss the direction of future development in Section 5.

2 The Bayesian Probit Model

While the proposed model framework is generally applicable to inter-rater agreement studies with ordinal scores, here we adopt the notations of the motivating example, where raters are called "graders" and rated items "images". Let y_{ijt} be the observed ordinal score by the ith grader ($i = 1, \ldots, N$) on the jth image ($j = 1, \ldots, J$) at time t ($t = 0/1$ for pre-/post-adoption of the grading scale). The value of y_{ijt} is determined by a continuous latent variable x_{ijt} and a set of cutoff points $c_1 < c_2 < \cdots < c_{K-1}$:

$$y_{ijt} \mid x_{ijt}, \mathbf{c} = \begin{cases} 1, & \text{if } x_{ijt} < c_1 \\ 2, & \text{if } c_1 \leqslant x_{ijt} < c_2 \\ \vdots & \vdots \\ K, & \text{if } x_{ijt} \geqslant c_{K-1}, \end{cases} \tag{2.1}$$

where $\mathbf{c} = (c_1, \ldots, c_{K-1})'$. We model x_{ijt} by a linear mixed effect model:

$$x_{ijt} = \mu + \alpha_{jt} + \beta_{it} + \epsilon_{ijt}, \tag{2.2}$$

where μ is a fixed effect representing the overall mean, α_{jt} the random effect of image-time interaction, β_{it} the random effect of grader-time interaction, and ϵ_{ijt} the residual error. Furthermore, we model α_{jt} by

$$\begin{pmatrix} \alpha_{j0} \\ \alpha_{j1} \end{pmatrix} \mid \sigma^2_{\alpha 0}, \sigma^2_{\alpha 1}, \rho_\alpha \overset{iid}{\sim} N\left[\mathbf{0}, \begin{pmatrix} \sigma^2_{\alpha 0} & \rho_\alpha \sigma_{\alpha 0}\sigma_{\alpha 1} \\ \rho_\alpha \sigma_{\alpha 0}\sigma_{\alpha 1} & \sigma^2_{\alpha 1} \end{pmatrix}\right], \text{ for } j = 1, \ldots, J. \tag{2.3}$$

Under the above model, within period t ($t = 0, 1$), α_{jt}'s are independently and identically distributed with $N(0, \sigma^2_{\alpha t})$. Across periods, each pair of $(\alpha_{j0}, \alpha_{j1})'$ are dependent with a correlation coefficient ρ_α. Similarly, β_{it}'s are modeled by

$$\begin{pmatrix} \beta_{i0} \\ \beta_{i1} \end{pmatrix} \mid \sigma^2_{\beta 0}, \sigma^2_{\beta 1}, \rho_\beta \overset{iid}{\sim} N\left[\mathbf{0}, \begin{pmatrix} \sigma^2_{\beta 0} & \rho_\beta \sigma_{\beta 0}\sigma_{\beta 1} \\ \rho_\beta \sigma_{\beta 0}\sigma_{\beta 1} & \sigma^2_{\beta 1} \end{pmatrix}\right], \text{ for } i = 1, \ldots, N. \tag{2.4}$$

The residual errors ϵ_{ijt}'s are assumed to be independent but with different variances across periods, $\epsilon_{ijt} \mid \sigma^2_{\epsilon t} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon t})$ for $t = 0, 1$. A non-informative flat prior is specified for the overall mean μ. Finally, we assume independent Uniform(0,1) priors for ρ_α and ρ_β; and independent inverse gamma priors for the variance components:

$$\begin{aligned} \sigma^2_{\alpha t} &\sim IG(a_\alpha, b_\alpha), \\ \sigma^2_{\beta t} &\sim IG(a_\beta, b_\beta), \\ \sigma^2_{\epsilon t} &\sim IG(a_\epsilon, b_\epsilon), \end{aligned}$$

for $t = 0, 1$. Here IG(a, b) denotes an Inverse-Gamma distribution with parameterization such that the mean is $b/(a-1)$.

The ICC that measures inter-rater agreement at time t is defined by

$$\text{ICC}_t = \frac{\sigma_{\alpha t}^2}{\sigma_{\alpha t}^2 + \sigma_{\beta t}^2 + \sigma_{\epsilon t}^2}, \ t = 0, 1. \tag{2.5}$$

Inference about ICC_t is based on samples of $(\sigma_{\alpha t}^2, \sigma_{\beta t}^2, \sigma_{\epsilon t}^2)$ from the full conditional distribution [15]. Specifically, let $(\sigma_{\alpha 0}^{2(q)}, \sigma_{\alpha 1}^{2(q)}, \sigma_{\beta 0}^{2(q)}, \sigma_{\beta 1}^{2(q)}, \sigma_{\epsilon 0}^{2(q)}, \sigma_{\epsilon 1}^{2(q)})$ be the random samples obtained at the qth $(q = 1, \ldots, Q)$ iteration of MCMC simulation and calculate $\text{ICC}_0^{(q)}$ and $\text{ICC}_1^{(q)}$ following Equation (2.5). Then the posterior mean and posterior variance of ICC_t are estimated by the sample mean and sample variance of $\{\text{ICC}_t^{(q)}, q = 1, \ldots, Q\}$, $t = 0, 1$. The posterior probability that the new grading scale improves inter-rater agreement is estimated by

$$\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y}) = \sum_{q=1}^{Q} I\left(\text{ICC}_1^{(q)} > \text{ICC}_0^{(q)}\right)/Q. \tag{2.6}$$

Here $\mathbf{Y} = \{y_{ijt}\}$ denotes the collection of observed data. A large value of $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ provides strong evidence that the new grading scale improves inter-rater agreement.

When the main goal is to compare ICC_0 versus ICC_1, to avoid introducing undue bias, it is important to assume identical priors for $(\sigma_{\alpha t}^2, \sigma_{\beta t}^2, \sigma_{\epsilon t}^2)$ across periods $t = 0, 1$. Furthermore, when there is no strong knowledge regarding the relative contribution of variability from different sources (i.e., raters $(\sigma_{\beta t}^2)$, images $(\sigma_{\alpha t}^2)$, and measurement error $(\sigma_{\epsilon t}^2)$), one option is to set $a_\alpha = a_\beta = a_\epsilon = a$ and $b_\alpha = b_\beta = b_\epsilon = a$. Assigning weakly informed and identical priors to the variance components ensures that the estimation and comparison of ICCs fully reflect learning from observed data. On the other hand, when there are reliable data from previous studies or expert opinion that should be incorporated into the inference of ICCs, through the specification of prior distributions, Bayesian modeling provides a principled and coherent mechanism for integrating prior information (Gelman et al., 1995).

To analyze the motivating example, due to lack of prior information, we assign identical $IG(a, b)$ priors to $(\sigma_{\alpha 0}^2, \sigma_{\alpha 1}^2, \sigma_{\beta 0}^2, \sigma_{\beta 1}^2, \sigma_{\epsilon 0}^2, \ , \sigma_{\epsilon 1}^2)$. Importantly, they are variance components of latent variables x_{ijt}'s, whose values are only meaningful in their relation to the cutoffs $\mathbf{c} = (c_1, \ldots, c_{K-1})'$. In this study the latent variable x_{ijt} is assumed to follow the $N(0, 1)$ distribution, hence the cutoffs $\mathbf{c} = (c_1, \ldots, c_{K-1})'$ can be viewed as quantiles of the $N(0, 1)$ distribution, corresponding to the cumulative probabilities of $\mathbf{Y}$ at observed ordinal categories. We employ a data driven approach to specifying the prior for $\mathbf{c}$. First we empirically estimate the cumulative distribution of $\mathbf{Y}$ and obtain $(h_1, \ldots, h_{K-1})$, where h_k is the observed proportion of y_{ijt}'s with values less than or equal to k $(k = 1, \ldots, K-1)$. Then we assume the prior distribution of c_k to be centered around the $(100h_k)$th percentile of $N(0, 1)$, by specifying a truncated distribution:

$$c_k \sim N(0, 1)I(l_k, u_k), \ k = 1, \ldots, K-1.$$

Here $l_k = \Phi^{-1}(h_k - \delta)$ and $u_k = \Phi^{-1}(h_k + \delta)$, with $\Phi^{-1}(\cdot)$ being the inverse cumulative distribution function $N(0, 1)$, and δ is a tuning parameter controlling how much c_k's can deviate from the empirical quantiles. The above prior specification for $\mathbf{c}$ takes two key factors into consideration: (1) the assumed $N(0, 1)$ distribution of the latent variable; (2) the empirical distribution of observed ordinal scores. Since the cutoffs determine how the latent variable is mapped to the observed categories, poorly specified priors of $\mathbf{c}$ can cause the latent variable x_{ijt} to deviate from the $N(0, 1)$ distribution and misfit between model and observed data, eventually leading to bias

in ICC estimation. On the other hand, as parameters defined on the latent scale, non-informative priors on $\mathbf{c}$ would lead to poor mixing of MCMC when observed data is limited. In practice, when the sample size (number of images and raters) is relatively small, we would recommend a data driven approach similar to what was described above.

The specified prior distribution of $\mathbf{c}$ effectively anchors the distribution of latent variable x_{ijt} around $N(0, 1)$. For the prior distributions of variance components, accordingly, we set the values of a and b so that $(\sigma^2_{\alpha t} + \sigma^2_{\beta t} + \sigma^2_{\epsilon t})$ has a prior mean of 1 for $t = 0, 1$. Specifically, we set $a = 2$ and $b = 0.333$. Note that $IG(2, 0.333)$ has a mean of 0.333 but the distribution is so diffuse that its variance does not exist.

Hence we complete the specification of the Bayesian model. By jointly modeling $(\alpha_{j0}, \alpha_{j1})$ and (β_{j0}, β_{j1}), we properly account for correlation between pre- and post-intervention ratings. Although the effect of such correlation on the inference of ICC is theoretically intractable, it is appropriately channeled through MCMC simulation to the Bayesian estimation and comparison of ICCs. Because separate priors are specified on each component of the variance matrices in (2.3) and (2.4), we call it the "individual components" model.

Alternatively, we can re-parameterize $(\sigma^2_{\alpha 0}, \sigma^2_{\alpha 1}, \rho_\alpha)$ and $(\sigma^2_{\beta 0}, \sigma^2_{\beta 1}, \rho_\beta)$ as covariance matrices and employ the inverse Wishart prior (Zhang, 2021). Define

$$\mathbf{\Sigma}_\alpha = \begin{pmatrix} \sigma^2_{\alpha 0} & \rho_\alpha \sigma_{\alpha 0} \sigma_{\alpha 1} \\ \rho_\alpha \sigma_{\alpha 0} \sigma_{\alpha 1} & \sigma^2_{\alpha 1} \end{pmatrix} \text{ and } \mathbf{\Sigma}_\beta = \begin{pmatrix} \sigma^2_{\beta 0} & \rho_\beta \sigma_{\beta 0} \sigma_{\beta 1} \\ \rho_\beta \sigma_{\beta 0} \sigma_{\beta 1} & \sigma^2_{\beta 1} \end{pmatrix},$$

and assume $\mathbf{\Sigma}_\alpha \sim IW(\mathbf{\Omega}, \nu)$ and $\mathbf{\Sigma}_\beta \sim IW(\mathbf{\Omega}, \nu)$, where $IW(\mathbf{\Omega}, \nu)$ denotes an inverse Wishart distribution with a degree of freedom ν and mean $\mathbf{\Omega}/(\nu - 2 - 1)$. Following consideration similar to that in the "individual components" model, we set $\nu = 4$ and

$$\mathbf{\Omega} = \begin{pmatrix} 0.333 & 0 \\ 0 & 0.333 \end{pmatrix}.$$

Under such specification the prior means of $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$ equal $\begin{pmatrix} 0.333 & 0 \\ 0 & 0.333 \end{pmatrix}$ and their variances do not exist (Von Rosen, 1988). The other components of the Bayesian model remain unchanged. We refer to this approach as the "inverse Wishart" model. Under this approach, $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$ are sampled as random matrices from the posterior distribution. Let $\mathbf{\Sigma}^{(q)}_\alpha$ and $\mathbf{\Sigma}^{(q)}_\beta$ be the posterior samples obtained at the qth $(q = 1, \ldots, Q)$ iteration of MCMC simulation, then $\left(\sigma^{2(q)}_{\alpha 0}, \sigma^{2(q)}_{\alpha 1}\right)$ and $\left(\sigma^{2(q)}_{\beta 0}, \sigma^{2(q)}_{\beta 1}\right)$ are extracted from the diagonal elements of $\mathbf{\Sigma}^{(q)}_\alpha$ and $\mathbf{\Sigma}^{(q)}_\beta$, respectively, and plugged into Equation (2.5) to calculate $\text{ICC}^{(q)}_0$ and $\text{ICC}^{(q)}_1$.

Because the inverse gamma and inverse Wishart distributions have distinct properties, we have not identified a specification of the "inverse Wishart" model under which the prior distributions of $(\sigma^2_{\alpha 0}, \sigma^2_{\alpha 1}, \sigma^2_{\beta 0}, \sigma^2_{\beta 1})$ are equivalent to those under the "individual components" model. Nevertheless, we believe the "individual components" model and the "inverse Wishart" model specified here are comparable in key respects, as they share two features: 1) for $t = 0, 1$ the prior mean of $(\sigma^2_{\alpha t} + \sigma^2_{\beta t} + \sigma^2_{\epsilon t})$ equals to 1; 2) weakly informed priors are assumed for $\sigma^2_{\alpha 0}, \sigma^2_{\alpha 1}, \sigma^2_{\beta 0}, \sigma^2_{\beta 1}, \sigma^2_{\epsilon 0}, \sigma^2_{\epsilon 1}$ under which the prior variances do not exist.

The "inverse Wishart" model is appealing for its ease of coding and computational convenience, since inverse-Wishart priors are conjugate to normal models. In contrast, the "individual components" model allows separate prior specification for $(\sigma^2_{\alpha 0}, \sigma^2_{\alpha 1}, \rho_\alpha, \sigma^2_{\beta 0}, \sigma^2_{\beta 1}, \rho_\beta)$, offering greater flexibility in incorporating prior information. For example, assuming degenerate priors

that fix $\rho_\alpha = \rho_\beta = 0$ makes the "individual components" model equivalent to analyzing pre- and post-intervention ratings independently—an extension that is difficult to achieve directly within the "inverse Wishart" framework. We have conducted extensive simulation to assess performance of these two approaches.

3 Simulations

We conduct extensive simulation to evaluate performance of the proposed Bayesian method in assessing inter-rater agreement. The research question is whether by properly accounting for correlation between ordinal ratings in pre-post studies, we can improve the estimation and comparison of ICCs. Besides the "individual components" model and the "inverse Wishart" model, two additional models are considered: the "unpaired" model which is a degenerative version of the "individual components" model with correlation parameters fixed at 0, i.e., $\rho_\alpha = \rho_\beta = 0$; and the "numeric" model which treats the observed ordinal ratings as if they were continuous. That is, we directly apply the linear mixed effect model (2.2) to y_{ijt}'s.

The simulation adopts a setting similar to that of the motivating example, where $N = 106$ raters score $J = 14$ images by $K = 5$ ordinal categories. The true values of variance components are specified as

$$\begin{aligned} \sigma^2_{\alpha 0} = 0.5, \quad \sigma^2_{\beta 0} = 0.3, \quad \sigma^2_{\epsilon 0} = 0.2, \\ \sigma^2_{\alpha 1} = 0.7, \quad \sigma^2_{\beta 1} = 0.2, \quad \sigma^2_{\epsilon 1} = 0.1, \end{aligned}$$

and correlation parameters $\rho_\alpha = 0.7$ and $\rho_\beta = 0.6$. The cutoffs $\mathbf{c} = (c_1, c_2, c_3, c_4)$ are set at the 20th, 40th, 60th, and 80th percentiles of the $N(0, 1)$ distribution. Note that the variance components sum to 1 within each period and the implied true ICCs are $\text{ICC}_0 = 0.5$ and $\text{ICC}_1 = 0.7$. The simulation proceeds as follows:

1. Given the true values of parameters, generate $(\alpha_{j0}, \alpha_{j1})'$ for $j = 1, \ldots, J$ and $(\beta_{i0}, \beta_{i1})'$ for $i = 1, \ldots, N$ from (2.3) and (2.4), respectively.
2. The latent variable x_{ijt}'s are generated following the linear mixed effect model (2.2).
3. The ordinal ratings $\mathbf{Y} = \{y_{ijt}\}$ are obtained following (2.1)
4. Feed dataset $\mathbf{Y}$ to each of the four Bayesian models ("individual components", "inverse Wishart", "unpaired", "numeric"), conduct MCMC simulation, obtain the posterior estimates of means and standard deviations of ICCs, as well as the probability of improvement in inter-rater agreement, $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$.
5. Repeat Steps 1–4 for $M = 100$ times. Obtain the overall mean, bias, and rMSE (root of mean squared error) for the Bayesian estimators of ICC_0, ICC_1, as well as the overall mean and standard deviation of $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$.

The simulation results are summarized in Table 1. Each row corresponds to one of the models evaluated, where "Ind Comp" denotes the "individual components" model and "IW" the "inverse Wishart" model. Columns "ICC_t Mean" and "ICC_t rMSE" report the overall mean and rMSE of the Bayesian estimator of ICC_t $(t = 0, 1)$ across 100 simulations. Columns "P.P Mean" and "P.P SD" report the overall mean and standard deviation of the estimated posterior probability $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$. Several observations can be drawn from Table 1. First, the performances of the "individual components" and "inverse Wishart" models are comparable in terms of their accuracy in estimating the ICCs and $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$. The small observed differences are generally negligible relative to the random variability inherent in simulation. Second, when the correlation between pre- and post-intervention observations is ignored,

Table 1: Simulation: performance of estimated ICCs and $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ based on correlated data.

Model	ICC_0 Mean	ICC_1 Mean	ICC_0 rMSE	ICC_1 rMSE	P.P Mean	P.P SD
Ind Comp	0.460	0.651	0.074	0.080	0.934	0.109
IW	0.457	0.654	0.075	0.079	0.937	0.097
Unpaired	0.440	0.634	0.084	0.090	0.883	0.101
Numeric	0.415	0.605	0.096	0.104	0.930	0.105

the "unpaired" model performs substantially worse than the "individual components" model. The ICCs are estimated with less accuracy: the rMSE increases from 0.074 to 0.084 for ICC_0 and from 0.08 to 0.09 for ICC_1. More importantly, the estimated probability of improvement in inter-rater agreement, $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$, decreases on average by 0.051 (from 0.934 to 0.883). Since estimated posterior probabilities form the basis in Bayesian hypothesis testing, an underestimated $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ under the "unpaired" model indicates reduced power to detect intervention effects on inter-rater agreement. The loss of power when ignoring correlation in pre–post studies is well documented in the clinical trial literature (Ahn et al., 2014; Zhang et al., 2018), where hypotheses often concern means (continuous outcomes), proportions (binary outcomes), or hazards (survival outcomes). Our simulations extend this conclusion to settings where the outcome of interest is inter-rater agreement. Finally, the ICCs estimated under the "numeric" model are severely biased, underscoring the importance of respecting the ordinal nature of observed ratings and employing an appropriate model. Interestingly, because the "numeric" model does not ignore pre–post correlation, its estimates of $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ are close to those obtained from the "individual components" and "inverse Wishart" models. Nonetheless, due to the incompatibility between model and data, we regard the probabilities estimated by the "numeric" model as untrustworthy.

Compared with the simulation truths ($\text{ICC}_0 = 0.5$ and $\text{ICC}_1 = 0.7$), the Bayesian estimates exhibit a noticeable downward bias. Two factors contribute to this bias. First, the prior means of ICCs are centered around 0.333. Taking ICC_0 for example, recall that the prior distributions of $\sigma^2_{\alpha 0}$, $\sigma^2_{\beta 0}$, and $\sigma^2_{\epsilon 0}$ are specified so that each has a prior mean of 0.333. By definition, $\text{ICC}_0 = \sigma^2_{\alpha 0}/(\sigma^2_{\alpha 0} + \sigma^2_{\beta 0} + \sigma^2_{\epsilon 0})$, so its prior distribution is centered around 0.333. Second, the small sample size in terms of the number of images increases the tendency of posterior ICC estimates to shrink toward their prior means. In our simulations, we assume 106 raters and 14 images, with ICC defined as the proportion of total variance due to between-image variability. Given only 14 images, even if the number of raters (106) and the total number of observed ratings ($14 \times 106 \times 2 = 2968$) are relatively large, the information about between-image variability in the data is limited. To verify this intuition, we conduct a second simulation study with all settings the same as those in Table 1, except that we reversed the sample sizes, assuming $N = 14$ graders and $J = 106$ images. The results are presented in Table 2. We observe a universal reduction in the biases and rMSEs of Bayesian ICC estimators across all models, although the total number of observed ratings (2968) remains unchanged. Equation (2.5) shows that the image variance component $\sigma^2_{\alpha t}$ appears in both the numerator and denominator of ICC_t, playing a more important role than $\sigma^2_{\beta t}$ and $\sigma^2_{\epsilon t}$ in the calculation of ICCs. Increasing the number of images directly improves the estimation of $\sigma^2_{\alpha t}$, which can more effectively improve the inference about ICCs than increasing the number of raters. This observation provides an important insight

Table 2: Simulation: performance of estimated ICCs and $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ with reversed sample sizes.

Model	ICC_0 Mean	ICC_1 Mean	ICC_0 rMSE	ICC_1 rMSE	P.P Mean	P.P SD
Ind Comp	0.491	0.693	0.039	0.042	0.985	0.023
IW	0.491	0.694	0.040	0.043	0.987	0.021
Unpaired	0.495	0.690	0.039	0.039	0.973	0.033
Numeric	0.456	0.646	0.057	0.064	0.990	0.022

Table 3: Simulation: performance of estimated ICCs and $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ based on independent case.

Model	ICC_0 Mean	ICC_1 Mean	ICC_0 rMSE	ICC_1 rMSE	P.P Mean	P.P SD
Ind Comp	0.477	0.650	0.086	0.080	0.838	0.206
IW	0.480	0.657	0.088	0.081	0.838	0.204
Unpaired	0.472	0.637	0.083	0.079	0.834	0.203
Paired Numeric	0.447	0.607	0.087	0.083	0.826	0.220

to researchers planning studies of inter-rater agreement: increasing the number of rated items is more effective than increasing the number of raters for improving inference about ICCs. In practice, cost considerations will inevitably influence the choice between the two strategies, but this result offers a valuable guiding principle.

We conduct a third simulation where the simulation truths are the same as those in Table 1 except that the true correlations are set at zero, $\rho_\alpha = \rho_\beta = 0$. That is, the simulated pre- and post-intervention observations are in truth independent. Our goal is to evaluate the robustness of the proposed Bayesian method when they are misapplied to data that are actually independent. The results are presented in Table 3. Under this setting the "unpaired" model is the true model, as corroborated by its rMSEs being the smallest among four models. Nevertheless, the "individual components" and "inverse Wishart" models' performances are only slightly inferior. For practical purposes, the differences in estimated $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ among the three models are negligible. This simulation provides reassurance to practitioners: when the "individual components" and "inverse Wishart" models are correctly applied to correlated data, they yield substantial gains in inference about ICCs, while in the unfortunate case that they are misapplied to independent data, the cost is minimal.

Under $\rho_\alpha = \rho_\beta = 0$, this simulation in effect represents a parallel randomized study where each arm involves a distinct set of 106 raters and a distinct set of 14 images. It is noteworthy that given the same group size (the same numbers of raters and images) and the same effect size (the same true values of ICC_0 and ICC_1), the posterior probability of detecting improvement in inter-rater agreement, $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$, under a parallel randomized design is 0.834, which is considerably smaller than that under a pre-post design (0.934). This observation suggests that when properly designed and analyzed, a pre- and post-intervention design can achieve greater power in detecting improvement in inter-rater agreement.

In the supplementary material, we have presented an additional scenario where there is no improvement in inter-rater agreement ($\text{ICC}_0 = \text{ICC}_1 = 0.5$). The estimated posterior probabilities $P(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y})$ are very close to 0.5, which is expected because we have assumed iden-

Table 4: Application to the motivating example.

Method	ICC_0 (SD)	ICC_1 (SD)	$P(ICC_1 > ICC_0 \mid \mathbf{Y})$
Ind Comp	0.252 (0.070)	0.560 (0.087)	>.999
IW	0.253 (0.071)	0.554 (0.087)	.999
Unpaired	0.245 (0.064)	0.532 (0.084)	0.996

tical priors for variance components pre- and post-intervention. Furthermore, the 95% credible intervals of ICCs and the difference between ICC_1 and ICC_0 were presented in the supplementary material.

4 Application to the Motivating Example

With the accuracy and robustness of the proposed Bayesian method well established, we apply the "individual components" and "inverse Wishart" models to the motivating example. A prospective validation study was conducted to assess whether a novel grading scale improved inter-rater agreement in scoring the degree of conjunctival inflammation on patients with OCP. 106 raters were recruited, who were instructed to score the clinical images of 14 OCP patients by 5 ordinal categories. Afterwards, the grading scale was presented to the raters, and they were asked to re-score the 14 images. Based on the empirical distribution of observed ratings, each of the cutoffs $\mathbf{c} = (c_1, \ldots, c_4)$ is assigned a truncated $N(0, 1)$ prior with $h_1 = 0.09$, $h_2 = 0.365$, $h_3 = 0.694$, $h_4 = 0.889$, and tuning parameter $\delta = 0.03$. The analysis results are presented in Table 4. For illustrative purposes, results of the "unpaired" model are also presented. Under the "individual components" model, the estimated ICC has a posterior mean (standard deviation) of 0.252 (0.07) before implementing the novel grading scale, and the ICC increases to 0.56 (0.087) after implementation. The posterior densities of ICC_0 and ICC_1 are presented in Figure 1. With $\hat{P}(ICC_1 > ICC_0 \mid \mathbf{Y}) > 0.999$, there is overwhelming evidence that the novel grading scale improved inter-rater agreement. Consistent with what we have observed in simulation, the results of the "inverse Wishart" model are very close to those of the "individual components" model. The "unpaired" model, however, underestimates the probability of the novel grading scale improving inter-rater agreement.

The traceplots of the variance components $(\sigma^2_{\alpha 0}, \sigma^2_{\beta 0}, \sigma^2_{\epsilon 0})$ are presented in Figure 2. They indicate good mixing of the MCMC chains for variance components on the latent scale, suggesting that our strategy of specifying the priors of cutoffs $(\mathbf{c})$ and variance components $(\sigma^2_{\alpha 0}, \sigma^2_{\beta 0}, \sigma^2_{\epsilon 0}, \sigma^2_{\alpha 1}, \sigma^2_{\beta 1}, \sigma^2_{\epsilon 1})$ to anchor the distributions of latent variables x_{ijt}'s around $N(0, 1)$ is effective in enhancing identifiability of the latent variables and improving the efficiency of MCMC simulation. As a result, the parameters of primary interest, ICC_0 and ICC_1, also show good mixing in Figure 3.

In the frequentist paradigm, Nelson and Edwards (2015) employed a generalized linear mixed model to assess inter-rater agreement of ordinal scores among multiple raters. A Kappa-like measure, K_{ma}, was proposed as the metrics of agreement. This approach does not accommodate the pre-post design and the agreement metrics is different from ICC. Nonetheless, we can analyze pre- and post-intervention data separately and report the 95% confidence intervals of K_{ma} for each period. Inference about improvement in agreement can be made by examining whether the two confidence intervals overlap. Such assessment is statistically valid although it

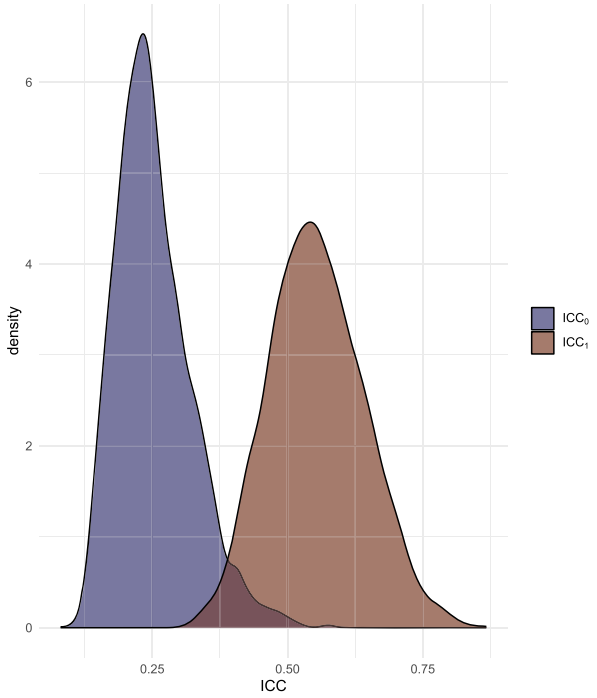
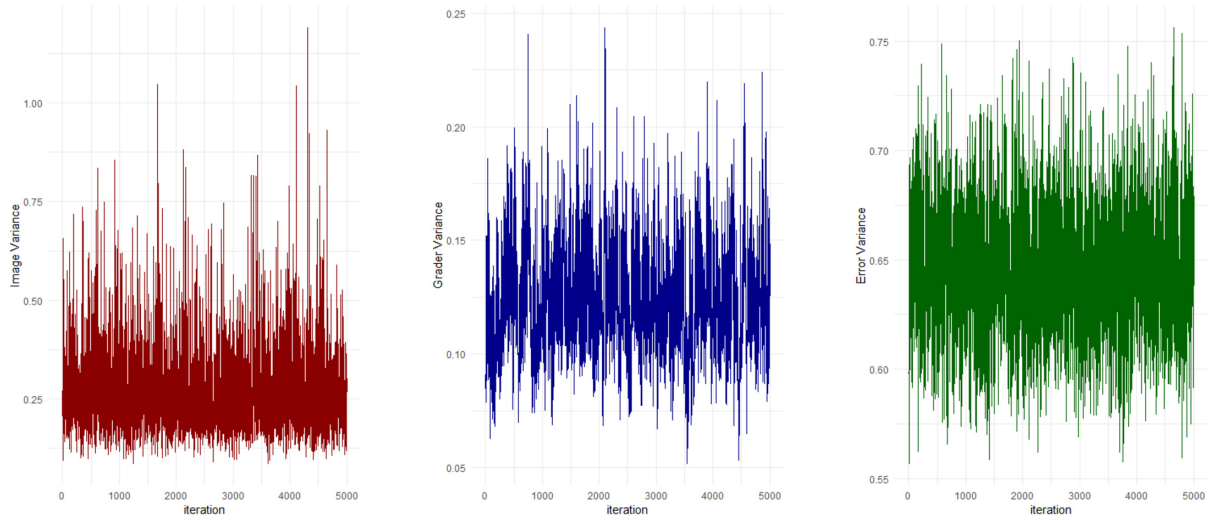


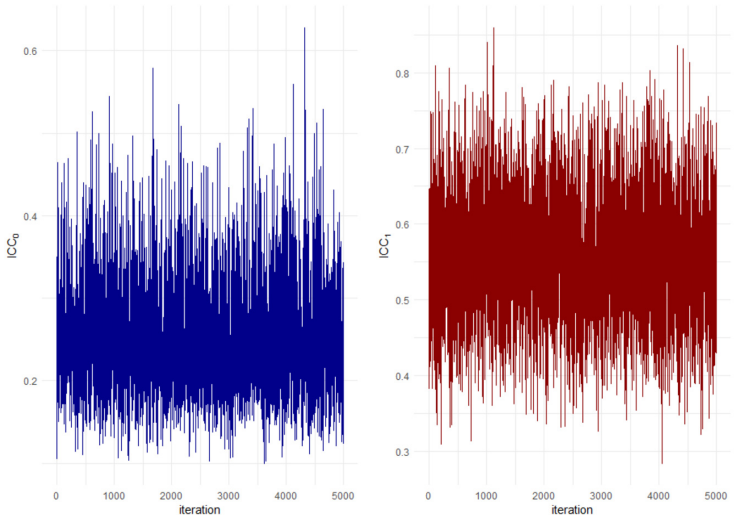
Figure 1: Posterior distributions of ICC_0 and ICC_1.



(a) (b) (c)

Figure 2: Traceplots of variance components $\sigma^2_{\alpha 0}, \sigma^2_{\beta 0}, and \sigma^2_{\epsilon 0}$.

loses power because pre-post correlation is ignored. We have conducted the analysis and the estimated K_{ma} in the pre- and post-training periods are (k_{ma} = 0.15, 95% CI 0.06–0.24) and (k_{ma} = 0.4, 95% CI 0.26–0.54), respectively. The two confidence intervals do not overlap, indicating significant improvement in agreement after the training. It is noteworthy that the 95%



(a) (b)

Figure 3: Traceplots of ICC_0 and ICC_1.

confidence intervals of k_{ma} are closely separated, suggesting marginal statistical significance. The proposed Bayesian method reports $\hat{P}(\text{ICC}_1 > \text{ICC}_0 \mid \mathbf{Y}) > 0.999$, indicating that by properly accounting for pre-post correlation, we can achieve greater power in detecting improvement in ICC.

5 Discussion

The pre–post design has been widely used to evaluate the effects of interventions. To date, however, no statistical methodology has been developed for pre–post studies where the effectiveness metric is inter-rater agreement based on ordinal ratings. This study introduces a Bayesian probit modeling framework for assessing pre–post change in inter-rater agreement when the rating scores are ordinal. By incorporating the correlation between pre- and post-intervention ratings, our approach improves estimation of ICCs and posterior probabilities of improvement, offering greater power compared to models that ignore such correlation. Simulations confirm that both the "individual components" and "inverse Wishart" models perform well, with negligible differences in practice. However, the "individual components" model affords greater flexibility in specifying priors for variances and correlations, making it more adaptable in applications where prior knowledge is available. In the motivating example every rater graded every image. The proposed Bayesian approach, however, naturally accommodates scenarios with incomplete observations where each rater only grades a random subset of images. Such scenarios are common in practice due to time and resource constraints. In such cases, the missing ratings arise by design rather than non-response, and can be considered as missing completely at random (MCAR). The proposed Bayesian model can handle such incomplete rating designs in the same fashion as frequentist mixed-effect models handle incomplete observations. The MCAR assumption ensures that the variance components and ICCs estimated based on observed data are valid, and that the missingness does not introduce bias in the estimation process.

Several insights emerge. First, treating ordinal ratings as continuous produces biased ICC estimates and unreliable inference, underscoring the need for ordinal-specific modeling. Second, study design strongly influences inference: increasing the number of items is more effective than increasing the number of raters in reducing bias and mean squared error for the assessment of ICCs. Third, the proposed Bayesian method properly account for pre–post correlation and improves inference about inter-rater agreement. Our simulation show that, however, even when the Bayesian method is misapplied to independent data, the resulting loss in accuracy is minimal. Together, these findings reassure practitioners that the proposed framework is robust and practically advantageous.

The application to grading of conjunctival inflammation in patients with OCP illustrates the clinical relevance of this methodology, providing strong evidence that a visual analog grading scale improves rater reliability. Our future research interests include evaluating the effects of rater and image characteristics on inter-rater agreement, exploring alternative prior specification for ICCs, and accounting for hierarchical study designs. Such extensions will further enhance the utility of Bayesian methods for evaluating inter-rater agreement.

Supplementary Material

The supplementary material includes supplementary tables and R codes.

References

Ahn C, Heo M, Zhang S (2014). *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. CRC Press.

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679. https://doi.org/10.1080/01621459.1993.10476321

Atenafu EG, Hamid JS, To T, Willan AR, M Feldman B, Beyene J (2012). Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. *BMC Medical Research Methodology*, 12(126): 1–8. https://doi.org/10.1186/1471-2288-12-126

Calle-Alonso F, Perez Sanchez CJ (2015). A Monte Carlo–based Bayesian approach for measuring agreement in a qualitative scale. *Applied Psychological Measurement*, 39(3): 189–207. https://doi.org/10.1177/0146621614554080

Cohen J (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46. https://doi.org/10.1177/001316446002000104

Cohen J (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4): 213–220. https://doi.org/10.1037/h0026256

Eziama E, Nguyen C, Foster CS, Heydinger S, Cao JH (2025). Novel grading scale for conjunctival inflammation in cicatrizing conjunctivitis associated with pemphigoid. *Ocular Immunology and Inflammation*, 33(4): 649–653. https://doi.org/10.1080/09273948.2024.2434128

Fanshawe TR, Lynch AG, Ellis IO, Green AR, Hanka R (2008). Assessing agreement between multiple raters with missing rating information, applied to breast cancer tumour grading. *PLoS ONE*, 3(8): e2925–e2936. https://doi.org/10.1371/journal.pone.0002925

Fisher RA (1921). On the "probable error"' of a coefficient of correlation deduced from a small sample. *Metron*, 1: 3–32.

Fleiss JL, Levin B, Paik MC (2013). *Statistical Methods for Rates and Proportions*. John Wiley & Sons.

Gajewski BJ, Hart S, Bergquist-Beringer S, Dunton N (2007). Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*, 26(25): 4602–4618. https://doi.org/10.1002/sim.2877

Gelman A, Carlin JB, Stern HS, Rubin DB (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Giraudeau B, Mary J (2001). Planning a reproducibility study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Statistics in Medicine*, 20(21): 3205–3214. https://doi.org/10.1002/sim.935

Hallgren KA (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1): 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Konishi S (1985). Normalizing and variance stabilizing transformations for intraclass correlations. *Annals of the Institute of Statistical Mathematics*, 37(1): 87–94. https://doi.org/10.1007/BF02481082

Müller R, Büttner P (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 13(23–24): 2465–2476. https://doi.org/10.1002/sim.4780132310

Nelson KP, Edwards D (2015). Measures of agreement between many raters for ordinal classifications. *Statistics in Medicine*, 34(23): 3116–3132. https://doi.org/10.1002/sim.6546

Olkin I, Lou Y, Stokes L, Cao J (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, 10(1): 4–30. https://doi.org/10.1017/jwe.2014.26

Shrout PE, Fleiss JL (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2): 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Tran QD, Demirhan H, Dolgun A (2021). Bayesian approaches to the weighted kappa-like inter-rater agreement measures. *Statistical Methods in Medical Research*, 30(10): 2329–2351. https://doi.org/10.1177/09622802211037068

Van Oest R, Girard JM (2022). Weighting schemes and incomplete data: A generalized Bayesian framework for chance-corrected interrater agreement. *Psychological Methods*, 27(6): 1069–1088.

Von Rosen D (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 15(2): 97–109.

Wang C, Yandell B, Rutledge J (1991). Bias of maximum likelihood estimator of intraclass correlation. *Theoretical and Applied Genetics*, 82(4): 421–424. https://doi.org/10.1007/BF00588594

Yue C, Chen S, Sair HI, Airan R, Caffo BS (2015). Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models. *Computational Statistics & Data Analysis*, 89: 126–133. https://doi.org/10.1016/j.csda.2015.02.012

Zhang S, Cao J, Ahn C (2018). Sample size calculation for before–after experiments with partially overlapping cohorts. *Contemporary Clinical Trials*, 64: 274–280. https://doi.org/10.1016/j.cct.2015.09.015

Zhang Z (2021). A note on Wishart and inverse Wishart priors for covariance matrix. *Journal of Behavioral Data Science*, 1(2): 119–126. https://doi.org/10.35566/jbds/v1n2/p2