Assessing Inter-rater Reliability With Heterogeneous Variance Components Models: Flexible Approach Accounting for Contextual Variables

Patrícia Martinková
Institute of Computer Science of the Czech Academy of Sciences
Charles University

František Bartoš
Institute of Computer Science of the Czech Academy of Sciences
University of Amsterdam

Marek Brabec
Institute of Computer Science of the Czech Academy of Sciences

Inter-rater reliability (IRR), which is a prerequisite of high-quality ratings and assessments, may be affected by contextual variables, such as the rater’s or ratee’s gender, major, or experience. Identification of such heterogeneity sources in IRR is important for the implementation of policies with the potential to decrease measurement error and to increase IRR by focusing on the most relevant subgroups. In this study, we propose a flexible approach for assessing IRR in cases of heterogeneity due to covariates by directly modeling differences in variance components. We use Bayes factors (BFs) to select the best performing model, and we suggest using Bayesian model averaging as an alternative approach for obtaining IRR and variance component estimates, allowing us to account for model uncertainty. We use inclusion BFs considering the whole model space to provide evidence for or against differences in variance components due to covariates. The proposed method is compared with other Bayesian and frequentist approaches in a simulation study, and we demonstrate its superiority in some situations. Finally, we provide real data examples from grant proposal peer review, demonstrating the usefulness of this method and its flexibility in the generalization of more complex designs.

Keywords: Bayesian inference; inter-rater reliability; mixed-effect models; heterogeneous variance components; grant peer review

1. Introduction

Inter-rater reliability (IRR) has been used to assess the quality of ratings and assessments in psychology, education, health, hiring, proposal and journal peer
review, and with other areas involving multiple raters. From a measurement perspective, individual ratings (such as scores applicants receive from a hiring committee) may be thought of as imprecise estimates of the true underlying quality of a measured subject or object. IRR enumerates the consistency among raters, and it may be described as the correlation between scores of different raters given to the same subject or object of measurement (Webb et al., 2006).

A notable portion of research is focused on the identification of heterogeneity sources in IRR with respect to contextual variables, such as rater or ratee characteristics, with the goal of identifying policies with the potential to generally decrease measurement error and to increase the IRR especially for the lower IRR subgroups. For example, IRR was found to vary for different research areas of grant-proposal peer review (Mutz et al., 2012) and to increase after reviewer training (Sattler et al., 2015). In the context of teacher hiring, IRR was found to be lower for internal than external applicants (Martinková et al., 2018) and lower for novice than experienced applicants (Goldhaber et al., 2021). In the context of teacher assessment, the IRR was found to be higher for the ratings of live versus recorded lectures (Casabianca et al., 2013).

Different estimation techniques were considered in the past to account for heterogeneity in IRR with respect to groups. The most common approach involves stratification of the data and separate estimation of IRR in subgroups with analysis of variance (ANOVA) or mixed-effect models (Sattler et al., 2015). More complex mixed-effect models allowing for heterogeneous variance components were shown to detect group differences in IRR even in those cases where no difference was detected using the stratification approach (Martinková et al., 2018). Another method is based upon generalized-estimation equations (Mutz et al., 2012). Bartoš et al. (2020) compared different methods for IRR estimation under heterogeneity with respect to a single grouping variable in a simulation study, where the data-generating model was known, showing that both frequentist and Bayesian mixed-effect models, as well as general additive models, can provide accurate estimates of group-dependent IRR.

However, further methodological complexities arise in real-life situations, which were not solved by previous studies. The variance components of the ratings may be affected by a combination of contextual factors, such as the rater’s or ratee’s age, gender, major, or internal versus external status, which may be of different types (besides binary also nominal, ordinal, or metric). Furthermore, the model specification—inclusion or exclusion of the different contextual factors—might need to be inferred from the data. Researchers might also be interested in whether a particular contextual factor does or does not affect the variance components, that is, testing a hypothesis whether the effect of a given factor differs from zero. To our best knowledge, there has been no study estimating IRR or reliability with heterogeneous variance mixed-effects models in cases of heterogeneity due to a combination of covariates. None have dealt with model selection, nor is any general approach available.
To fill this existing research gap, we propose a flexible general approach to IRR estimation and hypothesis testing using Bayes factors (BF) in those cases of variance heterogeneity due to covariates and unknown data-generating models. Our work builds upon studies of Bayesian mixed-effects models with heterogeneous variance components (Williams et al., 2020; Williams et al., 2019), which were previously shown to provide a richer understanding of the psychological processes in various contexts, and upon the work of Dablander et al. (2020), who recently introduced a default BF test for the inequality of variances. We also consider model-averaged estimates, which incorporate uncertainty in the model selection process for the final estimate (Depaoli et al., 2020; Hoeting et al., 1999; Raftery, 1996; Raftery et al., 1995).

This article proceeds as follows: We first introduce the IRR in the multilevel modeling framework. We extend it to heterogeneous variance components and introduce Bayesian hypothesis testing and model averaging in Section 2. Second, we describe a simulation study and compare the proposed methodology to alternative approaches in Section 3. Third, we illustrate the methodology on real data sets from ratings of the grant proposals in Section 4. Finally, in Section 5, we conclude with a discussion of the results and of further computational aspects of IRR estimation, including the aspects of generalizations in more complex designs. Sample R code and additional tables and figures are provided in electronic Supplementary Material at https://osf.io/bk8a7/.

2. Methods

2.1. IRR in Multilevel Modeling Framework

In the simplest case of a multilevel linear model with a one-way ANOVA, rating \( j \) of subject \( i \), denoted \( Y_{ij} \), is modeled as

\[
Y_{ij} = \mu + \gamma_i + \epsilon_{ij},
\]

where \( \mu \) is the grand mean rating, \( \gamma_i \) is the ratee-specific deviation from the grand mean—the random intercept that together with \( \mu \) represents the ratees’ true ratings—with structural variance \( \sigma^2_\gamma \), and finally, \( \epsilon_{ij} \) is the random error of the rating with a residual variance \( \sigma^2_\epsilon \). For the sake of estimation, it is standard to assume independent and identically distributed (IID) \( \gamma_i \sim N(0, \sigma^2_\gamma) \) and \( \epsilon_{ij} \sim N(0, \sigma^2_\epsilon) \), and random errors \( \epsilon_{ij} \) to be uncorrelated with the ratee effects \( \gamma_i \). Note that the assumption of IID error implies that there are different raters at each rating or that the rater effect is neglected.

Under the model defined by Equation 1, IRR is defined as the ratio of the true variance due to ratees, \( \sigma^2_\gamma \), to the total variance \( \sigma^2_\gamma + \sigma^2_\epsilon \), that is,

\[
\text{IRR} = \frac{\sigma^2_\gamma}{\sigma^2_\gamma + \sigma^2_\epsilon},
\]
corresponding to the intraclass correlation coefficient denoted as ICC(1,1) (see McGraw & Wong, 1996; Shrout & Fleiss, 1979, for more details and further possibilities).

The variance components in IRR defined by Equation 2 can be estimated using various frequentist and Bayesian approaches to provide an estimate of IRR. The maximum likelihood (ML) estimates are found as parameters maximizing the likelihood function given the data (Searle, 1997). The REstricted (or REsidual) ML (REML) method is an adaptation, in which the marginal likelihood function is maximized with respect to variance components \((\sigma^2_y, \sigma^2_e)\), while the third parameter \(\mu\) is integrated out of the likelihood function. For a balanced design of the model defined by Equation 1, that is when the same number of ratings is given to all ratees, the REML estimates of variance components are identical to those from the one-way ANOVA method of moments (Searle et al., 2006).

Finally, the Bayesian estimation (e.g., Gelman & Hill, 2006) starts with specifying prior distributions for all model parameters (the variance components \(p(\sigma^2_y, \sigma^2_e)\) and mean \(p(\mu)\)). The posterior distribution is then obtained through the Bayes rule, by multiplying the prior distributions by the likelihood and standardizing by the probability of data (i.e., the marginal likelihood).

2.1.1. Incorporating heterogeneous variance components. The multilevel model defined by Equation 1 can be further generalized. The generalization suggested here involves the variance terms, possibly together with the mean, depending on covariates. For \(Y_{ij}\) being the rating \(j\) of subject \(i\), we assume the following multilevel model:

\[
Y_{ij} = \mu_i + \gamma_i + \epsilon_{ij},
\]

where the mean \(\mu_i\) is modeled as a regression on covariates \(x_i\)

\[
\mu_i = \alpha_{\mu} + \beta_{\mu}^T x_i.
\]

\(\gamma_i\) is a random effect of subject \(i\), and \(\epsilon_{ij}\) is a random error term for rating \(j\) on subject \(i\) as in Equation 1. We moreover allow the covariates to influence the variance terms. In other words, we assume that \(\gamma_i \sim N(0, \sigma_{\gamma i}^2)\) and \(\epsilon_{ij} \sim N(0, \sigma_{\epsilon ij}^2)\) with subject-specific variance terms \(\sigma_{\gamma i}^2\) and \(\sigma_{\epsilon ij}^2\) being modeled as a regression on possibly a different set of covariates \(u_i\) and \(v_i\):

\[
\sigma_{\gamma i} = \sigma_{\gamma} e^{\beta_{\gamma}^T u_i},
\]

\[
\sigma_{\epsilon ij} = \sigma_{\epsilon} e^{\beta_{\epsilon}^T v_i}.
\]

In this equation, \(\beta_{\epsilon}\) are linear effects attributed to covariates \(u_i\) explaining the variability in variance terms \(\sigma_{\epsilon ij}\), while \(\beta_{\gamma}\) are linear effects attributed to covariates \(v_i\) explaining the variability in variance terms \(\sigma_{\gamma i}\). The logarithmic link
transforms the linear predictor into a multiplicative factor of the square root of the variance components’ grand means, ensuring that the resulting variance is positive, given that the grand means of the variance components are positive.

Under a model defined by Equation 3, the IRR in Equation 2 from Section 2.1 is generalized to depend on the set of covariates $u_i$ and $v_i$:

$$\text{IRR}_i = \text{IRR}(u_i, v_i) = \frac{\sigma_{yi}^2}{\sigma_{yi}^2 + \sigma_{ei}^2} = \frac{\sigma_{yi}^2 e^{\beta_i^v v_i}}{\sigma_{yi}^2 e^{\beta_i^v v_i} + \sigma_{ei}^2 e^{\beta_i^e u_i} }.$$  \hspace{1cm} (6)

Note that when including covariates $x_i$ in the fixed part of the model defined in Equation 4, the estimates as well as the meaning of the variances in Equations 5 change, and the interpretation of IRR in Equation 6 changes as well. More specifically, by including overall effects of ratee characteristics in a model, the between-ratee variance will typically be reduced in some groups, and the group differences in $\sigma_i^2$ will become smaller. As a practical consequence, the interpretation of the IRR in such a case relates to an instance, whereby the final judgment takes into consideration the covariates. As an example, Goldhaber et al. (2021) assumed that hiring officials were likely to take the rater type into consideration when interpreting the professional references’ ratings of teacher applicants, and thus, their IRR estimates were adjusted for these sources of variation by including the rater type into the fixed part of the model. If, on the contrary, the final judgments are completed based solely upon the ratings, a simpler version of the model given by Equation 3 should be considered, in which $\mu$ is a constant, that is, $\beta_{\mu}$ in Equation 4 is restricted to 0.

Also note that we generally allow a possibly different set of covariates $x_i$, $u_i$, and $v_i$ to explain the means $\mu_i$, structural variances $\sigma_{yi}^2$, and residual variances $\sigma_{ei}^2$ in Equation 3, that we will use in the real data example in Section 4.3. However, in certain situations, a simpler and more restrictive model may be assumed, in which the same set of covariates is used, as we will do in our simulation study in Section 3 and in the real data examples in Sections 4.1 and 4.2.

2.2. Bayesian Hypothesis Testing and Model Averaging

With Equation 3 denoting the most complex model, a number of submodels can be considered as special cases based on restricting some of the effects in $\beta_\mu$, $\beta_\gamma$, and $\beta_e$ to zero. With a number of models to select from, we can select the best fitting model (as discussed in this subsection) and use parameter estimates for the best-fitting model for the final estimate of IRR. Alternatively, we can incorporate the uncertainty of the model selection process and calculate the model-averaged parameter estimates.

2.2.1. Bayes factors. We consider here the Bayesian hypothesis testing framework of Jeffreys (1931), which evaluates the evidence in support of/against any
model by the usage of BFs. BFs are computed as a ratio of the marginal like-
lihoods of the competing models (Etz & Wagenmakers, 2017; Kass & Raftery,
1995; Rouder & Morey, 2019; Wrinch & Jeffreys, 1921)

$$BF_{10} = \frac{p(\text{data}|M_1)}{p(\text{data}|M_0)},$$

(7)

with the marginal likelihood $p(\text{data}|M_m)$ quantifying the model’s $m$ relative
predictive performance by integrating the likelihood over the parameter space
(Jefferys & Berger, 1992).

The BF is a continuous measure of evidence in favor of $M_1$ and against $M_0.$
For ease of interpretation, we can label the resulting BFs as weak ($BF_{10}$ between
1 and 3), moderate (between 3 and 10), strong (between 10 and 100), and very
strong (larger than 100; Jeffreys, 1939, Appendix I; Kass & Raftery, 1995; Lee &
Wagenmakers, 2013, p. 105).

2.2.2. Bayesian model averaging. In addition to model selection and using the
single best-fitting model for parameter estimation, we should consider Bayesian
model averaging (Hoeting et al., 1999; Kass & Raftery, 1995; Leamer, 1978).

Bayesian model averaging accounts for the uncertainty of model selection by
weighting the posterior model estimates by posterior model probabilities. First,
we need to assign prior model probabilities $p(M_m)$ to the individual models $m$
and update them with the Bayes rule into posterior model probability
$p(M_m|\text{data})$ according to the Bayes rule (Fragoso et al., 2018; Hinne et al.,
2020; Hoeting et al., 1999)

$$p(M_m|\text{data}) = \frac{p(\text{data}|M_m) \times p(M_m)}{\sum_{m=1}^{M} p(\text{data}|M_m) \times p(M_m)}.$$  

(8)

We then combine the posterior parameter estimates $p(\theta|\text{data}, M_m)$ from the
$m = 1, \ldots, M$ individual models based on posterior model probabilities
$p(M_m|\text{data})$

$$p(\theta|\text{data}) = \sum_{m=1}^{M} p(\theta|M_m, \text{data}) \times p(M_m|\text{data}),$$

(9)

which allows us to acknowledge the uncertainty about the considered models.
We follow a common convention in Bayesian model averaging and assign an
equal prior model probability to models assuming the absence and presence of
the difference between the groups for either the mean, structural, or residual
variance, resulting in $p(M_m) = 1/M$ (Gronau et al., 2021; Kass & Raftery,
1995; Madigan et al., 1994; Raftery et al., 1995).

Furthermore, Bayesian model averaging allows us to quantify evidence in
favor of including a specific parameter across the whole set of specified models
with a comparable structure. For example, for a difference between the groups in
a residual variance $\sigma^2_{\varepsilon}$, the BF from Equation 7 is extended into inclusion BF (Gronau et al., 2021; Hinne et al., 2020):

$$
\text{BF}_{\varepsilon \varepsilon} = \frac{\sum_{a \in A} p(M_a | \text{data})}{\sum_{b \in B} p(M_b | \text{data})} \frac{\sum_{a \in A} p(M_a)}{\sum_{b \in B} p(M_b)},
$$

where $A$ represents a set of models for which the groups differ in the $\sigma^2_{\varepsilon}$ parameter, and $B$ represents a set of models for which they do not differ.

2.2.3. Parametrization and choice of priors. To employ the Bayesian framework consisting of BFs and Bayesian model averaging, we need to complete the models by specifying prior distributions for all the parameters in Equations 4 and 5. Here, we restrict ourselves to consideration of binary covariates, testing for and quantifying the differences between groups (see Section 5 for suggestions on dealing with other types of covariates). We use effect coding (i.e., we assign the values of $-0.5$ and $0.5$ for the two levels), so the prior distribution on the regression coefficients $\beta$ corresponds to the difference (for the mean rating) or standard deviation ratio (for the structural and residual variances) between the groups. Consequently, the intercept parameters $\alpha$ represent the unweighted grand means, that is, they have common interpretation across all possible submodels.

For the simulation study and the real-data example, we use the following priors:

$$
\alpha_{\mu} \sim \text{Normal}(0, 1),
$$

$$
\alpha_{\gamma}, \alpha_{\varepsilon} \sim \text{Normal}_+(0, 1),
$$

$$
\beta_{\mu}, \beta_{\gamma}, \beta_{\varepsilon} \sim \text{Normal}(0, 0.5^2),
$$

where $\text{Normal}_+$ stands for the half-normal distribution, and

$$
\gamma_i \sim \text{Normal}(0, \sigma^2_{\mu}),
$$

$$
y_{ij} \sim \text{Normal}(\mu_i + \gamma_i, \sigma^2_{\varepsilon i}),
$$

where $\mu_i$ is defined by Equation 4 with $x = -0.5$ for the first group and $x = 0.5$ for the second group, and $\sigma^2_{\mu i}$ and $\sigma^2_{\varepsilon i}$ are defined by Equation 5 for each participant $i$.

Our reasoning behind the choice of priors is as follows: Since the intercept parameters are common across all models (i.e., we are not going to test for the presence or absence of the intercept), we can specify weakly informative prior distributions on them (Gelman & Hill, 2006). Here, we use standard normal prior
distribution for the grand mean intercept, $\alpha_\mu \sim N(0, 1)$ and half normal prior distributions for the structural and residual standard deviation intercepts, $\alpha_\gamma, \alpha_\varepsilon \sim N_+(0, 1)$. This setting corresponds to the expectation that the outcome variable is somewhat standardized, that is, the grand mean is located around zero and the overall variance of the data is around one. If the outcome variable corresponded to a differently scaled measure, we would adjust the means and standard deviations of the prior distributions to reflect the overall expectations (e.g., we could use $\alpha_\mu \sim N(100, 15^2)$ and $\alpha_\gamma, \alpha_\varepsilon \sim N_+(0, 15^2)$ if we were working with IQ scores).

In contrast to the common intercepts, the regression parameters $\beta$ can differ between the submodels (omitting a predictor equals to setting the corresponding $\beta = 0$). Subsequently, the prior distribution on the regression parameters defines the hypothesis test for the presence or the absence of the effect for a given predictor. Here, we use informed Normal $(0, \sigma^2)$ prior distributions on the regression coefficients, where $\sigma^2$ parameter controls informativeness (i.e., deviations from the null hypotheses we are interested in) of the test. This corresponds to specifying a two-sided hypothesis on the regression parameters for the means and standard deviations. In our view, the choice of $\sigma^2 = 0.5^2$ used in the simulation study and the real data example corresponds to testing for “medium-sized” differences in means and standard deviation ratios (i.e., mean differences lower than 1 and standard deviation ratios lower than 2.7).

To assess the robustness of our results to the prior distribution specifications, we use two other choices of $\sigma^2$ in the real data example. In our view, the choices of $\sigma^2 = 0.25^2$ and $\sigma^2 = 1^2$ correspond to “small-sized” and “large-sized” differences in means and standard deviation ratios, respectively. See Figure 1 for the considered resulting prior distributions of the mean differences (left panel) and the resulting prior distributions of the ratios of standard deviations (right panel) obtainable by taking the exponent of the prior distribution.

### 3. Simulation Study

We perform a simulation study to assess the performance of the outlined methodology. We are specifically interested in an estimation and hypothesis testing in consideration to the differences between groups in any of the modeled parameters (i.e., means, structural standard deviations, and residual standard deviations) and IRR. We keep simulation settings simple, in order to compare the outlined methodology to other model selections and model-averaging techniques, considering that many of the other methods could not deal with more complex data settings.

For the simulation, we consider IRR in a group-specific variance components model; in other words, we assume a single binary covariate, the group
membership denoted by index $g \in \{0, 1\}$. For $Y_{ijg}$ being the rating $j$ of subject $i$ from group $g$, the original model defined by Equation 3 simplifies to

$$Y_{ijg} = \mu_g + \gamma_{ig} + \varepsilon_{ijg},$$

where $\mu_g$ is the group-specific mean rating, $\gamma_{ig} \sim N(0, \sigma^2_{\gamma g})$ is the ratee-specific deviation from the group mean with a group-specific (structural) variance $\sigma^2_{\gamma g}$, and finally, $\varepsilon_{ijg} \sim N(0, \sigma^2_{\varepsilon g})$ is the random error of the rating with group-specific residual variance $\sigma^2_{\varepsilon g}$. We assume normal distributions.

Under the model specified by Equation 12, the group-specific IRR$_g$ is then defined in the special case of Equation 6 as

$$\text{IRR}_g = \frac{\sigma^2_{\gamma g}}{\sigma^2_{\gamma g} + \sigma^2_{\varepsilon g}},$$

and it takes the two values of IRR$_0$ and IRR$_1$ depending upon the group, which is the only covariate assumed in this design.

Possible submodels of the model defined by Equation 12 are derived as special cases based upon restricting the group-specific parameters under a combination of conditions:

A. $\mu_0 = \mu_1 = \mu$,

B. $\sigma^2_{\gamma 0} = \sigma^2_{\gamma 1} = \sigma^2_{\gamma}$,

C. $\sigma^2_{\varepsilon 0} = \sigma^2_{\varepsilon 1} = \sigma^2_{\varepsilon}$. 

FIGURE 1. Visualization of different options of prior distributions on the regression coefficients. We used the Normal ($0, 0.5^2$) prior distributions (in bold) for the simulation and the real data example and the remaining options as a robustness check. The left panel visualizes the resulting prior distribution for the mean differences and the right panel for standard deviation ratios.
Altogether, the combination of conditions A, B, and C leads to seven possible submodels denoted as M1 through M7, and one unrestricted model (Equation 12) denoted as M8:

\[
\begin{align*}
\text{M1: } & \quad [ABC] \quad Y_{ijg} = \mu + \gamma_i + \epsilon_{ijg}, \\
\text{M2: } & \quad [AB\overline{C}] \quad Y_{ijg} = \mu + \gamma_i + \epsilon_{ijg}, \\
\text{M3: } & \quad [A\overline{BC}] \quad Y_{ijg} = \mu + \gamma_{ig} + \epsilon_{ij}, \\
\text{M4: } & \quad [\overline{A}BC] \quad Y_{ijg} = \mu + \gamma_{ig} + \epsilon_{ijg}, \\
\text{M5: } & \quad [\overline{A}B\overline{C}] \quad Y_{ijg} = \mu_g + \gamma_i + \epsilon_{ij}, \\
\text{M6: } & \quad [\overline{A}\overline{B}C] \quad Y_{ijg} = \mu_g + \gamma_{ig} + \epsilon_{ij}, \\
\text{M7: } & \quad [\overline{A}\overline{B}\overline{C}] \quad Y_{ijg} = \mu_g + \gamma_{ig} + \epsilon_{ijg}. \\
\end{align*}
\]

### 3.1. Data Generation

Data generation was inspired by real data encountered in the context of teacher applicant ratings (Martinková et al., 2018). Specifically, in Equation 12, we used two values for the standardized mean differences between the groups \((\mu_2 - \mu_1 = 0 \text{ or } 0.4)\), for the structural variance ratios \((\sigma^2_{\gamma_1}/\sigma^2_{\gamma_2} = 1 \text{ or } 1.5)\), and for residual variance ratios \((\sigma^2_{\epsilon_1}/\sigma^2_{\epsilon_2} = 1 \text{ or } 1.5)\), while we constrained the overall mean variance across groups to \(1/G\sum_{g=1}^{G} \sigma^2_{\gamma_g} + \sigma^2_{\epsilon_g} = 1\), and the mean IRR across groups to \(1/G\sum_{g=1}^{G} \text{IRR}_{rg} = 0.45\). This led to eight simulation scenarios, with scenarios 4 and 8 split into two subscenarios depending upon whether the structural and residual variance ratios differed in the same or the opposite direction (Table 1).

Moreover, we manipulated the number of times the ratees were rated \((J = 3 \text{ or } 5)\) and the number of ratees per group \((I = 25, 50, 100, \text{ or } 200)\) in each scenario. In total, 10 (scenarios including subscenarios) \(\times 2\) (number of ratings) \(\times 4\) (number of ratees) = 80 conditions were simulated, 1,000 times each, implying 80,000 randomly generated data sets.

### 3.2. Compared Methods

We compared the Bayesian hypothesis testing and model-averaging methodology outlined in Section 2 to alternative frequentist and Bayesian ways of estimating and testing for the differences in group-specific mixed-effects...
location scale models defined by Equation 12. Specifically, we used the ML and REML estimation in the frequentist framework for linear mixed models and Markov chain Monte Carlo estimation in the Bayesian framework.

3.2.1. Model selection. To assess the performance of the model selection with BFs specified in Section 2.2.1, we considered four frequentist model selection approaches and two Bayesian approaches.

The frequentist approaches include two stepwise selection procedures (backward and forward) and two model space selection procedures based upon Akaike information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978). In the forward stepwise selection procedure, we started with the simplest model. We first tested for adding the difference in means (with REML) and then gradually tested expanding the model with differences in structural and/or residual variances (with ML; see the left panel of Figure A1 in Appendix A for diagram). In the backward stepwise selection procedure, we started with the most complex model. We first tested for gradually removing differences in the structural and/or residual variances (with ML) and then tested for removing the difference in means (with REML; see the right panel of Figure A1 in Appendix A for the diagram). In the model selection procedures based upon information criteria, we estimated all the specified models (with ML) and subsequently selected the best-fitting model based on the lowest information criteria (AIC or BIC).

The Bayesian approaches include two model space selection procedures based upon posterior predictive performance—the Watanabe-AIC (WAIC; Watanabe & Opper, 2010) and leave-one-out (LOO) cross-validation (Vehtari et al., 2017). WAIC and LOO approximate the LOO prediction error using the log-likelihood evaluated with posterior parameter distribution (McElreath, 2018). While they

# Table 1

Data Generation Scenarios

| Scenario | $\mu_1$ | $\mu_2$ | $\sigma_{\gamma 1}$ | $\sigma_{\gamma 2}$ | $\sigma_{e 1}$ | $\sigma_{e 2}$ | IRR$_1$ | IRR$_2$ |
|----------|---------|---------|----------------------|----------------------|----------------|----------------|--------|--------|
| 1        | .00     | .00     | .67                  | .67                  | .74            | .74            | .45    | .45    |
| 2        | .00     | .00     | .67                  | .67                  | .67            | .82            | .50    | .40    |
| 3        | .00     | .00     | .60                  | .74                  | .74            | .74            | .40    | .50    |
| 4.1      | .00     | .00     | .60                  | .73                  | .66            | .81            | .45    | .45    |
| 4.2      | .00     | .00     | .73                  | .60                  | .66            | .81            | .55    | .35    |
| 5        | -.20    | .20     | .67                  | .67                  | .74            | .74            | .45    | .45    |
| 6        | -.20    | .20     | .67                  | .67                  | .67            | .82            | .50    | .40    |
| 7        | -.20    | .20     | .60                  | .74                  | .74            | .74            | .40    | .50    |
| 8.1      | -.20    | .20     | .60                  | .73                  | .66            | .81            | .45    | .45    |
| 8.2      | -.20    | .20     | .73                  | .60                  | .66            | .81            | .55    | .35    |

Note. IRR = inter-rater reliability.
are asymptotically equivalent, LOO usually performs better in small samples and under weak prior distributions (Vehtari et al., 2017). In our case, we use the individual ratings $Y_{ij}$ as the basis of LOO predictions.\(^1\)

### 3.2.2. Model averaging
To assess the performance of the parameter estimation with Bayesian model averaging, we used two frequentist and two alternative Bayesian approaches.

The frequentist methods combine the estimates $\hat{\theta}_m$ from $M$ individual models with weights $\omega_m$:

$$\hat{\theta} = \sum_{m=1}^{M} \hat{\theta}_m \times \omega_m. \quad (14)$$

The two frequentist approaches are based on information criteria (AIC and BIC), and specify the weight $\omega_m$ for model $m$ as

$$\omega_{\text{AIC},m} = \frac{\exp(-1/2 \Delta_m(\text{AIC}))}{\sum_{i=1}^{M} \exp(-1/2 \Delta_i(\text{AIC}))}, \quad (15)$$

$$\omega_{\text{BIC},m} = \frac{\exp(-1/2 \Delta_m(\text{BIC}))}{\sum_{i=1}^{M} \exp(-1/2 \Delta_i(\text{BIC}))},$$

with

$$\Delta_m(\text{AIC}) = \text{AIC}_m - \min(\text{AIC}),$$

$$\Delta_m(\text{BIC}) = \text{BIC}_m - \min(\text{BIC}),$$

where $\text{AIC}_m$ and $\text{BIC}_m$ correspond to the AIC and BIC value of the $m^{th}$ model and $\Delta_m(\text{AIC}_m)$ and $\Delta_m(\text{BIC}_m)$ correspond to the difference between the AIC or BIC, respectively, of the $m^{th}$ and the best-fitting model (Hjort & Claeskens, 2003; Wagenmakers & Farrell, 2004).

As an alternative Bayesian approach, the pseudo-Bayesian model averaging (BMA), similarly to the frequentist model averaging, uses information criteria to compute the model weights in Equation 15 (Geisser & Eddy, 1979; Gelfand, 1996). In contrast to Bayesian model averaging, it does not require specification of prior model probabilities, since the weights are based entirely on the LOO information criteria.

The last alternative approach, the Bayesian stacking of predictive distributions (Yao et al., 2018), is based upon stacking, which combines models in order to minimize LOO mean square error (MSE; Breiman, 1996; LeBlanc & Tibshirani, 1996; Wolpert, 1992). The stacking of posterior distribution is then based upon the LOO predictive distribution computed with LOO. Similar to pseudo-BMA,
Bayesian stacking does not require specification of prior model probabilities; however, it oftentimes does not allow inference about the true data structure and is unable to provide compelling evidence in favor of simple models (Gronau & Wagenmakers, 2019a, 2019b).

3.3. Evaluation of the Simulation Results

We first evaluate the proportion of selecting the correct model based upon BF$s and we compare it with other approaches. We evaluate the proportion of correct model selection averaged across all conditions and separately for each of the data generating model and sample size.

Next, we compare our approach with other approaches in the precision of estimates of model parameters and of IRR. As a measure of precision, we evaluate the root mean square error (RMSE). As a (relative) measure of bias, we also evaluate the bias$^2$/MSE ratio. This is again evaluated when averaged across all conditions, as well as for all the individual model generating designs.

Finally, we evaluate the calibration of our prior distributions and of inclusion BF$s with the so-called Bayes factor design analysis (see, e.g., Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019, for more details). Our goal for this simulation was (1) to verify that the inclusion BF$s found evidence in favor of the difference in parameters in those conditions, where the parameters differed and evidence in favor of no difference in conditions where the parameters do not differ, (2) to evaluate the proportion of misleading evidence, that is, how often would the inclusion BF$s find strong evidence in favor of a difference in scenarios with no difference present, and finally (3) to verify that the evidence is increasing with an increasing sample size.

3.4. Implementation

The simulation was carried out in R version 3.5.1. (R Core Team, 2018). We used nlme R package version 3.1 (Pinheiro et al., 2021) to estimate the frequentist version of the models, and we have written a custom Rstan model with the usage of rstan R package version 2.18.2 (Stan Development Team, 2018) for the Bayesian models. We further used the bridgesampling R package version 3.1 (Gronau et al., 2020) to compute the marginal likelihoods via bridge sampling (e.g., Gronau et al., 2017; Meng & Wong, 1996), and we used loo R package version 2.0 (Vehtari et al., 2020) to compute WAIC, LOO, and pseudo-BMA and stacking weights with the usage of Pareto-smoothed importance sampling (Vehtari et al., 2017).

3.5. Simulation Results

3.5.1. Model selection. We first take a look at the averaged results across all conditions. The probability of selecting the correct model is summarized
in Table 2. BFs, AIC, WAIC, and LOO are able to identify the correct model with precision around 26% of the time in the smallest group sizes \((N = 25)\); however, while BFs and AIC steadily improve with increasing sample sizes (up to 70% and 65% respectively), WAIC and LOO start lagging behind. Furthermore, the step-wise forward and backward selection procedures start catching up with the BFs with increasing sample sizes (69%).

Figure 2 displays model selection performance for individual data-generating designs. The first column shows the proportion of correctly selected models, the second column shows the proportion of selecting a more complex incorrect model (i.e., a model containing all true parameter differences plus some additional incorrect differences), and the last column shows the proportion of other incorrect models (models missing at least one parameter difference). We see a well-known behavior of BIC being biased toward the simpler model, resulting in “better” performance under data-generating scenario 1, and a bias of WAIC and LOO toward more complex models, resulting in a lower proportion of correct model selection and increased selection of incorrect more complex models in the simpler data-generating designs. The same trends are visible in the case of \(j = 5\) ratings per rate (see the electronic Supplementary Material).

### 3.5.2. Parameter estimation.

The RMSE of the residual SD estimates, averaged across all conditions, is summarized in Table 3. We can see that with small samples (e.g., \(N = 25\)), the model averaging leads to more precise estimates of residual variance than the estimates based upon the selected best-performing model. For example, Bayesian parameter estimation in the models selected with BFs resulted in RMSE of 0.075, while the Bayesian model averaging resulted in RMSE of 0.069. With a growing sample size, the uncertainty in model selection disappears, and the RMSE of the two approaches converges to the same value. The same trend can be seen for the IRR estimates in Table 4;
However, the benefits of model averaging are less pronounced than in the case of residual variances.

Analogous tables with the RMSE of the mean estimates and structural variances, which are not the main focus of IRR and our study, are available in the electronic Supplementary Material, as well as the corresponding results for \( j = 5 \) ratings. The electronic Supplementary Material also provides figures with more detailed results for individual data generating models. Namely, for the case of \( j = 3 \) and \( j = 5 \) ratings per ratee, we depict the bias, RMSE, and ratio of bias\(^2\)/MSE for estimates of the means, structural and residual variances, as well as IRR under different data-generating models. We can see a fairly similar performance

**FIGURE 2.** Proportion of correct, incorrect—more complex (MC, containing all true parameter differences plus some additional incorrect differences), and incorrect (missing at least one parameter difference) model selection. Red: Bayesian. Blue: Frequentist model selection techniques. BF: Model selection with Bayes factor proposed here.
in the terms of bias and RMSE across methods with a decreasing bias and RMSE with sample size and higher benefits of model averaging in smaller samples. The bias²/MSE then illustrates the bias variance trade-off between model averaging and model selection, where the decrease in RMSE is accompanied with a relative increase in bias.

3.5.3. Inclusion BF calibration. Finally, we turn our attention to performance of the inclusion BFs in providing evidence for differences in the mean, the structural, and the residual variance between the two groups. Table 5 summarizes the proportion of inclusion BFs correctly favoring the true data generating model for each parameter in both types of scenarios (with difference vs. without difference in a given parameter), averaged across data-generating scenarios of a given type, and the rate of misleading strong evidence (BF₁₀ > 10 in the case of no difference or BF₁₀ < 1/10 in the case of a difference between the groups), in brackets. The table suggests that the inclusion BFs are well calibrated for providing evidence about the differences in means but are noticeably biased when providing evidence toward no difference in structural variances and slightly biased toward models with no difference in residual variances in small samples. However, the
TABLE 4.
RMSE (and Standard Error of RMSE) of the IRR Estimates (Averaged across Conditions, Number of Ratings per Rated Subject $j = 3$)

| Scenario | $N = 25$ | $N = 50$ | $N = 100$ | $N = 200$ |
|----------|----------|----------|-----------|-----------|
| Model selection | | | | |
| BF | .106 (.001) | .082 (.000) | .060 (.000) | .044 (.000) |
| AIC | .119 (.001) | .086 (.000) | .061 (.000) | .043 (.000) |
| BIC | .110 (.001) | .084 (.000) | .064 (.000) | .048 (.000) |
| Forward | .112 (.001) | .085 (.000) | .062 (.000) | .045 (.000) |
| Backward | .112 (.001) | .085 (.000) | .062 (.000) | .045 (.000) |
| WAIC | .106 (.001) | .081 (.000) | .059 (.000) | .043 (.000) |
| LOO | .106 (.001) | .081 (.000) | .059 (.000) | .043 (.000) |
| Model averaging | | | | |
| BMA | .100 (.001) | .076 (.000) | .056 (.000) | .042 (.000) |
| AIC | .108 (.001) | .080 (.000) | .057 (.000) | .041 (.000) |
| BIC | .102 (.001) | .078 (.000) | .059 (.000) | .045 (.000) |
| WAIC | .100 (.001) | .076 (.000) | .055 (.000) | .041 (.000) |
| Pseudo-BMA | .099 (.001) | .075 (.000) | .055 (.000) | .041 (.000) |
| Stacking | .103 (.001) | .077 (.000) | .056 (.000) | .041 (.000) |
| Full model | .122 (.001) | .087 (.000) | .061 (.000) | .043 (.000) |

Note. IRR = inter-rater reliability; BF = Bayes factor; AIC = Akaike information criterion; BIC = Bayesian information criterion; WAIC = Watanabe-AIC; LOO = leave-one-out; RMSE = root mean square error; BMA = Bayesian model averaging.

TABLE 5.
Proportion (and Standard Error of Proportion) of Inclusion Bayes Factors Favoring the Correct Data-Generating Mechanism and the Rate of Misleading Strong Evidence (in Brackets); Averaged across Different Data-Generating Scenarios, Number of Ratings per Rated Subject $j = 3$

| Scenario | $N = 25$ | $N = 50$ | $N = 100$ | $N = 200$ |
|----------|----------|----------|-----------|-----------|
| Mean ($\mu$) | | | | |
| Difference | .626 (.000) | .814 (.000) | .961 (.000) | 1.000 (.000) |
| No difference | .846 (.006) | .888 (.006) | .923 (.005) | .944 (.003) |
| Structural variances ($\sigma_T$) | | | | |
| Difference | .277 (.000) | .316 (.000) | .402 (.000) | .543 (.000) |
| No difference | .805 (.004) | .858 (.006) | .894 (.006) | .933 (.004) |
| Residual variances ($\sigma_e$) | | | | |
| Difference | .399 (.000) | .586 (.000) | .809 (.000) | .971 (.000) |
| No difference | .907 (.004) | .936 (.004) | .952 (.002) | .968 (.003) |
rate of misleading strong evidence is minimal, with less than 0.6% across the parameters and conditions. The proportion of BF s correctly favoring the true data generating process is quickly increasing with the sample size, reaching the probability of 95% for differences in means and for differences in residual variances with \( n = 200 \). The case of \( j = 5 \) ratings depicted in the electronic Supplementary Material then shows that the bias toward no difference in a case of differences in structural variances is improving with the increased number of ratings, pointing to a lack of information about the structural variances themselves in cases of a low number of ratings, which favors the simpler models (i.e., models assuming no difference in structural variance).

A more detailed behavior of the inclusion BF s is depicted in Figure 3, visualizing the distribution of (\( \log_{10} \)) inclusion BF s for a difference in each parameter (columns) with increasing sample sizes (rows) across all simulation conditions. Inclusion BF s from data-generating scenarios with a difference in a given

FIGURE 3. Visualization of \( \log_{10} \) of the inclusion Bayes factors for each parameter aggregated across sample size \( n \). Positive values of \( \log_{10} \) inclusion Bayes factors correspond to evidence in favor of difference in the parameter, and negative values correspond to evidence in favor of no difference in the parameter. Red: aggregated inclusion Bayes factors for scenarios with difference for a given parameter. Blue: distribution visualizing the aggregated inclusion Bayes factors for scenarios with no difference for a given parameter conditions.
parameter are depicted in red and inclusion BFs from scenarios with no difference in a given parameter are depicted in blue.

We can see that inclusion BFs for differences in means BF$\mu_1,\mu_2$ and residual variances BF$\epsilon_1,\epsilon_2$ quickly converge at the correct solution, that is, to the right side of the figure in case of a difference (depicted in red) and to the left side of the figure in case of no difference (depicted in blue). However, the inclusion BFs for the differences in structural variances BF$\gamma_1,\gamma_2$ converge at the correct solution much slower due to the considerably smaller amount of information about structural variances provided by only three ratings per subject. See the electronic Supplementary Material for a comparison with the case of five ratings per subject that shows a slightly improved performance.

### 4. Real Data Examples

We demonstrate the estimation of IRR with two datasets of grant proposal peer-review analyzed by Erosheva et al. (2021) and available in the ShinyItemAnalysis R package (Martinková & Drabinová, 2018).

Unlike in the simulation presented in Section 3, with the real data, the true generating model is unknown. However, the practical examples provide an illustration regarding the application of the proposed method; for the first two examples incorporating a single covariate, we can also compare the results with those provided by other methods presented in the simulation study.

In our analysis, we assume that the selection panels may take the applicant’s gender and career stage into consideration when interpreting the ratings, that is, we allow covariates to explain the fixed part $m_i$ in Equation 3. In Appendix C and in the Supplementary Material, we include the results if no covariate adjustment is expected when interpreting the ratings by the selection panel (i.e., when only models with a constant $\mu$ are considered).

#### 4.1. AIBS Grant Proposal Review Data With Single Covariate

The first example involves peer-review ratings of the American Institutes of Biological Sciences (AIBS), analyzed by Erosheva et al. (2021). In the AIBS data set, each grant proposal was rated three times on the number of criteria as well as on the overall merit score considered here as the dependent variable. The gender ($n_{\text{female}} = 25$ and $n_{\text{male}} = 47$) of the principal investigator was used as a covariate/grouping variable.

The BF, as well as other model selection methods (both Bayesian and frequentist), indicated the simplest model (M1) was the most suitable for the data (Table 6). However, there was a relatively large uncertainty about the selected model, as can be seen from Table 7, which summarizes the weights for the individual models. The simulation study suggested that with this small sample size, all methods have a low probability of selecting the correct model, and it is hard to judge which of the models
is true. The simulation also demonstrated that the model selection based upon BIC more often prefers simpler models, which can also be observed in our practical example, where the model weight based upon BIC is much higher (0.78) for model M1 than the weights for this model based upon other criteria.

Quantifying the evidence across all models with inclusion BFs resulted only in weak evidence in support of the absence of difference in the means, $BF_{\mu\mu} = 2.74$, weak evidence in support of the absence of difference in structural variances $BF_{\gamma\gamma} = 1.38$, and moderate evidence in support of the absence of
difference in residual variances, $BF_{ee} = 3.52$. In other words, there was no clear evidence supporting differences in the means or in structural variances between the two gender groups, but the data were more consistent with differences in residual variances between the two gender groups.

### 4.2. NIH Grant Proposal Review Data With Single Covariate

For the second example, we used the data of peer-review ratings of the National Institutes of Health (NIH), analyzed by Erosheva et al. (2020) and further discussed in Erosheva et al. (2021). We used the preliminary Investigator criterion scores and the gender ($n_{female} = 574$ and $n_{male} = 1310$) of the principal investigator as a covariate.

All model selection methods selected models which assumed a difference in residual variances: BFs and BIC suggested Model M2, the model selection based upon LOO, WAIC, forward, and backward selection also suggested a difference in means (Model M6), while the method based upon AIC selected the most complicated model also suggesting a difference in structural variances. According to Models M2 and M6, the single-rater IRR for grant proposals by male PIs is 0.36 ($0.33 - 0.39$), whereas for grant proposals by female PIs it is 0.33 ($0.29/0.30 - 0.36$) with a difference in IRR of 0.04 ($0.01/0.02 - 0.06$). However, according to model M8, the difference of IRRs for male and female PIs is insignificant $-0.01$ and with a wider CI of $(-0.07 - 0.05)$ also covering zero (see the top panel of Table 8).

Despite a much larger sample size than in the previous example, there is still considerable uncertainty about the selected model (see Table 9). We again use model averaging to account for the model uncertainty, which results in wider confidence intervals. Most of the approaches find differences in IRR between the two groups insignificant, with a confidence interval of $\Delta IRR$ covering zero, see the last three columns of the bottom part of Table 8.

Quantifying the evidence across all models with inclusion BFs resulted only in weak evidence in support of the absence of difference in the means, $BF_{\mu} = 1.15$, weak evidence in support of the absence of difference in structural variances $BF_{\gamma} = 2.21$, and moderate evidence in support of the presence of difference in residual variances, $BF_{ee} = 7.25$. In other words, there was again no clear evidence in favor of differences in the means or in structural variances between the two gender groups, but the data were more consistent with differences in residual variances between the two gender groups. Acknowledging that the differences in residual variances may be caused by differences in career stage, we investigate further in the next section.

### 4.3. NIH Grant Proposal Review Data With More Covariates

We finally considered a more complex situation of IRR being dependent upon two covariates: gender and binarized career stage. In this case, the model given
by Equation 12 has 64 submodels. Model selection using BFs identified that the data were best predicted by a model in which the means, structural variances, and residual variances differed by career stage, with the posterior model probability of 0.23. Table 10 shows 10 models, which were best at predicting the data, accumulating a total of 95% of the posterior model probabilities. We can see
that despite the large sample size and the considerable uncertainty about the best model, most of the best-performing models consider the career stage to be an important predictor for all parameters. However, gender does not seem to play a central role and is slightly more often not even considered.

An overall picture is provided using Bayesian model averaging, which combines estimates and evidence across all models. We find weak evidence against difference in residual variance between the two gender groups, BF\(_{e,e}\) = 1/1.82 = 0.55; however, we find very strong evidence for the difference between the two career stage groups in residual variance, BF\(_{e,e}\) = 1.09 \times 10^{17} . The model-averaged estimates of the residual standard deviation ratios are 0.98 (95\% central credible interval: 0.91–1.00) for the two gender groups and 0.78 (0.74–0.83) for the two career stage groups.

While the residual variance is a parameter of central importance for the assessment of measurement error and IRR, we may also derive from the results conclusions regarding structural variance, the mean, and regarding between-group differences in these parameters. We find weak evidence against difference in structural variance between the two gender groups, BF\(_{\gamma,\gamma}\) = 1/1.43 = 0.70, and we find moderate evidence for difference between the two career stage groups, BF\(_{\gamma,\gamma}\) = 4.74. The model-averaged posterior mean estimates of the structural standard deviation ratios are 0.96 (0.81–1.00) for the two gender groups and 0.85 (0.72–1.00) for the two career stage groups. We find moderate evidence in favor of no difference in means by gender BF\(_{\mu,\mu}\) = 1/0.18 = 5.57 and very strong evidence for difference in mean by career stage

**TABLE 10.** Model Structure for the 10 Best-Performing Models in the NIH Data Set When Considering Both the Gender and Career Stage as Predictors

| \(\mu\)   | \(\sigma_\gamma\) | \(\sigma_e\) | Marg. Lik. | \(p(M_i)\) | \(p(M_i|\text{data})\) |
|-----------|-----------------|-------------|------------|------------|-----------------|
| Stage     | Stage           | Stage       | −7,672.06  | .02        | .23             |
| Stage     | Gender and stage| Stage       | −7,672.15  | .02        | .21             |
| Stage     | Stage           | Gender and stage | −7,672.28 | .02        | .19             |
| Stage     | Gender and stage| Gender and stage | −7,673.39 | .02        | .06             |
| Stage     | Gender          | Stage       | −7,673.47  | .02        | .06             |
| Gender and stage | Stage       | Stage       | −7,673.77  | .02        | .04             |
| Stage     | None            | Stage       | −7,673.79  | .02        | .04             |
| Gender and stage | Gender and stage | Stage       | −7,673.87  | .02        | .04             |
| Stage     | None            | Gender and stage | −7,673.98 | .02        | .03             |
| Gender and stage | Stage       | Gender and stage | −7,673.99 | .02        | .03             |

*Note.* First three columns describe the model in terms of predictors of each parameter (\(\mu\), \(\sigma_\gamma\), and \(\sigma_e\)). Marg. Lik denotes the marginal likelihood, \(p(M_i)\) the prior model probability, and \(p(M_i|\text{data})\) the posterior model probability of each model. NIH = National Institutes of Health.
BF_{μ,μ} = 1.24 \times 10^{31}. The model-averaged posterior mean estimates of the differences in mean are −0.01 (−0.08−0.00) for the two gender groups, and −0.56 (−0.65−0.48) for the two career stage groups.

Higher residual variance in the nonexperienced group is accompanied by only slightly higher structural variance, and it leads to a somewhat lower IRR (see Table 11). Note that when only the models with no covariate effect on the mean are considered, the structural variance in the nonexperienced group is higher, leading to higher IRR in this group (see Table A4 in Appendix C).

We further conducted a sensitivity analysis to assess how our conclusions would change if we specified different prior distributions, that is, tested different hypotheses. We used the two remaining prior distributions depicted in Figure 1; (1) a more concentrated prior distribution around no effect with the standard deviation \( \sigma = 0.25 \)—testing for the presence of smaller differences between the groups and (2) a wider prior distribution with standard deviation \( \sigma = 1 \)—testing for the presence of larger differences between the groups.

We will only discuss here the residual variances (see Appendix B for more details). Between the two gender groups, we found strong evidence for the absence of larger differences in the residual variances and no evidence supporting the presence or the absence of small differences. For the groups segregated by career stage, we found very strong evidence for presence of the effect, regardless of the specification of the alternative hypothesis. In other words, the data were more consistent with no or small differences in residual variances between the two gender groups and there was clear evidence in favor of differences in residual variances between the two career stage groups.

5. Discussion

In this work, we have presented a new flexible approach for assessing the IRR in cases, where variance components, depending upon covariates, are assumed to differ. We used mixed effect models with heterogeneous variances and employed the Bayesian framework with BF\textsubscript{s} and Bayesian model averaging. In a simulation study, we compared the methodology to other frequentist and Bayesian
approaches and shown comparable or superior performance to those of the other methods. More importantly, flexibility in the proposed methodology allows researchers to straightforwardly extend the presented models to cases with more covariates, whereas BFs can be used to select a single model, researchers can further account uncertainty in the model structure with Bayesian model averaging when drawing inferences about either the presence or absence of the effect via inclusion.

The suggested methodology—Bayesian hypothesis testing and model averaging—is, of course, not the only option researchers can pursue. Authors with different philosophical views would advocate for different approaches, such as estimation only or inference based upon confidence intervals (e.g., Cumming, 2014; Gelman & Hill, 2006). We prefer the Bayesian hypothesis testing since we believe that BFs (and likelihood ratio tests) are the only coherent method of testing for the presence versus absence of the effect. The problem with hypothesis testing based upon posterior credible intervals (or \( p \) values) is the assumption of either the presence (for credible intervals) or absence (for \( p \) values) of the effect at the onset of the analysis. In other words, it is impossible to provide evidence for/against an assumption that is already taken for granted (e.g., Jeffreys, 1939).

The advantages of Bayesian hypothesis testing and model averaging, however, come at an additional cost: the specification of prior distributions. Prior distributions are especially important upon the parameters of interest, where they define the hypotheses about the presence versus the absence of an effect. Different prior distributions equal to different hypotheses—different questions—being asked. Subsequently, different questions might lead to different answers (e.g., a one-sided vs. two-sided test). Nonetheless, as shown by the sensitivity analysis in Appendix B, similar prior distributions correspond to similar questions, which subsequently result in similar answers. To define a prior distribution, researchers must be able to define the degree of effects they are interested in (see, e.g., Johnson et al., 2010; Mikkola et al., 2021; O’Hagan et al., 2006, for detailed information about prior distribution elicitation).

The proposed method was further demonstrated when assessing IRR in a grant proposal peer-review with respect to the applicant’s gender and career stage. The results suggested that the IRR is not likely dependent upon the gender of the principal investigator, while it may be lower with a lower career stage. When demonstrated in this specific example of grant peer review, it is worth noting the importance and the wide range of possible applications using the proposed method. Our methods may be used to identify gaps in the IRR for various rating situations (applicant hiring or promotion, classroom observation of teachers, journal peer-review, etc.) and with respect to the different types of rater and ratee characteristics (dichotomous such as internal/external status, factors such as social status or marital status, and continuous such as age).
We also discussed how hypotheses regarding specific variance components may be addressed with the inclusion BFs. This is especially important because IRR is influenced by range restriction (Erosheva et al., 2021), meaning that for a fixed residual variance, different values of IRR are obtained depending upon the structural variance. For this reason, the difference between residual variances may be of greater interest than the difference in the IRR itself. This aspect was also demonstrated with the NIH example by comparing the case of when the grand mean was allowed to vary with covariates to the case of no covariate adjustment: In the latter case, the structural variance for the nonexperienced group was higher, leading to a somewhat higher IRR than in the experienced group, while when the ratings accounted for the stage, the structural variance was somewhat lower for the nonexperienced group, leading to a somewhat lower IRR. Unlike IRR which provided contradictory conclusions, the inclusion BFs provided more coherent information and in both cases unanimously concluded there is a significant difference in the residual variance and more error in the ratings from applicants in the nonexperienced group.

Several limitations of the current study and possible directions for future research are worth mentioning: First, a simulation study will always cover only a finite and rather limited number of parameter setups, and our simulation study involved only one binary covariate. Nevertheless, the current simulation study was already extensive in terms of the number of methods compared and the simulation time needed. Besides the computation time, some aspects of the frequentist approaches would need to be solved. As an example, the parametrization is not straightforward in the \texttt{lme()} and \texttt{lmer()} functions for cases involving additional covariates of variance components. The model selection outlined in Appendix A also becomes more complicated with the addition of covariates of variance components and the actual size of the entire space for possible models increases exponentially. Our real data examples show how a model with one covariate can be applied to models with multiple covariates.

Secondly, we considered binary covariates only. However, our approach could be easily extended to other types of covariates. For example, in the case of factors with multiple groups, one has to decide whether an ANOVA-like test for at least one difference between the factor levels should be specified with orthonormal contrasts asserting that the prior marginal levels are identical, making the levels interchangeable (Rouder et al., 2012), or, whether a multiple treatments versus control-like test should be specified with dummy coding, asserting that the control condition corresponds to the grand mean and the coefficients for each treatment condition to differences/standard deviation ratios. Similarly, priors on continuous covariates simply correspond to the unit change in the covariate, with the grand mean/variance corresponding to the covariate value of 0, while the centering or rescaling of the covariate prior to the analysis might simplify the prior specification.
Thirdly, we considered only the simplest model with the ratee being the single structural source of error, while all other possible sources of error (such as rater, occasion, etc.) were encompassed in the residual error. This model is appropriate and widely used when most of the raters rate only a single ratee. In real-life applications, the hierarchy may be more sophisticated (respondents nested within institutions, which themselves are nested within towns or countries), and there may be more sources of error such as raters, the so-called facets in the context of the generalizability theory (Brennan, 2001). Different generalizability and dependability coefficients may then be defined in such cases, and the IRR of interest may need to be defined by more complex ratios with different interpretations. However, the cases of heterogeneity would then be treated analogously, and the Bayesian approach suggested here would be easily applied to more complex situations.

Regardless of the limitations, the study offers a flexible method for assessing the heterogeneity in IRR with respect to rater and ratee characteristics. This may help identifying the subgroups with lower IRR and improving the IRR of these groups and in general. This in turn may be of great importance to those designing the ratings and to policymakers whose interest is to improve assessment systems and the selection processes.

Appendix

A. Model Selection Diagrams

FIGURE A1. Diagram depicting the forward (left) and backward (right) stepwise selection procedures. Arrows visualize the selection flow, nodes in wrapped in full line visualize the non-terminal steps, and nodes wrapped in dashed line visualize the terminal steps.
B. Sensitivity to Alternative Prior Distributions

For sensitivity analysis, we used the NIH grant proposal example with two covariates, discussed in Section 4.3. We re-estimated the models under the two additional prior distributions depicted in Figure 1: (1) a tighter prior distribution with $\sigma = 0.25$, specifying expectation of smaller differences between the groups and (2) a wider prior distribution with $\sigma = 1$, specifying expectation of larger differences between the groups. Table A1 describes the resulting inclusion Bayes factors under each parameterization and Table A2 describes the resulting model-averaged parameter estimates under each parameterization. The results from the middle column are identical to those presented in Section 4.3.

**TABLE A1.**
Inclusion Bayes Factors for Differences Between the Groups for Each Parameter (Rows) Across the Different Prior Distribution Specifications

| Prior Distribution Setting | $\sigma = 0.25$ | $\sigma = 0.50$ | $\sigma = 1.00$ |
|-----------------------------|-----------------|-----------------|-----------------|
| Residual variances ($\sigma_x$) |                |                |                |
| Gender                      | 0.96            | 0.55            | 0.31            |
| Stage                       | $2.00 \times 10^{17}$ | $1.09 \times 10^{17}$ | $7.84 \times 10^{16}$ |
| Structural variances ($\sigma_y$) |                |                |                |
| Gender                      | 1.06            | 0.70            | 0.43            |
| Stage                       | 7.80            | 4.74            | 2.60            |
| Mean ($\mu$)                |                |                |                |
| Gender                      | 0.36            | 0.18            | 0.09            |
| Stage                       | $7.77 \times 10^{30}$ | $1.24 \times 10^{31}$ | $9.67 \times 10^{30}$ |

**TABLE A2.**
Posterior Model-Averaged Estimates (and 95\% Credible Intervals) of Differences/Standard Deviation Ratios Between the Groups for Each Parameter (rows) across the Different Prior Distribution Specifications

| Prior Distribution Setting | $\sigma = 0.25$ | $\sigma = 0.50$ | $\sigma = 1.00$ |
|-----------------------------|-----------------|-----------------|-----------------|
| Residual standard deviation ratio |          |                |                |
| Gender                      | 0.97 (0.91, 1.00) | 0.98 (0.91, 1.00) | 0.99 (0.92, 1.00) |
| Stage                       | 0.78 (0.74, 0.83) | 0.78 (0.74, 0.83) | 0.78 (0.74, 0.83) |
| Structural standard deviation ratio |          |                |                |
| Gender                      | 0.95 (0.81, 1.00) | 0.96 (0.81, 1.00) | 0.97 (0.81, 1.00) |
| Stage                       | 0.84 (0.72, 1.00) | 0.85 (0.72, 1.00) | 0.87 (0.72, 1.00) |
| Mean difference             |               |                |                |
| Gender                      | $-0.01 (-0.10, 0.00)$ | $-0.01 (-0.08, 0.00)$ | $-0.00 (-0.07, 0.00)$ |
| Stage                       | $-0.55 (-0.63, -0.47)$ | $-0.56 (-0.65, -0.48)$ | $-0.57 (-0.65, -0.48)$ |
As expected, the different prior distributions had a noticeable effect on the degree of evidence either for/against the differences in the means/standard deviation ratios. Table A1 suggests more evidence for the null hypotheses with increasing standard deviation of the prior distribution; however, the qualitative conclusions remain essentially unchanged. Evidence against the presence of differences in the residual variances was, at most, between weak and moderate evidence, the later only under a prior distribution expecting large effects. The best performing models (not displayed) are also similar under all model specifications, although the tightest prior distributions ($\sigma = 0.25$) included both gender and stage as a predictor for the structural variances in the model with the highest posterior model probability (0.21), whereas the models with the medium ($\sigma = 0.50$) and wide ($\sigma = 1.00$) prior distributions included only stage for all three parameters in the model with the highest posterior model probability (0.23 and 0.34, respectively). Most importantly, as suggested by Table A2, the posterior model-averaged estimates remain essentially unchanged under the different prior distribution specifications.

C. Mean-Unadjusted Results for the NIH Data

Here, we present results assuming constant overall mean, that is, restricting $\beta_\mu$ in Equation 4 to 0, which corresponds to evaluating variance components and IRR for the covariate unadjusted ratings. The best-performing model suggests differences in the structural and also in the residual variance for the two stage groups, but no effect of gender (Table A3).

As in the covariate adjusted ratings (Section 4.3), we find a very strong evidence for the difference between the two career stage groups in residual variance, $BF_{e,e} = 9.84 \times 10^{16}$, and we find weak evidence against the difference in residual variance between the two gender groups, $BF_{e,e} = 1/2.09 = 0.48$. The model-averaged estimates of the residual standard deviation ratios are 0.79 (0.75

As in the covariate adjusted ratings (Section 4.3), we find a very strong evidence for the difference between the two career stage groups in residual variance, $BF_{e,e} = 9.84 \times 10^{16}$, and we find weak evidence against the difference in residual variance between the two gender groups, $BF_{e,e} = 1/2.09 = 0.48$. The model-averaged estimates of the residual standard deviation ratios are 0.79 (0.75

TABLE A3.

Model Structure for the Best-Performing Models in the NIH Data Set When Considering Both the Gender and Career Stage as Predictors for Variance Components, and Assuming Constant Overall Mean

| $|\mu$ | $\sigma_\gamma$ | $\sigma_\varepsilon$ | Marg. Lik. | $p(M_i)$ | $p(M_i|data)$ |
|-----|-------------|----------------------|------------|----------|--------------|
| None | Stage       | Stage                | $-7,743.49$ | .06      | .44          |
| None | Stage       | Gender and stage     | $-7,744.02$ | .06      | .26          |
| None | Gender and stage | Stage                | $-7,744.11$ | .06      | .24          |
| None | Gender and stage | Gender and stage     | $-7,745.41$ | .06      | .06          |

Note. First three columns describe the model in terms of predictors of each parameter ($\mu$, $\sigma_\gamma$, and $\sigma_\varepsilon$). Marg. Lik denotes the marginal likelihood, $p(M_i)$ the prior model probability, and $p(M_i|data)$ the posterior model probability of each model. NIH = National Institutes of Health.
– 0.83) for the two career stage groups and 0.98 (95% central credible interval: 0.92 – 1.00) for the two gender groups.

However, unlike in the case of covariate adjusted means presented in Section 4.3, for the structural variance, we find very strong evidence for the difference between the two career stage groups, BF$_{g;G}$ = 2.99 × 10$^{16}$, with the model-averaged posterior mean estimates of the structural standard deviation ratios of 0.79 (0.75 – 0.83) for the two career stage groups.

The high values of structural variance in the nonexperienced group subsequently lead to higher IRR values (see Table A4).

**TABLE A4.**
Estimated Model-Averaged Marginal Means and 95% CI for Each of the Parameters

| Gender | Stage   | $\mu$     | $\sigma_y$ | $\sigma_e$ | IRR   |
|--------|---------|-----------|------------|------------|-------|
| Female | nExp    | -.06 [-.09, -.02] | .81 [.73, .91] | .97 [.93, 1.03] | .41 [.36, .47] |
| Male   | nExp    | -.06 [-.09, -.02] | .79 [.71, .87] | .95 [.91, .99] | .41 [.35, .46] |
| Female | Exp     | -.06 [-.09, -.02] | .53 [.48, .60] | .76 [.73, .81] | .33 [.27, .39] |
| Male   | Exp     | -.06 [-.09, -.02] | .51 [.47, .56] | .75 [.72, .78] | .32 [.27, .37] |

*Note.* Exp = experienced; nExp = nonexperienced; IRR = inter-rater reliability.

**Authors’ Note**

Additional tables and figures, and accompanying R scripts are available at https://osf.io/bk8a7/.

**Acknowledgments**

The authors appreciate the computing and storage facilities of the Institute of Computer Science (Czech Republic RVO 67985807) and access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the program “Projects of Large Research, Development, and Innovations Infrastructures” (CESNET LM2015042). They are thankful to Martin Otava, Jon Kern, and anonymous reviewers for their helpful comments and suggestions on earlier versions of this article.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: The study was funded by the Czech Science Foundation grant number 21-03658S.
Note

1. We found that leave-one-rating cross-validation performed better than leave-one-ratee cross-validation in our simulations; therefore, we show results only for leave-one-rating-based metrics.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Bartoš, F., Martinková, P., & Brabec, M. (2020). Testing heterogeneity in inter-rater reliability. In M. Wiberg, D. Molenaar, J. González, U. Bockenholt, & J.-S. Kim (Eds.), *Quantitative psychology: The 84th annual meeting of the psychometric society, Santiago, Chile, 2019* (pp. 347–364). Springer (Proceedings in Mathematics & Statistics). https://doi.org/10.1007/978-3-030-43469-4_26

Breiman, L. (1996). Stacked regressions. *Machine Learning, 24*(1), 49–64.

Brennan, R. L. (2001). *Generalizability theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3456-0

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Dablander, F., Bergh, D. v. d., Ly, A., & Wagenmakers, E.-J. (2020). Default Bayes factors for testing the (in) equality of several population variances. *arXiv*. https://arxiv.org/abs/2003.06278v1

Depaoli, S., Lai, K., & Yang, Y. (2020). Bayesian model averaging as an alternative to model selection for multilevel models. *Multivariate Behavioral Research, 56*, 920–940.

Erosheva, E. A., Grant, S., Chen, M.-C., Lindner, M. D., Nakamura, R. K., & Lee, C. J. (2020). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances, 6*(23), eaaz4868.

Erosheva, E. A., Martinková, P., & Lee, C. (2021). When zero is not 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. https://doi.org/10.1111/rssa.12681

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science, 32*, 313–329. https://doi.org/10.1214/16-STS599

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review, 86*(1), 1–28. https://doi.org/10.1111/insr.12243

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association, 74*(365), 153–160.
Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov Chain Monte Carlo in Practice, 4*, 145–161.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Goldhaber, D., Grout, C., Wolff, M., & Martinková, P. (2021). Evidence on the dimensionality and reliability of professional references’ ratings of teacher applicants. *Economics of Education Review, 83*, 102130. https://doi.org/10.1016/j.econedurev.2021.102130

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science, 4*(3), 25152459211031256.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgessampling: An R package for estimating normalizing constants. *Journal of Statistical Software, 92*, 1–29.

Gronau, Q. F., & Wagenmakers, E.-J. (2019a). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior, 2*, 1–11.

Gronau, Q. F., & Wagenmakers, E.-J. (2019b). Rejoinder: More limitations of Bayesian leave-one-out cross-validation. *Computational Brain & Behavior, 2*, 35–47.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*(464), 879–899.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(4), 382–401.

Jeffreys, W. H. (1931). *Scientific inference*. Cambridge University Press.

Jeffreys, W. H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

Jefferys, W. H., & Berger, J. O. (1992). Ockham’s razor and Bayesian analysis. *American Scientist, 80*, 64–72.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., & Feldman, B. M. (2010). Methods to elicit beliefs for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology, 63*(4), 355–369. https://doi.org/10.1016/j.jclinepi.2009.06.003

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data* (Vol. 53). Wiley New York.

LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association, 91*(436), 1641–1650.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Madigan, D., Raftery, A. E., York, J. C., Bradshaw, J. M., & Almond, R. G. (1994). Strategies for graphical model selection. In P. Cheeseman & R. W. Oldford (Eds.),
Martinková et al.

Selecting models from data (pp. 91–100). Springer-Verlag. https://doi.org/10.1007/978-1-4612-2660-410

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. The R Journal, 10(2). https://doi.org/10.32614/RJ-2018-074

Martinková, P., Goldhaber, D., & Erosheva, E. A. (2018). Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. PLoS One, 13(10), e0203002. https://doi.org/10.1371/journal.pone.0203002

McElreath, R. (2018). Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman; Hall/CRC.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30. https://doi.org/10.1037/1082-989X.1.1.30

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. Statistica Sinica, 6, 831–860.

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., & Klami, A. (2021). Prior knowledge elicitation: The past, present, and future. https://arxiv.org/abs/2112.01380

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. https://doi.org/10.1371/journal.pone.0048509

O’Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). Uncertain judgements: Eliciting experts’ probabilities. John Wiley & Sons.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D.R Core Team. (2021). nlme: Linear and nonlinear mixed effects models [R package version 3.1]. https://CRAN.R-project.org/package=nlme

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika, 83(2), 251–266.

Raftery, A. E., Madigan, D., & Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. Bayesian Statistics 5, 323–349.

Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes’ theorem: Strength of evidence as predictive accuracy. The American Statistician, 73(2), 186–190. https://doi.org/10.1080/00031305.2017.1341334

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology, 56(5), 356–374.

Sattler, D. N., McKnight, P. E., Nancy, L., & Mathis, R. (2015). Grant peer review: Improving inter-rater reliability with training. PLoS One, 10(6), e0130450.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. Psychonomic Bulletin & Review, 25(1), 128–142.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464. https://doi.org/10.1214/aos/1176344136

Searle, S. R. (1997). Linear models. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118491782
Assessing Inter-rater Reliability With Heterogeneous Variance Components Models

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). Variance components. John Wiley & Sons.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420. https://doi.org/10.1037/0033-2909.86.2.420

Stan Development Team. (2018). RStan: The R interface to Stan [R package version 2.17.3]. http://mc-stan.org/3

Stefán, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. Behavior Research Methods, 51(3), 1042–1058.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [R package version 2.4.1]. https://mc-stan.org/loo/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave–one–out cross–validation and WAIC. Statistics and Computing, 27, 1413–1432.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. Psychonomic Bulletin & Review, 11(1), 192–196. https://doi.org/10.3758/BF03206482

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(12), 3571–3594.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C. Rao & S. Sinharay (Eds.), Psychometrics (pp. 81–124). Elsevier. https://doi.org/10.1016/S0169-7161(06)26004-8

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2020). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. Psychological Methods, 26(1), 74–89. https://doi.org/10.1037/met0000270

Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. Behavior Research Methods, 51(5), 1968–1986. https://doi.org/10.3758/s13428-019-01255-9

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. Philosophical Magazine, 42, 369–390. https://doi.org/10.1080/14786442108633773

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Analysis, 13(3), 917–1007.

Authors

PATRÍCIA MARTINKOVÁ is a senior researcher and a Vice-Chair at the Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, and a Researcher at the Institute for Research and Development in Education, Faculty of Education, Charles University; e-mail: martinkova@cs.cas.cz. Her research interests include statistical and computational aspects of psychometric methods.

FRANTIŠEK BARTOŠ is a PhD fellow at the Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences and a PhD Student
Martinková et al.

at the Department of Psychological Methods, University of Amsterdam; e-mail: bartos@cs.cas.cz. His research focuses on Bayesian modeling and meta-analyses.

MAREK BRABEC is a senior researcher at the Department of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences; e-mail: mbrabec@cs.cas.cz. His research interests include statistical methods in general and semiparametric and Bayesian modeling in particular.

Manuscript received November 22, 2021
Revision received June 14, 2022
Accepted December 3, 2022