Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients

ten Hove, D.; Jorgensen, T.D.; van der Ark, L.A.

DOI
10.1007/978-3-030-43469-4_7

Publication date
2020

Document Version
Final published version

Published in
Quantitative Psychology

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2020). Comparing hyperprior distributions to estimate variance components for interrater reliability coefficients. In M. Wiberg, D. Molenaar, J. González, H. Böckenholt, & J.-S. Kim (Eds.), *Quantitative Psychology: 84th Annual Meeting of the Psychometric Society, Santiago, Chile, 2019* (pp. 79-93). (Springer Proceedings in Mathematics & Statistics; Vol. 322). Springer. https://doi.org/10.1007/978-3-030-43469-4_7

Comparing Hyperprior Distributions to Estimate Variance Components for Interrater Reliability Coefficients

Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark

Abstract Interrater reliability (IRR) is often estimated by intraclass correlation coefficients (ICCs). Using Markov chain Monte Carlo (MCMC) estimation of Bayesian hierarchical models to estimate ICCs has several benefits over traditional approaches such as analysis of variance or maximum likelihood estimation. However, estimation of ICCs with small sample sizes and variance parameters close to zero, which are typical conditions in studies for which the IRR should be estimated, remains problematic in this MCMC approach. The estimation of the variance components that are used to estimate ICCs can heavily depend on the hyperprior distributions specified for these random-effect parameters. In this study, we explore the effect of a uniform and half-*t* hyperprior distribution on bias, coverage, and efficiency of the random-effect parameters and ICCs. The results indicated that a half-*t* distribution outperforms a uniform distribution but that slightly increasing the number of raters in a study is more influential than the choice of hyperprior distributions. We discuss implications and directions for future research.

Keywords Bayesian hierarchical modeling · Hyperprior distributions · Interrater reliability · Intraclass correlation coefficients · Markov chain Monte Carlo estimation · Random effects · Variance components

1 Introduction

In an ongoing research project (Ten Hove, Jorgensen, & Ten Hove et al. 2018; Ten Hove et al. 2019), we propose to estimate interrater reliability (IRR) with different types of intraclass correlation coefficients (ICCs) using Markov chain Monte Carlo (MCMC) estimation of Bayesian hierarchical models. MCMC estimation has several benefits over more traditional approaches, such as analysis

D. ten Hove (✉) · T. D. Jorgensen · L. A. van der Ark
Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands
e-mail: D.tenHove@uva.nl

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 322, https://doi.org/10.1007/978-3-030-43469-4_7

of variance (ANOVA) or maximum likelihood estimation (MLE). For example, MCMC estimation can easily accommodate missing at random data, which is a pitfall of ANOVA (Brennan 2001); for small sample sizes and parameters close to a boundary, it typically outperforms MLE (Gelman et al. 2013); and it provides Bayesian credible intervals, which quantify the uncertainty of the estimated ICCs (Hoekstra et al. 2014), whereas for both ANOVA and MLE, estimating confidence intervals is troublesome (Brennan 2001). However, we found that when using MCMC some ICCs were severely underestimated and inefficient when the number of raters was small or one of the variance components involved in the ICCs was close to zero. As a solution to these estimation difficulties, we proposed a planned missing data design, in which a subset of randomly drawn raters was assigned to each subject. This improved the estimation of ICCs using MCMC vastly, but some ICCs were still biased and inefficient due to biased and inefficiently estimated rater variances (Ten Hove et al. 2019).

The estimation difficulties in conditions with few raters and low variability are consistent with several studies on the performance of MCMC estimation for hierarchical models (Gelman and Hill 2006; McNeish and Stapleton 2016; Polson and Scott 2012; Smid et al. 2019). In the MCMC approach to estimating the ICCs, priors should be specified for the distribution of random effects. Because the variance of these random effects is itself estimated, a prior distribution should be specified for that parameter, called a hyperprior distribution (Gelman et al. 2013, p. 107–108). The performance (e.g., bias and efficiency) of the parameter estimates depends on the specification of these hyperpriors (Gelman and Hill 2006; Gelman et al. 2013; McNeish and Stapleton 2016; Polson and Scott 2012; Smid et al. 2019). The specification of hyperpriors thus provides an opportunity to improve the performance of parameter estimates of random effects. In our current research project, we followed Gelman's (2006, p. 527) advice to start with weakly informative uniform prior distributions on the random effects *SD*s. Several researchers (including Gelman himself) debated the use of these hyperpriors when the data provide little information about clusters (here raters) because uniform distributions may put too much probability mass on unreasonably high values. Various alternatives to these uniform hyperprior distributions were proposed and tested (Gelman 2006; McNeish and Stapleton 2016; Polson and Scott 2012; Spiegelhalter et al. 2004; Van Erp et al. 2019).

This study informs researchers which hyperprior distributions should be used to estimate ICCs. We investigated the effect of different hyperprior distributions on the bias, coverage rates, and efficiency of random-effect parameters and ICCs. The remainder of this paper is structured as follows. First, we briefly discuss the definition of IRR in terms of ICCs and their MCMC estimation. Second, we provide a short overview of (properties of) hyperprior distributions for random effects. Third, we present the results of a simulation study that tested the performance of the random-effect parameters and ICCs using different hyperprior distributions. We focus on conditions with very few raters, to draw attention to difficulties in estimating IRR in conditions that are typical for observational studies. Finally, based on the simulation results, we discuss implications for applied research and directions for future methodological research.

2 Interrater Reliability

2.1 Definition

Bartko (1966), Shrout and Fleiss (1979), and McGraw and Wong (1996) defined IRR in terms of ICCs. They identified raters and subjects as the main sources of variance in a rating process and decomposed each observation into the main and interaction effects of these raters and subjects. Let Y_{sr} be the score of subject s as rated by rater r on attribute Y. Y_{sr} is then decomposed into a grand mean (μ), a subject effect (μ_s), a rater effect (μ_r), an interaction effect, and random measurement error. In practice, the subject × rater interaction and random measurement error cannot be disentangled, so let μ_{sr} denote a combination of both these elements. The decomposition of Y_{sr} equals

$$Y_{sr} = \mu + \mu_s + \mu_r + \mu_{sr}. \tag{1}$$

If the raters are nested within subjects (i.e., a unique set of raters is used to rate each subject's attribute), μ_r and μ_{sr} cannot be disentangled. For simplicity, we ignore this situation in this paper. Each of the effects in Eq. 1 are assumed to be uncorrelated (Brennan 2001). The variance of Y can therefore be decomposed into the orthogonal variance components of each of these effects, resulting in

$$\sigma_Y^2 = \sigma_s^2 + \sigma_r^2 + \sigma_{sr}^2. \tag{2}$$

If it is assumed that raters and subjects are randomly drawn from a larger population of raters and subjects, respectively, the variances in Eq. 2 are modeled as random-effect variances components.

The variances components in Eq. 2 are used for several definitions of IRR. Each of these definitions is an ICC and defines IRR as the degree to which the ordering (consistency: C) or absolute standing (agreement: A) of subjects is similar across raters. In other words, the IRR is the degree to which subject effects can be generalized over raters. Assume we have k raters rating each subject. The most elaborated ICC (agreement based on the average rating of k raters) is defined as

$$\text{ICC(A}, k) = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2 + \sigma_{sr}^2}{k}}. \tag{3}$$

Other definitions of IRR are obtained by removing terms from Eq. 3, as is displayed in Table 1. For more information about these ICCs and the underlying variance decomposition, we refer to Bartko (1966), McGraw and Wong (1996), and Shrout and Fleiss (1979).

Table 1 Cross classification of ICCs in terms of type (agreement and consistency) and number of raters (single rater, $k > 1$ raters)

	Agreement	Consistency
Single rater	$\text{ICC(A, 1)} = \frac{\sigma_s^2}{\sigma_s^2+\sigma_r^2+\sigma_{sr}^2}$	$\text{ICC(C, 1)} = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2}$
Average of k raters	$\text{ICC(A,} k) = \frac{\sigma_s^2}{\sigma_s^2+(\sigma_r^2+\sigma_{sr}^2)/k}$	$\text{ICC(C,} k) = \frac{\sigma_s^2}{\sigma_s^2+\sigma_{sr}^2/k}$

2.2 MCMC Estimation

The ICCs from Table 1 can be estimated using MCMC estimation of a Bayesian hierarchical model. Let $\boldsymbol{\theta}$ denote a model's vector of parameters, and Y denote the data. In the MCMC approach, the posterior distribution of the model parameters given the data, $P(\boldsymbol{\theta}|\boldsymbol{Y})$, is estimated as proportional to the product of the prior probability distribution of the parameters, $P(\boldsymbol{\theta})$, and the likelihood of the data conditional on the parameters, $P(\boldsymbol{Y}|\boldsymbol{\theta})$, that is, $P(\boldsymbol{\theta}|\boldsymbol{Y}) \propto P(\boldsymbol{\theta})P(Y|\boldsymbol{\theta})$ (Gelman et al. 2013, p. 6–7).

The MCMC approach thus requires the specification of a prior probability distribution for each parameter. Because MCMC treats each of the random effects in Eq. 1 as parameters to estimate, their variance components in Eq. 2 are so-called hyperparameters (i.e., they are parameters that describe the distribution of other parameters). These hyperparameters require their own prior distribution (named hyperprior distribution), which we discuss in more detail in the following section. Depending on the software, the hyperparameters can be estimated in terms of either random-effect *SD*s (which should be squared to obtain the random-effect variances for the ICCs) or random-effect variances. For simplicity, we ignore the terms hyperparameters and variance components in the remainder of this paper and consistently use *random-effect variances* to refer to σ_s^2, σ_r^2, and σ_{sr}^2 or *random-effect* SD*s* to refer to the square roots of these random-effect variances.

MCMC estimation repeatedly samples from the posterior distributions, resulting in an empirical posterior distribution for each (hyper)parameter. When deriving ICCs, the posterior distributions of the random-effect variances are combined, yielding an empirical posterior distribution for each of the ICCs. From these empirical posterior distributions, Bayesian credible intervals (BCIs) can be derived that quantify the uncertainty about the random-effect variances and the ICCs that are calculated from these random-effect variances, for example, using percentiles as limits or kernel density estimators to obtain highest posterior density (HPD) limits.

The main difficulty in estimating the ICCs from Table 1 is rooted in the estimation of the random-rater effect variance, σ_r^2 (Ten Hove, Jorgensen, & Ten Hove et al. 2018; Ten Hove et al. 2019). Observational studies often involve few raters, and, when these raters have been trained well, they vary little in the average ratings that they provide. The data thus provide little information about, σ_r^2. As a result, its posterior is overwhelmingly influenced by the specified hyperprior distribution. This typically results in an over- and inefficiently estimated random-effect variance. Estimation difficulties of σ_r^2 due to influential hyperprior distributions can, in turn, result in under- and inefficiently estimated ICCs.

3 Hyperprior Distributions

The choice among hyperprior distributions for random-effect variances is frequently discussed (see e.g., Gelman 2006; Gelman et al. 2013; Smid et al. 2019; Van Erp et al. 2019). Prior and hyperprior distributions can be classified into informative or uninformative distributions, proper or improper distributions, and default, thoughtful or data-dependent distributions. For more information on these classifications, we refer to Gelman (2006), Gelman et al. (2013, chapter 2), and Smid et al. (2019).

When raters are skilled and the subjects can be scored objectively, it is reasonable to assume that raters differ little in their average ratings. We therefore believe that the hyperprior distribution for σ_r^2 should be weakly informative and put a relatively large weight on small values compared to large values. The other random-effect variances, σ_s^2 and σ_{sr}^2, are typically obtained from larger sample sizes (i.e., N (subjects) for σ_s^2 and N (subjects) × k (raters) for σ_{sr}^2). The data thus provide more information about these random-effect variances, making their hyperprior distributions less influential on the posterior compared to the hyperprior distribution of σ_r^2. Because σ_s^2 and σ_{sr}^2 are expected to be larger than σ_r^2, their hyperprior distributions should allow for large values.

Given these considerations, we prefer weakly informative, thoughtful hyperprior distributions. Moreover, we prefer hyperprior distributions that yield proper posterior distributions. We take these criteria into account while discussing three popular hyperprior distributions for variance parameters: a uniform distribution (Gelman 2006; McNeish and Stapleton 2016), the inverse-gamma distribution (Spiegelhalter et al. 2004), and the half-t or half-Cauchy distributions (Gelman 2006; McNeish and Stapleton 2016; Polson and Scott 2012).

3.1 Uniform Distribution

The uniform distribution is a popular hyperprior distribution with two parameters: a lower bound and an upper bound. This distribution implies a researcher's believe that all values within a specified range are equally likely. For random-effect *SD*s, the uniform hyperprior distribution can be specified as weakly informative by using the range $\left[0, \frac{max_Y - min_Y}{2}\right]$ (i.e., the smallest and largest possible *SD*), where max_Y and min_Y are the maximum and minimum value of Y, respectively. If max_Y and min_Y are estimated from the data, the uniform distribution is data dependent; if max_Y and min_Y are specified as the theoretically maximum and minimum values of Y, respectively (e.g., using the minimum and maximum possible scores, such as anchor points on a Likert scale), the uniform distribution is data independent. In practice, it is unlikely to find a random-effect *SD* near the upper bound of $\left[\frac{max_Y - min_Y}{2}\right]$. Such a large upper bound is unintentionally influential on the posterior when the data contain too little information about a parameter. Examples of little information include small sample sizes but also when the random-effect variance is nearly zero. However, it may be difficult to defend the choice of a hard upper bound of the uniform posterior distribution that is smaller than the maximum possible *SD*. A uniform hyperprior distribution performs best when it is specified for random-effect *SD*s (e.g., σ_r), rather than for random-effect variances (e.g., σ_r^2). To yield proper posteriors, a hyperprior distribution requires at least three clusters for random-effect *SD*s, or at least four clusters for random-effect variances (Gelman 2006).

3.2 Inverse-Gamma Distribution

The inverse-gamma distribution is another popular hyperprior distribution for random-effect variances, which is defined on a positive scale and has two parameters: a shape and scale parameter. This distribution is very sensitive to its specified shape and scale parameters when the estimated σ^2 is small, and its specification is therefore too influential for typical IRR studies in which σ_r^2 is expected to be low (Gelman et al. 2013, p. 315–316). Moreover, the inverse-gamma hyperprior distribution yields improper posteriors when the shape and scale parameters are specified as very uninformative (Gelman 2006). Therefore, although it is a commonly applied prior in many other settings (and potentially required as a conjugate prior for Gibbs sampling), we consider the inverse-gamma hyperprior inappropriate for our purpose.

3.3 Half-t or Half-Cauchy Distribution

The half-*t* and half-Cauchy hyperprior distributions were also proposed as hyperprior distributions for random-effect variances and are defined on a positive scale (Gelman 2019; Polson and Scott 2012). The half-*t* distribution has three parameters: a shape, location, and scale parameter. The half-Cauchy distribution is equivalent to a half-*t* distribution for $df = 1$ and thus only has a location and scale parameter. The half-Cauchy distribution has more kurtosis than *t* distributions having $df > 1$, allowing the greatest probability density for extreme values while still placing most probability density near the center of the distribution. If a wide range of possible values is specified for the random-effect variances, these distributions are specified as data independent and weakly informative. Especially for σ_r^2, we expect values near zero, so a half-*t* distribution with higher $df = 4$ is slightly more informative and is recommended for variance parameters that are expected to have values near the lower bound of zero (Gelman 2019). This may, however, be less beneficial for the other random-effect variances in Eq. 2.

4 Simulation Study

4.1 Methods

4.1.1 Data Generation

We generated data from Eq. 1 using the parameters in Eq. 2. We fixed μ to 0 and drew $N = 30$ values of μ_s from $\mathcal{N}(0, \sigma_{sr}^2 = \frac{1}{2})$, k values of μ_r from $\mathcal{N}(0, \sigma_r^2)$, and $30 \times k$ values from from $\mathcal{N}(0, \sigma_{sr}^2 = \frac{1}{2})$, and used Eq. 1 to obtain scores Y_{sr}. The choice to keep N, σ_s^2, and σ_{sr}^2 constant were arbitrary but believed to be realistic for observational studies with a small number of subjects.

4.1.2 Independent Variables

We varied the number of raters, the variance in the random-rater effects, the hyperprior distributions of the random-effect *SD*s, the type of estimator, and the type of BCI. The number of raters (k) had three levels: k = 2, 3, and 5. We selected k = 2 because two is the minimum number of raters required to estimate the IRR. And k = 2 is often used in the applied literature to estimate IRR. We also specifically incorporated this low number to draw attention to estimation difficulties. We selected k = 3 because a sample size of at least three is required to yield proper posteriors for uniform distributions. We used $k = 5$ to see how the results of different priors differed for slightly higher sample sizes.

The random-rater effect variance (σ_r^2) had two levels: $\sigma_r^2 = .01$ and $\sigma_r^2 = .04$. Random-effect variances extremely close to the lower bound of zero are typically poorly estimated but are less influential in the ICCs. Therefore, we used a value extremely close to zero and a slightly higher value to test whether increasing σ_r^2 improved the estimations. The population ICCs of ICC(A, 1) and ICC(A, k) ranged from 0.48 to 0.83. ICC(C, 1) and ICC(C, k) do not include σ_r^2 (Table 1) and are thus identical across the levels of σ_r^2. Therefore, we further ignored the ICC(C, 1) and ICC(C, k) in this study and focused on the estimation ICC(A, 1) and ICC(A, k).

The hyperprior distributions had three levels: uniform, half-t, and mixed. We specified each of these hyperprior distributions for the random-effect *SD*s (i.e., σ_s, σ_{sr}, and σ_r) rather than the random-effect variances. In the uniform condition, we specified uniform hyperprior distributions over the range $\left[0, \frac{max_Y - min_Y}{2}\right]$ for all random-effect *SD*s, with max_Y and min_Y being estimated from the data. A specified upper bound should not be data dependent, but we could not specify a reasonable data-independent upper bound because our simulated data have no natural boundaries such as Likert scales do. Nonetheless, this data-dependent upper bound will probably behave comparable to natural upper bounds for the random-effect *SD*s, because the specified range still places too much probability mass on unreasonably high values compared to the expected lower values. In the half-*t* conditions, we specified a half-$t(4,0,1)$ hyperprior distributions for all random-effect *SD*s. In the mixed conditions, we used a uniform hyperprior distribution for σ_s and σ_{sr} and a half-*t* hyperprior distribution for σ_r. We refer to Sect. 3 for the justification of these choices.

We obtained point estimates of σ_r and the ICCs using two approaches: posterior means (i.e., expected a posterior; EAPs) and posterior modes (i.e., maximum a posteriori; MAPs). Small sample sizes, such as the small number of raters, can result in skewed posterior distributions, especially for random-effect parameters near the lower bound of zero. Modal estimates resemble their MLE counterparts and are, especially for skewed distributions, preferred over the mean and median as a measure of central tendency (Gelman 2006).

We obtained 95% BCIs using two approaches: Using 2.5% and 97.5% percentiles as limits, and using the highest posterior density intervals (HPDIs). Percentiles are readily provided by most MCMC software but are only appropriate for symmetric unimodal distributions. HPDIs can be obtained from the empirical posterior distribution using kernel densities and accommodate skewness in the posterior distributions. When the posterior is bimodal or non-symmetrical, these approaches may thus yield very different results (for a brief discussion, see Gelman et al. 2013, p. 33).

The simulation design was fully crossed, resulting in 3 (k) × 2 (σ_r^2) × 3 (hyperprior) = 18 between-replications conditions, for each of which we simulated 1000 replications. Within each of the 18 between-replication conditions, we investigated bias for the two types of point estimate (EAPs and MAPs) and coverage rates for the two types of interval estimate (percentiles and HPDI).

4.1.3 Parameter Estimation

We used MCMC estimation of Bayesian hierarchical models and specified the hyperpriors of the random-effect *SD*s as discussed in the previous paragraph. We used three independent chains of 1000 iterations. The first 500 iterations per chain served as burn-in iterations, and the last 500 iterations of each chain were saved in the posterior. This resulted in a posterior of 1500 iterations to estimate each parameter. We checked convergence using the potential scale reduction factor, $\hat{R}$, and the effective sample size, N_{eff}, of each parameter (Gelman 2006). If any of the $\hat{R} < 1.10$, we doubled the number of burn-in iterations. This was repeated until the model converged, or did not converge after the limit of 10,000 burn-in iterations was reached, in which case the replication was discarded. Thereafter, we checked whether each parameter's N_{eff} exceeded 100. If a parameter or ICC showed an effective sample size that was too low, we increased the number of post burn-in iterations based on the lowest N_{eff} with a factor of $120/\min(N_{\text{eff}})$.

4.1.4 Software

We used the R software environment (R Core Team 2019) for data generation and analyses, and the Stan software (Stan Development Team 2017) with the R package `rstan` (Stan Development Team 2018) to estimate the Bayesian hierarchical models and the ICCs. We obtained the MAP estimates using the `modeest` package (Poncet 2019) and 95% HPDIs using the `HDInterval` package (Meredith and Kruschke 2018). Our software code is available on the Open Science Framework (OSF): https://osf.io/shkqm/

4.1.5 Dependent Variables

We evaluated the quality of the estimated ICCs and the random-rater effect *SD*, σ_r,[1] using four criteria: convergence, relative bias, 95% BCI coverage rates, and relative efficiency. We calculated the percentage of converged solutions per condition, which was preferably 100%. Let $\bar{\theta}$ denote the average EAP or MAP estimate of σ_r or the derived ICC across replications in a condition, and let θ denote the population parameter in that condition. We computed relative bias as $\frac{\bar{\theta}-\theta}{\theta}$, and we used relative bias $>$ .05 as indicating minor bias and $>$ .10 as indicating substantial bias. We tested the 95% BCI coverage rates of both percentiles and HPDIs, using a coverage rate $<$ 90% and $>$ 97% as a rule of thumb for defining the width of BCIs as, too narrow or wide BCIs, respectively. We calculated relative efficiency as the ratio of the average posterior *SD* of σ_r and the ICCs, relative to the *SD* of their posterior means.[2] A ratio of 1 indicates accurate estimates of variability. We used relative efficiency $<$ .90 or $>$ 1.10 as indicating minor under- or overestimation of the posterior *SD*s and relative efficiency $<$ .80 or $>$ 1.20 as indicating substantial under- or overestimation of the posterior *SD*s.

4.2 Results

We provide a summary of the simulation results and diverted the complete results to the authors' OSF account. The results for the conditions with mixed hyperprior distributions and the conditions with half-*t* hyperprior distributions for each random-effect *SD* were very similar. Therefore, we do not discuss the results

[1] We focused on σ_r instead of σ_r^2 because our Stan program estimated random-effect *SD*s, from which we derived the random-effect variances.

[2] We want to emphasize the difference between random-effect *SD*s, which quantify the variability of the random effects, and the posterior *SD*s, which quantify the uncertainty about the estimated parameters (the random-effect *SD*s and the ICCs).

of conditions with mixed hyperprior distributions. Similarly, the results for the estimated ICC(A,k) resembled the results for ICC(A,1), so we only present the results for the ICC(A,1).

4.2.1 Convergence

All replications in all conditions converged to a solution. The following results are therefore based on 18 × 1000 = 18,000 replications.

4.2.2 Relative Bias

Figures 1 and 2 show the relative bias of the estimated σ_r and ICC(A,1) across conditions. Both σ_r and ICC(A,1) showed less bias in conditions with a half-t hyperprior distribution than in those with a uniform hyperprior distribution. EAPs severely overestimated σ_r, whereas the MAP was an unbiased estimator of this parameter in all conditions with $k > 2$. The MAP and EAP estimates of ICC(A,1) were comparable. Neither σ_r or ICC(A,1) resulted in unbiased estimates in any condition with $k = 2$. MAPs of both σ_r and ICC(A,1) were unbiased in all conditions with $k = 5$.

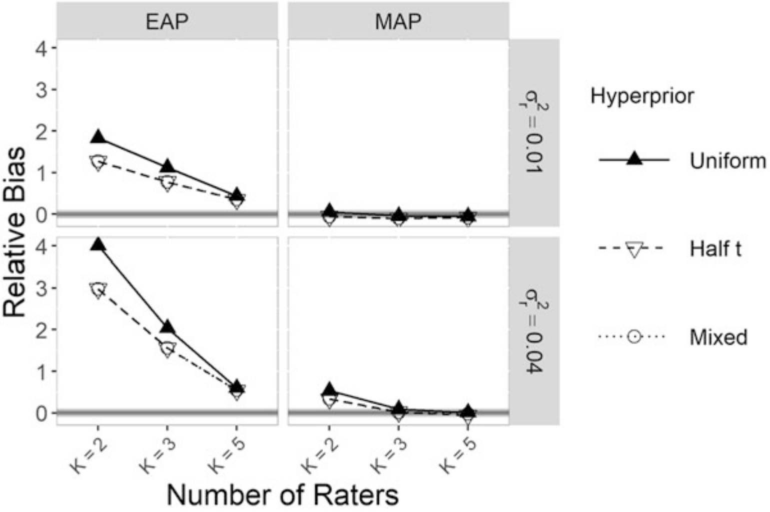

Fig. 1 Relative bias of σ_r under different conditions. White areas, large bias (>10%); light-gray areas, substantial bias (5–10%); dark-gray areas, minor bias (< 5%)

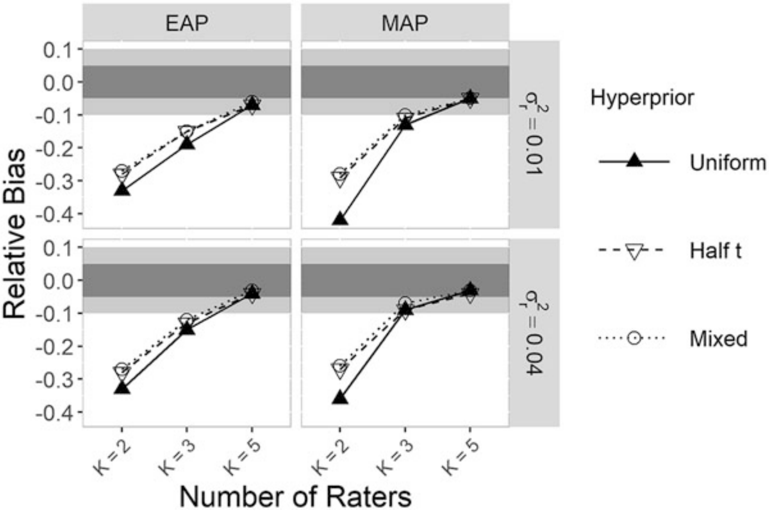

Fig. 2 Relative bias of ICC(A,1) under different conditions. White areas, large bias (>10%); light-gray areas, substantial bias (5–10%); dark-gray areas, minor bias (< 5%)

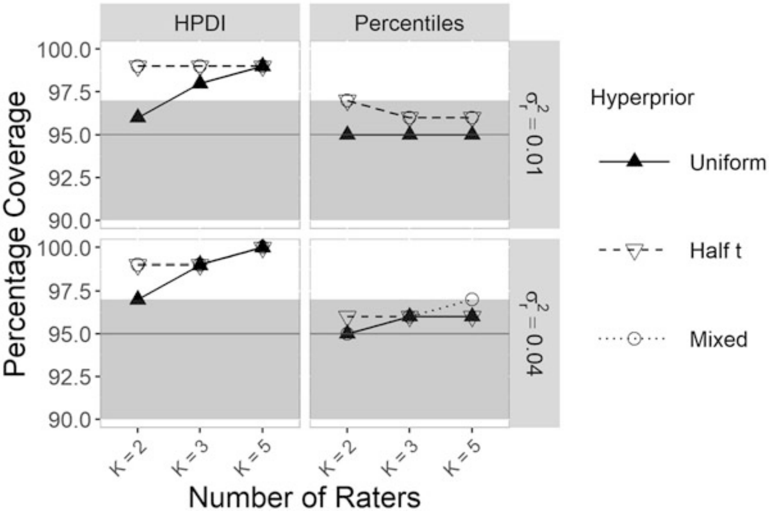

Fig. 3 95% BCI coverage rates of σ_r under different conditions. White areas, substantially too narrow (< 90%) or too wide BCIs (> 97%); light-gray areas, slightly too narrow ($90 \leq 95\%$) or too wide BCIs ($95 > 97\%$)

4.2.3 95% BCI Coverage

Figures 3 and 4 show the 95% BCI coverage rates of σ_r and ICC(A,1), respectively, across conditions. HPDIs were too wide for σ_r but yielded nominal coverage rates

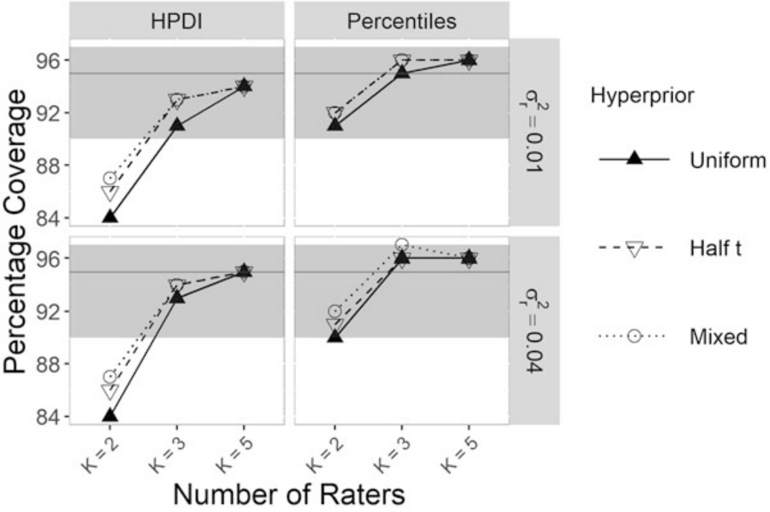

Fig. 4 95% BCI coverage rates of ICC(A,1) under different conditions. White areas, substantially too narrow (< 90%) or too wide BCIs (> 97%); light-gray areas, slightly too narrow ($90 \leq 95\%$) or too wide BCIs ($95 > 97\%$)

for the ICC(A,1) for more than two raters. Percentiles yielded nominal coverage rates for σ_r and for the ICC(A,1) but only for $k > 2$.

4.2.4 Relative Efficiency

Figure 5 shows the relative efficiency of the estimated σ_r and ICC(A,1) across conditions. Both hyperprior distributions yielded posterior *SD*s of both σ_r and ICC(A,1) that were considerably larger than the actual sampling variability of these estimates. The overestimation of posterior *SD*s decreased when k increased but remained severe even in conditions with $k = 5$. Overestimation of the posterior *SD*s was more severe for σ_r than for ICC(A,1) and comparable for both hyperprior distributions.

5 Discussion

The results of this study indicate that half-*t* hyperprior distributions have a slight advantage over uniform hyperprior distributions for estimating IRR with ICCs. The best performing condition combined MAP point estimates, percentiles based BCIs, half-*t* hyperprior distributions, and $k > 2$ raters. For $k = 2$, ICCs were underestimated and inefficient. This bias and inefficiency decreased as k

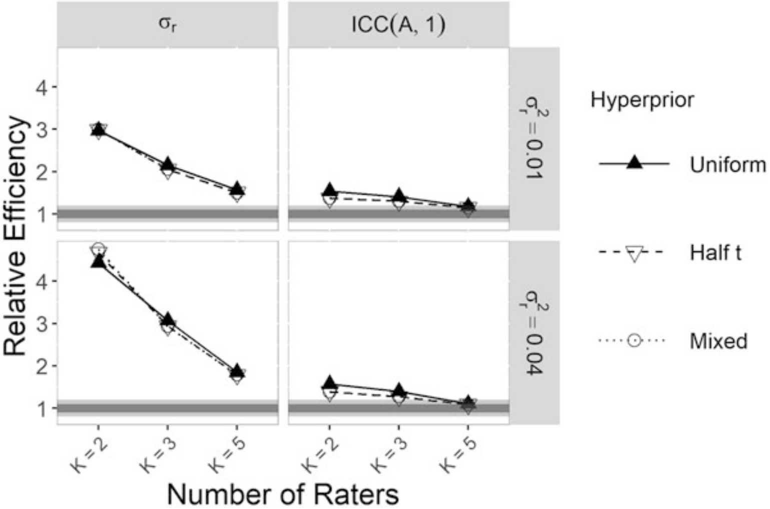

Fig. 5 Relative Efficiency of σ_r and ICC(A,1) under different conditions. White areas: highly inefficient (>20%); Light-gray areas: substantial inefficient (10–20%); Dark-gray areas: slightly inefficient (< 10%)

increased. For $k > 2$ (the conditions with unbiased estimates), the combination of a half-t hyperprior distribution with percentile BCIs yielded nominal coverage rates. Overall, the number of raters used to estimate IRR had a larger effect on the performance of the MCMC estimates than the choice of hyperprior distributions.

The results of this study are in line with earlier research indicating that random-effect variances cannot be properly estimated when the number of clusters (here raters) is as small as two (Gelman 2006). This should discourage researchers from estimating the IRR with ICCs when data are collected from as few as two raters, a situation that we observed frequently in the applied literature. Using $k > 2$ raters in an observational study may sound like a high burden for researchers. Fortunately, estimation of the IRR in conditions with scarce resources could already be improved by randomly sampling a subset of raters for each subject from a larger rater pool (Ten Hove et al. 2019). This would result in a larger rater-sample size, with missing at random data. This resembles an often seen practice (Viswesvaran et al. 2005), which diminishes the burden per rater and allows to keep the total number of observations at the same level as a fully crossed design in which each of two raters rates each subject. It would be interesting to test the combination of the half-t hyperprior distribution with such a planned missing data design in a future study.

Our simulation study was not comprehensive concerning the number of conditions. The performance of the ICCs in our simulation study may thus, for example, depend the population values of the other random-effect variances in the ICCs. Our statements about obtaining (in)appropriate estimates for these ICCs can therefore not readily be generalized to conditions with differing variability in each of the involved effects. However, the results on σ_r itself are promising, because its estimation seems to improve when the variability in the rater effects increases only slightly (at least for $k > 2$). With increasing magnitude, the rater variance had a larger effect on the ICCs, and arguably, the quality of its estimates has a larger influence on the quality of the ICC estimates. Presumably, the ICCs will thus be estimated more accurately and efficiently when the variability in rater effects increases.

In conclusion, we advise researchers to use an half-*t* hyperprior distribution for the random-rater effect *SD*, MAP point estimates, percentiles based BCIs, and, most importantly, at least three raters to estimate the IRR using an MCMC approach. However, we want to highlight Gelman's (2006) advice that every noninformative or weakly informative (hyper)prior distribution is inherently provisional, implying that researchers should always inspect whether their posterior forms a proper distribution. He argued that, if an approach yields improper posteriors, there is more prior information available that needs to be incorporated in the estimation procedure.

References

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*, 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*, 515–534. Retrieved from https://projecteuclid.org/euclid.ba/1340371048

Gelman, A. (2019). *Prior choice recommendations*. Retrieved from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall/CRC.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295–314. https://doi.org/10.1007/s10648-014-9287-x

Meredith, M., & Kruschke, J. (2018). *HDInterval: Highest (posterior) density intervals.* Retrieved from https://CRAN.R-project.org/package=HDInterval (Computer software)

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis, 7*(4), 887–902. https://doi.org/10.1214/12-BA730

Poncet, P. (2019). *modeest: Mode estimation.* Retrieved from https://CRAN.R-project.org/package=modeest (Computer software)

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/ (Computer software)

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2019). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*, 169–191. https://doi.org/10.1080/10705511.2019.1604140

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). New York: Wiley.

Stan Development Team. (2017). *Stan modeling language: User's guide and reference manuals.* Retrieved from https://mc-stan.org/users/interfaces/stan.html (Computer software)

Stan Development Team. (2018). *RStan: The R interface to Stan*. Retrieved from https://mc-stan.org/users/interfaces/rstan.html (Computer software)

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2018). *Interrater reliability for dyad-level predictors in network data.* (Paper presented at the XXXVIII Sunbelt 2018 Conference, Utrecht)

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2019). *Interrater reliability for multilevel data: A generalizability theory approach*. (Paper presented at the 84th annual International Meeting of the Psychometric Society, Santiago, Chile).

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2019). *Interrater reliability for multilevel data: A generalizability theory approach.* (Manuscript submitted for publication)

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology, 89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108–131. https://doi.org/10.1037/0021-9010.90.1.108