Describing posterior distributions of variance components: Problems and the use of null distributions to aid interpretation

Joel L. Pick[1,2,3*], Claudia Kasper[4], Hassen Allegue[5], Niels J. Dingemanse[6], Ned A. Dochtermann[7], Kate L. Laskowski[8], Marcos R. Lima[9], Holger Schielzeth[10], David F. Westneat[11], Jonathan Wright[1], Yimen G. Araya-Ajoy[1,3]

[1] Centre for Biodiversity Dynamics (CBD), Department of Biology, Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway.

[2] Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK

[3] Both authors contributed equally

[4] Animal GenoPhenomics group, Agroscope, Tioleyre 4, CH-1725 Posieux, Switzerland.

[5] Département des Sciences Biologiques, Université du Québec à Montréal, Montréal, QC, Canada

[6] Behavioural Ecology, Department of Biology, Ludwig-Maximilians University of Munich, Planegg-Martinsried, Germany

[7] Department of Biological Sciences, North Dakota State University, Fargo, ND, USA

[8] Department of Evolution and Ecology, University of California Davis, Davis, CA, USA

[9] Departamento de Biologia Animal e Vegetal, Centro de Ciências Biológicas, Universidade Estadual de Londrina, Londrina, Brazil

[10] Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

[11] Department of Biology, University of Kentucky, Lexington, KY, USA

[12] Members of the SQuID working group

[*] Corresponding author, email address: joel.l.pick@gmail.com

Running title: Null distributions for variance components

Keywords: hierarchical models, variance, null distribution, permutation, simulations, squidSim

Data and code availability All code and generated data for the simulated examples are deposited in https://github.com/squidgroup/null_distributions

1 Abstract

Assessing the biological relevance of variance components estimated using MCMC-based mixed-effects models is not straightforward. Variance estimates are constrained to be greater than zero and their posterior distributions are often asymmetric. Different measures of central tendency for these distributions can therefore be very different, and credible intervals cannot overlap zero, making it difficult to assess the the size and statistical support for among-group variance. This is often done through visual inspection of the whole posterior distribution, and so relies on subjective decisions for interpretation. We use simulations to demonstrate the difficulties of summarising the posterior distributions of variance estimates from MCMC-based models. We compare commonly used summary statistics of posterior distributions of variance components showing that the posterior median is predominantly the least biased. We also describe different methods for generating null distributions (i.e. a distribution of effect sizes that would be obtained if there was no among-group variance) that can be used to aid in the interpretation of variance estimates. We further show how null distributions could be used to derive a p-value that provides complimentary information to the commonly presented measures of central tendency and uncertainty and also facilitates the implementation of power analyses within an MCMC framework.

2 Introduction

Estimating variance components using mixed-effects models is common in ecology and evolution (Bolker *et al.*, 2009; Harrison *et al.*, 2018). Mixed-effect models are a flexible statistical tool used to study hierarchically structured data, including extensions for estimating quantitative genetic parameters (animal models; Henderson, 1988; Kruuk, 2004) and comparative analysis (meta-analysis and phylogenetic mixed models; Hadfield & Nakagawa, 2010). Markov chain Monte Carlo (MCMC) algorithms are increasingly used to fit mixed-effects models, due to their flexibility and availability of open-source software (e.g. winBUGS (Gilks *et al.*, 1994), JAGS (Plummer, 2003), MCMCglmm (Hadfield, 2010), Stan (Stan Development Team, 2022b)). MCMC algorithms are a collection of probabilistic simulation methods for generating observations from designated statistical distributions and are typically implemented within a Bayesian framework (Gelman *et al.*, 2013).

MCMC methods have many advantages in ecology and evolution. For instance, we are commonly interested in derived measures such as a standardised measure of variance (e.g. repeatability, heritability and evolvability Nakagawa & Schielzeth, 2010; Houle, 1992). These derived measures can be estimated using the whole posterior distribution of their components, allowing uncertainty to be propagated both within and among analyses. In contrast, in a maximum likelihood framework, the methods to estimate the uncertainty of derived metrics (such as the delta method) can be biased with small sample sizes (O'Hara *et al.*, 2008). Data in ecological and evolutionary studies are also commonly non-Gaussian, for example counts (e.g. number of offspring), binary and ratio data (e.g. survival, presence/absence, sex ratio) and categorical data (e.g. colour morphs, horn type in sheep). The performance of MCMC algorithms in generalized linear mixed-effects models has been found to be superior in terms of accuracy and precision compared with Restricted Maximum Likelihood (REML) approaches (O'Hara & Merilä, 2005; de Villemereuil *et al.*, 2013). Bayesian methods also allow existing information to be incorporated as a prior distribution, although this has rarely been used in ecological or evolutionary studies (Lemoine, 2019).

Despite these clear advantages, there are several issues that empiricists face when using MCMC mixed-effect models. Here we address the issue that variance estimates and their uncertainty can be hard to describe and interpret, especially when trying to assess their biological relevance. We highlight two problems that can occur when estimating variance components, both of which centre around the difficulty of describing the posterior distribution of variance components using summary statistics: (i) finding an appropriate measure of central tendency; and (ii) assessing the statistical support for non-zero among-group variance. These problems stem from variance estimates being constrained to be greater than zero and that their posterior distributions are often asymmetric.

In order to describe the posterior distribution, we often present some measure of central tendency alongside some measure of uncertainty (quantile-based intervals or Highest Posterior Density (HPD) intervals). The posterior mean, median and mode have all been used as measures of central tendency, and more recent works have suggested the general use of the posterior median (Gelman *et al.*, 2020; McElreath, 2020). There is, however, no clear guidance on which measure provides a more appropriate summary statistic for variance components, although in our experience the mode and mean are most commonly reported. When the posterior distribution of a variance component is far away from zero and is symmetric, then the mean, median and mode are approximately equal (Figure 1a) and inferences are robust to the choice of central tendency metric. However, when variances are small (relative to the total variance) and/or there are small sample sizes (both of which often occur in ecology and evolution), the posterior distributions can be close to zero. As variances are constrained to be greater than zero, these posterior distributions are typically asymmetric and can even be bimodal. Consequently, there can be a considerable difference between the mean, median and mode, with the mode often lying close to zero (Figure 1b). This discrepancy makes it is difficult to draw inference about the magnitude of the posterior variance estimate.

Use of the posterior mode is often justified as being the closest to the maximum likelihood estimate (MLE) when uninformative priors are used. However, this comparison refers to the joint posterior mode, rather than the marginal mode that is typically estimated and reported. In more complex models, the joint and marginal modes may differ (Held & Sabanés Bové, 2020, Section 6.5.4), meaning that the marginal mode and MLE are no longer the same. As shown in Figure S2, the convergence of the posterior mode and MLE also requires the use of uninformative improper priors, which are generally not advised (Gelman *et al.*, 2013) and are thus seldom used. The posterior mode is also hard to estimate; it is typically done using kernel density estimation and different methods may provide quite different estimates (Figure 2), thereby providing an additional source of hidden ambiguity. Furthermore, the mode requires a larger number of samples in the posterior distribution to be reliably estimated, and will show greater variation between models/chains run on the same dataset (Kruschke, 2015). In contrast, the mean is strongly affected by extreme values, and so by the long tail of an asymmetric distribution.

It is also often important to assess statistical support for among-group variance at a particular level. Typically 95% credible intervals (CRIs) are presented as a measure of uncertainty in parameter estimates derived from MCMC model fits. As variance components cannot overlap zero, CRIs give no information about the compatibility of the estimates with the null hypothesis (no among-group variance). Posterior distributions are often inspected visually, as histograms or density plots, in order to assess whether the distributions are biased towards zero, which is commonly assumed to signify that the estimated variance is not different from zero. What is seldom appreciated, however, is that the degree of smoothing that is applied in such plots (via the binning interval or bandwidth) can alter these conclusions. This means that the same distribution can be seen as uni- or bimodal, or peaking at zero or away from zero (Figure 2). Such assessments therefore tend to be highly subjective and lack a proper quantitative basis.

To address this, several methods for generating metrics for assessing the confidence in a result (such as p-values) have been suggested in a Bayesian framework (reviewed in Makowski *et al.*, 2019a). Two of these, Region of Practical Equivalence (ROPE) and Bayes factors, can be used for variance components. The ROPE approach identifies a range of values considered negligible or too small to be of any practical relevance (i.e. the Region of Practical Equivalence), and quantifies the proportion of overlap between the posterior distribution and the ROPE. This is similar to equivalence testing in a Frequentist framework, specifically to the two one-sided tests (TOST) approach (Lakens *et al.*, 2018). Bayes factors are analogous to Frequentist likelihood ratios, comparing different models (for example with and without the random effects of interest), but unlike likelihood ratios they incorporate information from the prior distributions into the comparison of the models (Morey *et al.*, 2016). Both of these metrics can be used to provide a measure of statistical support for estimates of variance components, but their implementation is complicated - ROPE requires the definition of a threshold, incorporating further subjectivity into the analysis, whilst the computation of Bayes factors can be challenging, and even not implementable in some commonly used programs (e.g. MCMCglmm). We discuss these two methods further in the discussion.

Here we suggest a complementary method to assess statistical support in mixed-effect models, which compares the estimated variance components to a null distribution in order to inform the statistical inferences made from the model. This involves creating a distribution of effect sizes that would be expected under the null hypothesis (no among-group variance) and comparing this null distribution with the observed among-group variance. This method has several advantages. Null distributions can be used to generate a p-value describing the probability that the observed estimate is as or more extreme than expected under the null hypothesis. Although often criticised through their association with Null Hypothesis Significance Testing (NHST; Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019), p-values have well understood and useful properties. When correctly interpreted, these test statistics provide a useful tool by indicating how consistent an observed effect size is with a scenario in which there is no among-group variance. In contrast to the ROPE method, the creation of a null distribution requires no subjective decisions about thresholds and, in contrast to Bayes Factors, they can be implemented using the output from any Bayesian model.

We present two methods, permutation and simulation, for generating null distributions for variance components. When generating a null distribution using permutation, some feature of the data or data structure is randomised to produce a new dataset that contains the structure of the original dataset, but where there is no relationship between the response variable and the variable of interest (the among-group variance in this case). This randomization is repeated a large number of times (e.g. 1000) to create many different permuted datasets. The same analysis is then carried out on the permuted datasets as on the original dataset, and a test statistic of interest (e.g. the estimate of among-group variance) is used to create a null distribution of test statistics (Figure 1c,d). A (one-tailed) p-value can then be derived as the proportion of permuted datasets with a test statistic greater than or equal to the test statistic observed with the real data set. Permutation tests have already been suggested as an alternative to likelihood ratio tests for frequentist analyses (Fitzmaurice *et al.*, 2007; Samuh *et al.*, 2012), although they are not commonly utilized in ecology and evolution (but see Araya-Ajoy & Dingemanse, 2017; Stoffel *et al.*, 2017). Permutation tests are a subclass of nonparametric tests (Pesarin & Salmaso, 2010; Lehmann & Romano, 2005) and do not rely on specific probability distributions, and so make few assumptions. However, as we show later in the manuscript, datasets can be permuted in several different ways when the data structure is complex, and the consequences of the choices involved in such cases are often not immediately obvious. An alternative method of creating a null distribution is using simulations. This process is similar to permutation, but instead of generating permuted datasets we can simulate datasets from the observed model parameters (in a similar way to parametric bootstrapping), whilst setting the variance in question to zero. This simulation method makes more assumptions about the data and model, but allows for more control of the manipulated features of the simulated datasets compared with permutations.

Finally, a crucial part of designing experiments and statistical analyses is assessing the power to detect an effect size of interest. Power is defined as the probability of rejecting the null hypothesis (i.e. no among-group variance) for a given effect size at a specified alpha level (typically 0.05), and so is dependent upon the generation of p-values. Although power relates to NHST and the often criticized alpha level (Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019), it and analogous metrics (Gelman & Carlin, 2014) remain an important tool for study design regardless of statistical philosophy, and this is because it provides a quantitative approach to calculating optimal sample sizes and designing sampling regimes. Power may also provide a more useful metric than precision when considering variance components. As their distributions are bounded at zero, standard errors will always decrease when distributions are close to zero (see Supplementary Figure S4). However, the concept of power for variance components in MCMC models is not well developed. As null distributions can be used to generate p-values, they also provide a convenient way of conducting power analysis.

Here, we first compare commonly used summary statistics of posterior distributions of variance components. We then demonstrate the utility of null distributions (i.e. a distribution of effect sizes that would be obtained if there was no among-group variance) to generate a complementary p-value statistic and aid the interpretation of the variance components. Comparison with a null distribution provides a quantitative measure of confidence that the observed variance component is larger than what might be expected under the null hypothesis, given the data structure and priors used. Importantly, we are not advocating that this approach should replace the presentation and use of effect sizes (e.g. posterior mean/median/mode) and credible intervals, but rather that it should be used as an additional and complementary statistic. Finally, we show how null distributions can be used to perform a power analysis within an MCMC framework.

3 Methods

3.1 Generation of Simulated Datasets

Simulated datasets were generated out in R (version 4.1.0 R Core Team, 2022) using the squidSim R package (version 0.1.0 Pick, 2022). We first simulated Gaussian data with one hierarchical level and varied the number of observations per group (2 and 4) and the number of groups (20, 40 and 80). We simulated a total variance of 1 and varied the among-group variance (0, 0.1, 0.2 and 0.4; since the total variance simulated was 1, these are also the respective intra-class correlations (ICCs)/repeatabilities). We simulated every combination of these parameters (24 parameters sets) and for each set we simulated 500 datasets. Power to detect among-group variance is known to be determined by effect size and sample size both within and among groups. We deliberately chose these parameter values and sample sizes to explore scenarios where power is low (Dingemanse & Dochtermann, 2013) to understand the impact on posterior distributions. These sample sizes also correspond to typical experimental designs in behavioral ecology or life history data collected on wild populations (Bell *et al.*, 2009).

We analysed each simulated dataset with a linear mixed-effect model specifying group level random effects in a Bayesian framework, using Stan with the rstan package (version 2.21.3 Stan Development Team, 2022a). We specified weakly informative priors on the among-group and residual standard deviations (half-Cauchy distribution with scale 2), and ran one chain for each model with 5000 iterations and a warm-up period of 2000 iterations. Across the majority of models (95%) this ensured an effective sample size in the posterior distribution of the among group variance of >500. For comparison, we also ran REML models using the lmer function of the lme4 package (version 1.1-29 Bates *et al.*, 2015), the results of which are shown in the Supplementary Figure S1.

As a demonstration that our findings hold with more complex data, we simulated Bernoulli data (binomial with one observation) with 80 groups and 4 observations per group. Among-group effects were simulated from a Gaussian distribution on the latent scale, with among-group variances of 0 and 0.2. The latent scale response variable was then transformed using the inverse logit function to provide the probabilities, and sampled with a Bernoulli process. We simulated 100 datasets for each variance, and analysed the data as outlined above.

3.2 Comparison of Posterior Distribution Summary Statistics

From the posterior distributions of the among-group variances, we calculated the posterior mean, median and mode, and compared these estimates with the simulated values.

While calculating the mean and median of the posterior distribution is straightforward, there are several ways of estimating the mode of the marginal posterior distribution, which involve some (hidden) assumptions. Commonly used functions in R include the `posterior.mode` function in the MCMCglmm package (Hadfield, 2010), the `Mode` function in the ggdist package (Kay, 2022), and the `map_estimate` function of the bayestestR package (Makowski *et al.*, 2019b). Typically these functions estimate the mode by estimating the parameter value at which the kernel density is maximised. Kernel density estimation essentially involves fitting a model to the distribution of posterior samples to estimate a density function. The maximum of this function (the estimated mode) is then calculated over a series of predicted values. One key parameter in kernel density estimation is the bandwidth, which essentially describes the amount of smoothing and is analogous to the number of breakpoints in a histogram (Figure 2). Common methods generally scale bandwidth generated by specific algorithms. MCMCglmm scales the bandwidth generated by Silverman's 'rule of thumb' algorithm (nrd0; eqn 3.31 in Silverman, 1986) by 0.1 (i.e. it is much less smoothed; Figure 2d). In contrast, ggdist and bayestestR use the default values of the nrd0 and SJ algorithms (Sheather & Jones, 1991), respectively (the default bandwidth of the nrd0 algorithm is also used by `density` function in R; Figure 2a). The impact on the potential inferences caused by the choice of scaling is demonstrated in Figure 2, with the degree of smoothing affecting where the posterior mode is estimated. To explore this impact of bandwidth, we estimated the posterior mode using these two bandwidth scalings (0.1 and 1). The kernel density was estimated using the SJ algorithm (Sheather & Jones, 1991), and the mode was estimated using 512 predicted values with a cut point at zero. These additional parameters all differ between commonly used functions, but have much smaller impacts upon the results than the bandwidth, and so we hold them constant here.

To ensure that our results, especially on the mode, were not driven by the choice of the prior, we ran additional models on a subset of the data (ICC=0.2, N groups=80, N within=2) with a half-Cauchy prior with scale 5 and 25, and uniform priors from 0 to 5 and 0 to 25 on the among-group standard deviation. The half Cauchy prior has been recommend for variance components (Gelman, 2006) and is commonly used (note it is equivalent to the parameter expanded priors in MCMCglmm). For demonstration purposes, we also ran models in MCMCglmm specifying uninformative improper priors. Given the simplicity of these models, the posterior mode is expected to correspond to the REML estimate. The different parametrizations of the half Cauchy and uniform priors resulted in no difference in the results (Figure S2). As expected, using an uninformative improper prior led to a concordance between REML and posterior mode, although the strength of this similarity differed between the different methods used to estimate the mode (Figure S2).

To compare these different measures of central tendency, we calculated the bias as $\frac{1}{n}\sum\hat{\theta}_i - \theta$ (where θ is the true simulated value, $\hat{\theta}_i$ is the model estimate from ith simulation in a parameter set, and n is the number of simulations). For the non-zero effect sizes, we also calculated relative bias $\frac{1}{n}\sum\frac{\hat{\theta}_i-\theta}{\theta}$ and absolute relative bias $\frac{1}{n}\sum\frac{|\hat{\theta}_i-\theta|}{\theta}$. We also calculated the precision as $1/\sqrt{\frac{1}{n}\sum(\hat{\theta}_i - \bar{\theta})^2}$, which we present in the Supplementary Figure S4.

3.3 Creation of null distributions and p-values

We created null distributions for each simulated dataset using two methods. First, we permuted the datasets by shuffling the group indices (IDs) to create 100 new datasets, each of which was analysed in the same way as the original dataset. From each permuted dataset, we extracted the same parameters (the estimates of central tendency in the posterior distribution of the among-group variance) as for models fitted to the original data and created the corresponding null distributions. Second, we used simulations to create the null distribution. To do this, we simulated datasets with no among-group variance. To ensure the same total variance we added the posteriors of the among-group and residual variances of the original model, and we used the median of the resulting distribution as our inputted value for the simulated residual variance in the null model. The choice of the median for this step should have little consequence, as this derived distribution will be estimated with much less uncertainty and so will be symmetric, meaning that the three measures of central tendency will be equivalent. Each simulated null dataset was analysed in the same way as the original dataset, and we extracted the same parameters to create the corresponding null distributions.

Although we recommend using a larger number of permutations/simulations to build up a null distribution in empirical studies (e.g. 1000), here we used 100 permutations and simulations to generate null distributions for these simulated datasets in order to reduce the computational burden (500 simulations for 4 variances, with 6 different sample sizes is 12000 datasets, for each of which we performed 100 permutations and 100 simulations). We then calculated a p-value for each original dataset, as the proportion of estimates in the null distribution that were higher than the estimate from the original data. We calculated p-values using each central tendency measure, and these are compared in Figure S5.

3.4 Power analysis

Using the simulated datasets outlined above, we compared two ways by which power can be calculated. Power is defined as the probability of rejecting the null hypothesis (i.e. no among-group variance in this case) for a given effect size and data structure at a specified alpha level (typically 0.05). To do this, we calculated the proportion of datasets in which the p-value was below a nominal threshold of 0.05. It is worth noting that, although power has a superficial connection with NHST, power can also be seen as a description of the distribution of p-values expected for a given effect size and data structure. Other descriptions of this distribution (e.g. the mean) would be simple functions of the power. We therefore chose to present power as a description of the distribution of p-values as it is conceptually well understood and frequently used, rather than due to any philosophical alignment with NHST.

First, we estimated power using the p-values generated though comparison with the null distributions from both permutation and simulation approaches outlined above ('full' method). We were also able to calculate the false positive rate for this method (essentially the power when the simulated value is 0). Second, we used the model estimates from the simulated datasets with zero among-group variance for each data structure (combination of among- and within-group sample sizes) as a null distribution, against which the estimates from simulated datasets with among-group variance could be tested ('reduced' method). This method of estimating power is similar to the simulation method of generating null distributions, but involves generating one null distribution for *all* datasets with the same data structure, instead of null distribution for *each* dataset. It is therefore massively less computationally intensive for power analyses, because to explore power within the parameter space presented here it only required the running 12,000 models, rather than 1,212,000. It is not possible to calculate a false positive rate for this method, as this would involve comparing the null distribution with itself, and so the false positive rate would be 5%, by definition.

4 Results

4.1 Comparing summary statistics of the posterior distribution

When the simulated among-group variance was zero, all summary statistics were upwardly biased to some extent (the posterior distribution cannot include 0; Figure 3a). Predictably, the posterior mean and median from datasets with zero variance were considerably more upwardly biased for small sample sizes, in contrast to the mode. The mean was the most biased as it is heavily influenced by the tail of the distribution. Consequently, this upward bias is stronger when the uncertainty is high (i.e. when the tail is large). Note, however, that this upward bias is also present in Frequentist analyses (see Figure S1), and is not just a feature of Bayesian analyses.

When the simulated among-group variance is non-zero, then the mean, median and mode all appeared to be consistent estimators, in that any bias occurred only at small sample and/or effect sizes. The posterior median generally converged on the simulated value at lower effect and sample sizes (Figure 3b), as compared with the posterior mean, which was upwardly biased, and the posterior mode that was biased towards zero (Figure 3b).

When considering the absolute relative bias (Figure 3c), the mean and median show very similar levels of bias, with exception of the lowest sample and effect size combination where the mean was more biased. This suggests that although the mean is more likely to be upwardly biased, the magnitude of the bias is similar in the two measures. However, the mode is consistently more biased than the other measures (Figure 3c), although this bias disappears at higher sample and effect sizes. Following the example shown in Figure 2, the bias in the mode depends upon the bandwidth that was used, with higher smoothing showing less bias across the two bandwidths tested. We found similar patterns in our Bernoulli simulations (Figure 5a).

4.2 Performance of the null distributions

As expected, both permutation and simulation methods produced a uniform distribution of p-values when applied to datasets where the simulated among-group variance was zero (Figures 4). The distribution of p-values from both tests then shifts towards zero as the sample size and the magnitude of the variance increase (Figure 4). Similar patterns were found in the Bernoulli simulations (Figure 5b).

Importantly, although the mean, median and mode were often quite different in magnitude (reflecting skew in the posterior distribution), the inference based upon the p-values did not differ between the different metrics. There were strong correlations between p-values estimated with the different metrics, with the exception of the mode estimated with less smoothing (see Figures S5 and S7). P-values were also strongly correlated between null distributions generated through simulation and permutation methods (see Figures S6 and S8).

4.3 Power analyses

When considering the full method of estimating power, both ways of generating null distributions (permutation and simulation) gave very similar results (Figure 6), with marginally higher power for the permutation method. These power estimates are very similar to previous published estimates for Frequentist models (Dingemanse & Dochtermann, 2013). These methods also displayed the expected false positive rates (5%) under all simulated conditions (black points in Figure 6). The reduced method for estimating power, using the same null distribution for all simulation datasets within a particular data structure, generally gave a similar power to the other methods (Figure S9).

As with the p-values, power was not particularly sensitive to the measure of central tendency used, the highest power being seen in the mode with higher smoothing and the lowest power with the mode with less smoothing (Figure S9).

5 Worked example - Random slopes

As is often the case, the examples presented above are simplistic and empiricists commonly encounter more complex questions and data structures in their studies. Here we outline a more realistically complex example where the permutation of datasets require some careful decisions.

Random slope models (where group-specific intercepts and slopes are modelled, also known as random regression) provide a good example of this complexity. We will focus here on generating a null distribution for the estimate of among-group variance in slopes. This estimate is based upon the relationship between the predictor variable and response, the distribution of the response variable across groups, and the distribution of the predictor variable within and across groups. This provides us with four possibilities for permutation: 1) permuting the response variable (retains data structure and breaks all relationships with response); 2) permuting the predictor (retains the group data structure, breaks link between predictor and response, and the distribution of the predictor across groups); 3) permuting the group identities (breaks the group data structure, but retains link between predictor and response); and 4) permuting the predictor within groups (retains the group data structure and the distribution of the predictor across groups, but breaks link between predictor and response). Additionally, we can also generate a null distribution through simulation, where we can simply simulate no among-group variance in slopes, adding the variance generated by the random slopes to the residual to ensure the same total phenotypic variance. Below we explore these different null distributions using a simulated and a real data set. Null distributions were generated based upon the analyzes of 100 null datasets.

5.1 Simulated dataset

We imagined a hypothetical researcher measuring the body mass of a bird species at different times of the day with the aim of studying how temperature affects body mass. The question of interest was to asses whether there is variation among individuals in how temperature affects their body mass. The (simulated) observed data set consisted of 300 individuals measured 4 times each. Body mass and temperature were both normally distributed. Temperature was scaled to have a mean of 0 and variance of 1, and has an effect on body mass of 0.2 for the average individual. The simulated among individual variance in the intercepts was 0.2 and the phenotypic variance generated by variation in slopes was 0.1 (with no correlation among random slopes and intercepts), while the residual variance was set to 0.7 to ensure a total phenotypic variance not explained by the average effect of the environment was 1. Formulas to estimate the total phenotypic variance in random slope models can be found in Allegue *et al.* (2017) There were no systematic differences in the average temperature experienced by the different individuals.

5.2 Real world dataset

For our example with real data, we used a study on variation in the plastic aggressive response to intruders of great tits (*Parus major*) in a nestbox population in southern Germany (Araya-Ajoy & Dingemanse, 2017). Aggressiveness data were collected over a 6-year period (2010–2015) for all male birds during their first breeding attempt each year. The aggression test started when a taxidermic mount of a male great tit was presented on a 1·2 m wooden pole with a playback song 1 m away from the subject's nest box. They subsequently recorded the behaviour of the focal male for a period of 3 min after it had entered a 15 m radius around the box. Simulated territorial intrusions were performed twice during the egg-laying stage and twice during the egg-incubation stage of each focal nest. Therefore, males had repeated measures both within- and among-years.

We analysed 2854 aggression tests performed to 1042 breeding attempts of 679 individuals. The average number of years for which we obtained an individual's reaction norm was 1·4, with 513, 142, 44, 8, 8 and 1 individual(s) sampled for one, two, three, four, five or six breeding attempt(s) (years), respectively. On average, we acquired 2·8 (out of 4) data points for male aggressiveness per breeding attempt (i.e. year), because males did not always respond to the territorial intrusion experiment (Araya-Ajoy & Dingemanse, 2017). Details of the experimental setup, and assayed behaviours, are provided in Araya-Ajoy & Dingemanse (2014). For the purpose of this paper, we used the subject's minimum distance to the mount as a measure of aggressiveness because previous work implies that this behaviour represents a reliable predictor of the intensity of an aggressive response in both stages of breeding (Araya-Ajoy & Dingemanse, 2014).

5.3 Random slope methods

Both datasets were analysed using random slope mixed-effects models, specifying the environmental predictor (temperature for the simulated example and breeding stage for the real example) as a fixed covariate, and random intercepts and environment slopes across individuals. We then generated 5 null distributions (4 permutations and 1 simulation), as outlined above, with which we compared the estimate of among individual variance in slopes from the observed data. Breeding stage (egg-laying versus egg-incubation) was first coded as zero (for laying) versus one (for incubation), and subsequently mean centred and standardized to standard deviation units. Models were fitted in a Bayesian framework, using Stan with the rstan package (version 2.21.3 Stan Development Team, 2022a). We specified weakly informative priors on the among-group and residual standard deviation. We ran three chains for the model of the simulated and real observed data with 5,500 iterations and a warm-up period of 500 iterations. To decrease computational burden, the models for the permuted/simulated data sets were run for only one chain. We then generated five null distributions of posterior medians for each dataset, using the methods described above.

5.4 Random slope results

The different types of null distributions provided the same qualitative results, supporting the conclusion regarding among-individual variation in slopes, in both the real and simulated datasets (Fig 7). For these datasets, permuting individual identity created null distributions with a larger mean value of random slope variance (see Discussion for an explanation). It is important to note that these results relate only to this specific example and may not generalize to other studies. We therefore recommend exploring the particular consequences of using different types of permutations for specific datasets, if a reader wishes to use a permutation method.

6 Discussion

Through the use of simulations, we demonstrate the difficulties of summarising the posterior distributions of variance estimates from MCMC-based models. We describe different methods for generating null distributions that provide useful complimentary information alongside the presentation of central tendency and uncertainty that are generally reported. We also show a way in which null distributions could be used to derive a p-value, which is an easy addition to the statistics presented when summarizing a posterior distribution and also facilitates power analysis.

6.1 Summary statistics

Our experience in ecology and evolution is that both posterior mean and mode are commonly, but inconsistently, presented without justification. For fixed effect parameter estimates, this is typically inconsequential, as the posteriors are usually symmetrically distributed. When estimating variance components, however, our simulations show that depending upon the underlying parameter value, both of these measures can show large biases in opposite directions. When posterior distributions are close to zero and there *is* among-group variance, the posterior mode is very biased towards zero, whereas the posterior median and mean perform much better. On the other hand, if there is no among-group variance, the mode is by far the least biased. The mode, however, suffers further from subjectivity in its estimation. Our simulations also show that the estimation of the mode depends on the underlying algorithm for mode estimation. Unfortunately, the method of mode estimation is rarely justified or even stated in empirical papers. The mode also requires larger posterior distributions to be reliably estimated and will show greater variation between models/chains (Kruschke, 2015). Given this hidden ambiguity in the estimation of the mode, we would therefore cautiously recommend the presentation of the posterior median, or both median and mean, as a measure of central tenancy for variance components. This recommendation is based upon the median being generally less biased than the mean when power is low. Presenting both allows the discrepancy to be seen, showing that the distribution is near to zero and not symmetric, further stressing the uncertainty in these measures.

Upward biases in variance components have been seen before when power is low, but the dependence on the choice of the central tendency metric has not been highlighted. For example, Fay *et al.* (2022) note overestimation of variance components in Bernoulli models, with this overestimation decreasing in size as sample size and effect size increase. Fay *et al.* (2022) use the posterior mean as a summary statistic, and (as we show in Supplementary Figure S10) this bias will decrease (although not disappear completely) through the use of a posterior median. This is not just a bias in Bernoulli models, or in fact MCMC models (Figure S1), but a general property of variance components estimated with low power.

It is often argued that rather than presenting summary statistics, we should present and interpret the whole posterior distribution, which are frequently presented using density plots. Again, the underlying parameters of the kernel density estimation are usually not presented alongside the density plots, meaning the amount of smoothing is not documented. A large degree of smoothing can hide asymmetry and/or bi-modality, and so change inferences. We therefore suggest the use of histograms over density plots in the presentation of posterior distributions, because although they are subject to the same smoothing problems, the degree of smoothing is explicit in the histogram, but hidden in the density plot. Alternatively, other plots that explicitly show the raw posterior samples (e.g. beeswarm plots) could be used (e.g. Figures 4 and 5).

6.2 Null distributions

The null distribution approaches outlined here are relatively easy to use, and allow quantification of confidence that a variance estimate is the result of a biological process rather than a consequence of the choice of priors and data structure. Importantly, the p-values based upon null distributions are not dependent upon what measure of central tendency is used. Such inferential statistics comparing the observed estimates with the null distributions can provide quantitative measures that can be reported alongside the observed estimates and uncertainty, and provides a useful tool for assessing the probability that variance components are non-zero and thereby supplement visual inspections of posterior distributions, or comparison of posterior mode, median and mean. Furthermore, they can serve as an objective and easy-to-communicate assessment of the biological relevance of an estimated variance component to the general public and policy makers, or for the statistical support of non-zero values for derived statistics like heritability, repeatability or evolvability. Common criticisms of p-values include that they are often misinterpreted or used for NHST. We would therefore recommend readers thinking of using the null distribution approach to acquaint themselves with the literature on these topics (some useful examples include Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019). Importantly, p-values cannot demonstrate absence of effect, just confidence in difference from the null hypothesis (here no among-group variance). We believe generating null distributions will help empiricists understand these concepts, as they can be used to give a visual representation of what a p-value signifies.

As we illustrate in our examples of random slopes, there are different ways of permuting datasets, which become more varied as the complexity of the data structure and model increase. Our example on random slope analysis demonstrated that these differences can lead to qualitatively similar results, although whether they always or usually do so would require a much broader set of simulations than we report here. Interestingly, permuting individual identity created null distributions with noticeably larger values of random slope variance. We believe this is due to the existence of random slopes in the simulated and real data set generating heterogeneous residuals (i.e. variance in response changed with the environmental predictor) that were confounded with random slope variation in the analyses of the null data sets (similar effects are also shown in Ramakers *et al.*, 2020). The other permutation methods break up the relationship between the predictor and response, and so the average estimate for the null distributions was lower. This illustrates how comparing the results of the different methods of null distributions generation may provide insights that may be used to inform the statistical inferences from estimated variance components.

In some instances, generating a null distribution using permutations may not be possible. For example, in event-history models of survival (where individuals have an entry for each time point where they are observed, in a sequence of 0's for time points they survive and a 1 for the time point after which they die). In this case, permuting the individual identifiers would fundamentally alter the data structure, meaning that some individuals had multiple deaths. This could be made to work in the context of an animal model, where the observed 0's and 1's could be interchanged between individuals, so that the same between individual structure was maintained, but the link with the pedigree was broken. This serves to demonstrate that some care needs to be taken when assessing the suitability of permutations and how they impact the data structure on a case-by-case basis. Overall, we are not advocating a specific recipe for permutations here - it is likely context and question dependent. We instead advocate a simulation approach at the planning stage, using simulations to check in advance that the permutation design gives desired properties with your likely data structure.

Generating null distributions through simulation avoids many of the issues with the permutation approach, although it does not account so well for the particularities of each data set. Simulation has the advantage that it allows the structure of the data to be fully retained, a more fine-scale alternation of the variances in question, and it makes no additional assumptions than those already being made by the statistical model itself. Reassuringly, in our random slope example, the null distributions generated using the simulation method were similar to the other methods. We therefore cautiously recommend the use of this simulation method, as it is the most flexible for complex models.

These null distribution approaches are, however, computationally intensive and applying them can take a long time depending upon the model complexity, the amount of data and the available computational resources. MCMC methods are often used for highly complex problems (e.g. double hierarchical GLMs; Cleasby *et al.*, 2015), where running a large number of permutations may not be an option. The number of permutations/simulations that are run affects the precision with which a p-value can be calculated and the minimum p-value that can be calculated - a null distribution of 100 can have a minimum p-value of 0.01 and vary by intervals of 0.01. This is why we would recommend a higher number of samples in the null distributions than we used here. However, we were able to produce meaningful results with 100 simulations, and even a few permutations/simulations would give some idea (although much less reliably) of the compatibility of the observed variance with the range expected under the null hypothesis.

6.3 Alternative approaches

A p-value is defined as the probability that an estimate equal to or more extreme than the observed estimate would occur under the null hypothesis (i.e. if the true among-group variance is zero). It relies upon the distribution of p-values being uniform when the null hypothesis is true, a property that is expected to be invariant to sample size (as we show in Figure 4). P-values therefore provide support for the alternative hypothesis, but they do not provide support for the null hypothesis. The ROPE value and Bayes factors aim to assess actual support for the null hypothesis, and therefore depend upon sample size. Below we outline the potential issues that empiricists may encounter when trying to employ these methods.

The ROPE introduces another source of subjectivity into the analysis, because it involves an arbitrary threshold that needs to be defined. This is not trivial in the case of variance components, as small variances can have large knock-on effects. For example, McFarlane *et al.* (2015) find that maternal genetic effects account for 2% of variation in fitness, but this small amount predicts a 56% increase in mean lifetime reproductive success in less than 10 generations, which is highly biologically meaningful. Bonnet *et al.* (2022) address this by using simulations to demonstrate the biological relevance of the thresholds they use (0.01 and 0.001, for the variances not ICC). There is also discussion about whether the overlap of the whole posterior or the 95% credible interval should be used with ROPE (Makowski *et al.*, 2019a; Schwaferts & Augustin, 2020). As with NHST, 95% is also an arbitrary cutoff, and so the ROPE would represent the overlap of two arbitrary thresholds. ROPE is often discussed in a context where a cost-benefit analysis can be used to work out the minimum effect size that warrants the use of a particular intervention, for example of medical interventions (Kruschke, 2018). Typically this is not relevant for research in ecology and evolution as, in many cases, it is of interest whether variance in a particular component exists, and if so its magnitude. We think there is clear application for using ROPE in fields like conservation, where interaction with stakeholders requires thresholds over which decisions need to be made, but for many empiricists, ROPE requires more subjective decisions to be made and justified.

Bayes factors can be used to test the 'significance' of parameters in Bayesian mixed-effect models. However, the calculation of Bayes factors that allow inferences to be made about variance components is not straightforward. They require large posterior distributions for stable estimation and are sensitive to both prior and model specification (Gelman *et al.*, 2013; Navarro, 2019; Schad *et al.*, 2022). Bayes factors are also not implementable in all programs, including commonly used programs in ecology and evolution (e.g. MCMCglmm). Our approach provides an alternative to this method, which is easily implemented and allows straightforward interpretation with reference to the probability that the estimate obtained is the consistent purely with the data structure and model specification.

6.4 Power analysis and possible alternatives

Power analysis is controversial as it relies on NHST. NHST is controversial because its misuse has been attributed to scientific misconduct and the replication crisis (Wasserstein & Lazar, 2016; Amrhein *et al.*, 2017; McShane *et al.*, 2019; Amrhein *et al.*, 2019), issues which relate to the use of p-values *after* data collection and analysis. Power analysis, however, serves a clear purpose in aiding experimental design, and is conducted *pre*-analysis, and so is perhaps not subject to the same criticisms. Suggested alternatives, such as Type M and Type S error, also rely upon calculation of p-values and definition of an arbitrary alpha value, and are both a simple function of power (Gelman & Carlin, 2014). Type S error (proportion of significant estimates that have the opposite sign) is not relevant for variance components. Type M (absolute relative bias of significant estimates) gives some additional information but, unlike power, it is affected by the measure of central tendency that is chosen (Figure S11). Power can also be seen as a description of the distribution of p-values expected for a given effect size and data structure. Other descriptions of this distribution (e.g. the mean) would be simple functions of the power, but the common use of this metric makes it more widely understood. An alternative to power would be to design studies around a desired level of precision in estimates. Although this works for unbounded parameters, precision is difficult to interpret for variance components, and SE will decrease as true value gets closer to zero, not because precision increases, but because it is limited by zero (see Figure S4). We would therefore suggest that power still provides a suitable metric for designing studies to estimate variance components.

We show two methods of power analysis based upon null distributions. The first (full) involves generating p-values for each simulated dataset by generating a null distribution for that dataset. This method is highly computationally intensive as it involves running a certain number of simulations multiplied by the number of permutations/simulations models, which could realistically be one million models per parameter. Our alternative method (reduced) is to generate a single null distribution for each data structure, and generate p-values by comparing the parameter estimates from the simulated datasets to this single null distribution. This method gives similar results to the first approach and is massively less computationally intensive (requiring running 2000 models rather than a million for each set of parameters). The disadvantage is that the false positive rate cannot be calculated.

Even if power is not the intended use (or there is an objection to arbitrary alpha values), these simulations can serve an extremely useful purpose before studies are conducted. First, these simulations allow a empiricist to consider the distribution of p-values expected under a given effect size and design (note power is essentially a descriptions of the shape of this distribution). Second, the null distribution of point estimates can be considered - this enables the distribution of effect sizes that can occur under the null hypothesis to be visualised. Even if an empiricist does not want to calculate a p-value, creating a null distribution is still a powerful way of seeing the distribution of estimates that would be generated with no among-group variance, and would serve to encourage caution in how results that lie within that distribution are interpreted.

6.5 Recommendations

1. Use of posterior median as a measure of central tendency for posterior distributions of variance components from MCMC-based models. Our results show that the median is the least biased estimate, but will overestimate variances when power is low. Reporting multiple measures of central tendency allows any asymmetry in the posterior to be made obvious.

2. Reporting of smoothing values in kernel estimation. Kernel density estimation is commonly used for estimating the posterior mode and creating density plots. The parameters used in this estimation are seldom reported, but can have a large impact on interpretation. We advise the reporting of parameters in the kernel density estimation, or the use of more explicit methods of plotting posterior distributions, such as histograms.

3. Using null distributions for inference. Null distributions provide a way of putting the observed parameter estimates into a context expected under an explicitly defined null hypothesis (i.e. no among-group variance). Null distributions can be created in multiple ways, but they are most easily controlled when generated using simulations. As with many aspects of statistical analysis, there are many decisions relating to generating null distributions that may have an affect on the results. Therefore, these methods should be defined pre-analysis, in order to reduce researcher degrees of freedom.

4. Using a null distribution to estimating power. As well as aiding *post-hoc* inference, null distributions can be used for power analysis. We provide details of a method for doing so that does not present a large computational burden.

Acknowledgments

We would like to thank the other members of the Statistical Quantification of Individual Differences (SQuID) working group, and the Wild Evolution and Statistics in Ecology and Evolution Discussion groups at the University of Edinburgh for valuable feedback on the ideas presented here. SQuID workshops in 2022 and JLP were funded by a Research Council of Norway INTPART project number 309356 grant to JW. YGA was supported by the Research Council of Norway project number 325826. HS were supported by the German Research Foundation (DFG, 215/543-1, 316099922). DFW was supported by the U.S. National Science Foundation. HA was supported by the Natural Sciences and Engineering Research Council of Canada (CGSD3-504399-2017) and the Fond de Recherche du Québec - Nature et Technologies (FRQNT; 283511).

References

Allegue, H., Araya-Ajoy, Y.G., Dingemanse, N.J., Dochtermann, N.A., Garamszegi, L.Z., Nakagawa, S., Réale, D., Schielzeth, H. & Westneat, D.F. (2017) Statistical Quantification of Individual Differences (SQuID): an educational and statistical tool for understanding multilevel phenotypic data in linear mixed models. *Methods in Ecology and Evolution*, **8**, 257–267. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12659, https://dx.doi.org/10.1111/2041-210X.12659.

Amrhein, V., Greenland, S. & McShane, B. (2019) Scientists rise up against statistical significance. *Nature*, **567**, 305–307. https://dx.doi.org/10.1038/d41586-019-00857-9.

Amrhein, V., Korner-Nievergelt, F. & Roth, T. (2017) The earth is flat ($p > 0.05$): Significance thresholds and the crisis of unreplicable research. *PeerJ*, **5**, e3544. https://dx.doi.org/10.7717/peerj.3544.

Araya-Ajoy, Y.G. & Dingemanse, N.J. (2014) Characterizing behavioural 'characters':

an evolutionary framework. *Proceedings of the Royal Society of London Series B*, **281**, 20132645. https://dx.doi.org/10.1098/rspb.2013.2645.

Araya-Ajoy, Y.G. & Dingemanse, N.J. (2017) Repeatability, heritability, and age-dependence of seasonal plasticity in aggressiveness in a wild passerine bird. *Journal of Animal Ecology*, **86**, 227–238. https://dx.doi.org/10.1111/1365-2656.12621.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://dx.doi.org/10.18637/jss.v067.i01.

Bell, A.M., Hankison, S.J. & Laskowski, K.L. (2009) The repeatability of behaviour: a meta-analysis. *Animal Behaviour*, **77**, 771–783. Publisher: Elsevier Ltd ISBN: 0003-3472, https://dx.doi.org/10.1016/j.anbehav.2008.12.022.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.S.S. (2009) Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135. https://dx.doi.org/10.1016/j.tree.2008.10.008.

Bonnet, T., Morrissey, M.B., de Villemereuil, P., Alberts, S.C., Arcese, P., Bailey, L.D., Boutin, S., Brekke, P., Brent, L.J.N., Camenisch, G., Charmantier, A., Clutton-Brock, T.H., Cockburn, A., Coltman, D.W., Courtiol, A., Davidian, E., Evans, S.R., Ewen, J.G., Festa-Bianchet, M., de Franceschi, C., Gustafsson, L., Höner, O.P., Houslay, T.M., Keller, L.F., Manser, M., McAdam, A.G., McLean, E., Nietlisbach, P., Osmond, H.L., Pemberton, J.M., Postma, E., Reid, J.M., Rutschmann, A., Santure, A.W., Sheldon, B.C., Slate, J., Teplitsky, C., Visser, M.E., Wachter, B. & Kruuk, L.E.B. (2022) Genetic variance in fitness indicates rapid contemporary adaptive evolution in wild animals. *Science*, **376**, 1012–1016. https://dx.doi.org/10.1126/science.abk0853.

Cleasby, I.R., Nakagawa, S. & Schielzeth, H. (2015) Quantifying the predictability of behaviour: Statistical approaches for the study of between-individual variation in the within-individual variance. *Methods in Ecology and Evolution*, **6**, 27–37. https://dx.doi.org/10.1111/2041-210X.12281.

de Villemereuil, P., Gimenez, O. & Doligez, B. (2013) Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: A simulation study for Gaussian and binary traits. *Methods in Ecology and Evolution*, **4**, 260–275. https://dx.doi.org/10.1111/2041-210X.12011.

Dingemanse, N.J. & Dochtermann, N.A. (2013) Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, **82**, 39–54. https://dx.doi.org/10.1111/1365-2656.12013.

Fay, R., Authier, M., Hamel, S., Jenouvrier, S., van de Pol, M., Cam, E., Gaillard, J.M., Yoccoz, N.G., Acker, P., Allen, A., Aubry, L.M., Bonenfant, C., Caswell, H., Coste, C.F.D., Larue, B., Le Coeur, C., Gamelon, M., Macdonald, K.R., Moiron, M., Nicol-Harper, A., Pelletier, F., Rotella, J.J., Teplitsky, C., Touzot, L., Wells, C.P. & Sæther, B.E. (2022) Quantifying fixed individual heterogeneity in demographic parameters: Performance of correlated random effects for Bernoulli variables. *Methods in Ecology and Evolution*, **13**, 91–104. https://dx.doi.org/10.1111/2041-210X.13728.

Fitzmaurice, G.M., Lipsitz, S.R. & Ibrahim, J.G. (2007) A Note on Permutation Tests for Variance Components in Multilevel Generalized Linear Mixed Models. *Biometrics*, **63**, 942–946. https://dx.doi.org/10.1111/j.1541-0420.2007.00775.x.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition.

Gelman, A., Hill, J. & Vehtari, A. (2020) *Regression and Other Stories*. Cambridge University Press, Cambridge.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

Gelman, A. & Carlin, J. (2014) Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, **9**, 641–651. https://dx.doi.org/10.1177/1745691614551642.

Gilks, W.R., Thomas, A. & Spiegelhalter, D.J. (1994) A Language and Program for Complex Bayesian Modelling. *Journal of the Royal Statistical Society Series D (The Statistician)*, **43**, 169–177. https://dx.doi.org/10.2307/2348941.

Hadfield, J.D. (2010) MCMC methods for multi-response generalized linear mixed models: The {MCMCglmm} {R} package. *Journal of Statistical Software*, **33**, 1–22. https://dx.doi.org/10.1002/ana.23792.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508. https://dx.doi.org/10.1111/j.1420-9101.2009.01915.x.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., Robinson, B.S., Hodgson, D.J. & Inger, R. (2018) A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, **6**, e4794. https://dx.doi.org/10.7717/peerj.4794.

Held, L. & Sabanés Bové, D. (2020) *Likelihood and Bayesian Inference*, volume 10. Springer.

Henderson, C.R. (1988) Theoretical basis and computational methods for a number ofdifferent animal models. *Journal of Dairy Science*, **71**, 1–16.

Houle, D. (1992) Comparing evolvability and variability of quantitative traits. *Genetics*, **130**, 195–204. https://dx.doi.org/citeulike-article-id:10041224.

Kay, M. (2022) *ggdist: Visualizations of Distributions and Uncertainty*. R package version 3.2.0.

Kruschke, J. (2015) *Doing Bayesian Data Analysis.* Acadmiec Press/Elsevier, second edition.

Kruschke, J. (2018) Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, **1**, 270–280. https://dx.doi.org/10.1177/2515245918771304.

Kruuk, L.E.B. (2004) Estimating genetic parameters in natural populations using the "animal model". *Philosophical Transactions of the Royal Society of London B*, **359**, 873–890. https://dx.doi.org/10.1098/rstb.2003.1437.

Lakens, D., Scheel, A.M. & Isager, P.M. (2018) Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, **1**, 259–269. https://dx.doi.org/10.1177/2515245918770963.

Lehmann, E.L. & Romano, J.P. (2005) *Testing Statistical Hypotheses.* Springer Texts in Statistics. Springer, New York, 3rd ed edition.

Lemoine, N.P. (2019) Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, **128**, 912–928. https://dx.doi.org/10.1111/oik.05985.

Makowski, D., Ben-Shachar, M.S., Chen, S.H.A. & Lüdecke, D. (2019a) Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, **10**, 2767. https://dx.doi.org/10.3389/fpsyg.2019.02767.

Makowski, D., Ben-Shachar, M.S. & Lüdecke, D. (2019b) bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, **4**, 1541. https://dx.doi.org/10.21105/joss.01541.

McElreath, R. (2020) *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Chapman and Hall/CRC, 2nd edition.

McFarlane, S.E., Gorrell, J.C., Coltman, D.W., Humphries, M.M., Boutin, S. & Mcadam, A.G. (2015) The nature of nurture in a wild mammal's fitness. *Proceedings of the Royal Society of London B*, **282**, 1–7.

McShane, B.B., Gal, D., Gelman, A., Robert, C. & Tackett, J.L. (2019) Abandon Statistical Significance. *The American Statistician*, **73**, 235–245. https://dx.doi.org/10.1080/00031305.2018.1527253.

Morey, R.D., Romeijn, J.W. & Rouder, J.N. (2016) The philosophy of bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, **72**, 6–18. https://dx.doi.org/10.1016/j.jmp.2015.11.001.

Nakagawa, S. & Schielzeth, H. (2010) Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society*, **85**, 935–56. https://dx.doi.org/10.1111/j.1469-185X.2010.00141.x.

Navarro, D.J. (2019) Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, **2**, 28–34. https://dx.doi.org/10.1007/s42113-018-0019-z.

O'Hara, R.B., Cano, J.M., Ovaskainen, O., Teplitsky, C. & Alho, J.S. (2008) Bayesian approaches in evolutionary quantitative genetics. *Journal of Evolutionary Biology*, **21**, 949–957. https://dx.doi.org/10.1111/j.1420-9101.2008.01529.x.

O'Hara, R.B. & Merilä, J. (2005) Bias and precision in QST estimates: Problems and some solutions. *Genetics*, **171**, 1331–1339. https://dx.doi.org/10.1534/genetics.105.044545.

Pesarin, F. & Salmaso, L. (2010) *Permutation Tests for Complex Data*. John Wiley & Sons, Ltd, first edition.

Pick, J.L. (2022) *squidSim: a flexible simulation tool for linear mixed models*. R package version 0.1.0.

Plummer, M. (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria*, **124**.

R Core Team (2022) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramakers, J.J.C., Visser, M.E. & Gienapp, P. (2020) Quantifying individual variation in reaction norms: Mind the residual. *Journal of Evolutionary Biology*, **33**, 352–366. https://dx.doi.org/10.1111/jeb.13571.

Samuh, M.H., Grilli, L., Rampichini, C., Salmaso, L. & Lunardon, N. (2012) The Use of Permutation Tests for Variance Components in Linear Mixed Models. *Communications in Statistics - Theory and Methods*, **41**, 3020–3029. https://dx.doi.org/10.1080/03610926.2011.587933.

Schad, D.J., Nicenboim, B., Bürkner, P.C., Betancourt, M. & Vasishth, S. (2022) Workflow techniques for the robust use of bayes factors. *Psychological Methods*, pp. No Pagination Specified–No Pagination Specified. Place: US Publisher: American Psychological Association, https://dx.doi.org/10.1037/met0000472.

Schwaferts, P. & Augustin, T. (2020) Bayesian decisions using regions of practical equivalence (rope): Foundations.

Sheather, S.J. & Jones, M.C. (1991) A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society Series B (Methodological)*, **53**, 683–690.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stan Development Team (2022a) RStan: the R interface to Stan. R package version 2.21.3.

Stan Development Team (2022b) Stan modeling language users guide and reference manual. Version 2.3.

Stoffel, M.A., Nakagawa, S. & Schielzeth, H. (2017) rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **8**, 1639–1644. https://dx.doi.org/10.1111/2041-210X.12797.

Wasserstein, R.L. & Lazar, N.A. (2016) The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, **70**, 129–133. https://dx.doi.org/10.1080/00031305.2016.1154108.

7 Figures

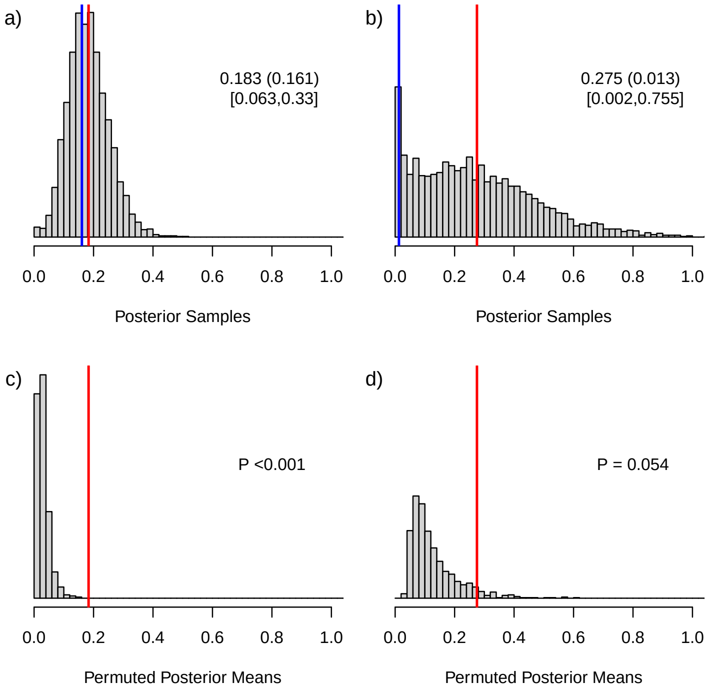

Figure 1: Posterior distributions of variance estimates for two different scenarios (a and b) and their respective null distributions (c and d) generated using permutations. Example a) shows a symmetric posterior distribution far away from zero with close agreement between the posterior mean (red lines) and mode (blue line), whilst b) shows an asymmetric posterior distribution close to zero, with clear divergence between the posterior mean and mode. c) and d) show null distributions of posterior means generated through permuting the datasets, and corresponding p-values, of a) and b), respectively. The values given in a) and b) correspond to mean (mode) [CRIs]. Both datasets were simulated with among-group variances of 0.2, but with differing sample sizes; a) with 80 groups and 4 observations per group; b) with 40 groups and 2 observations per group.

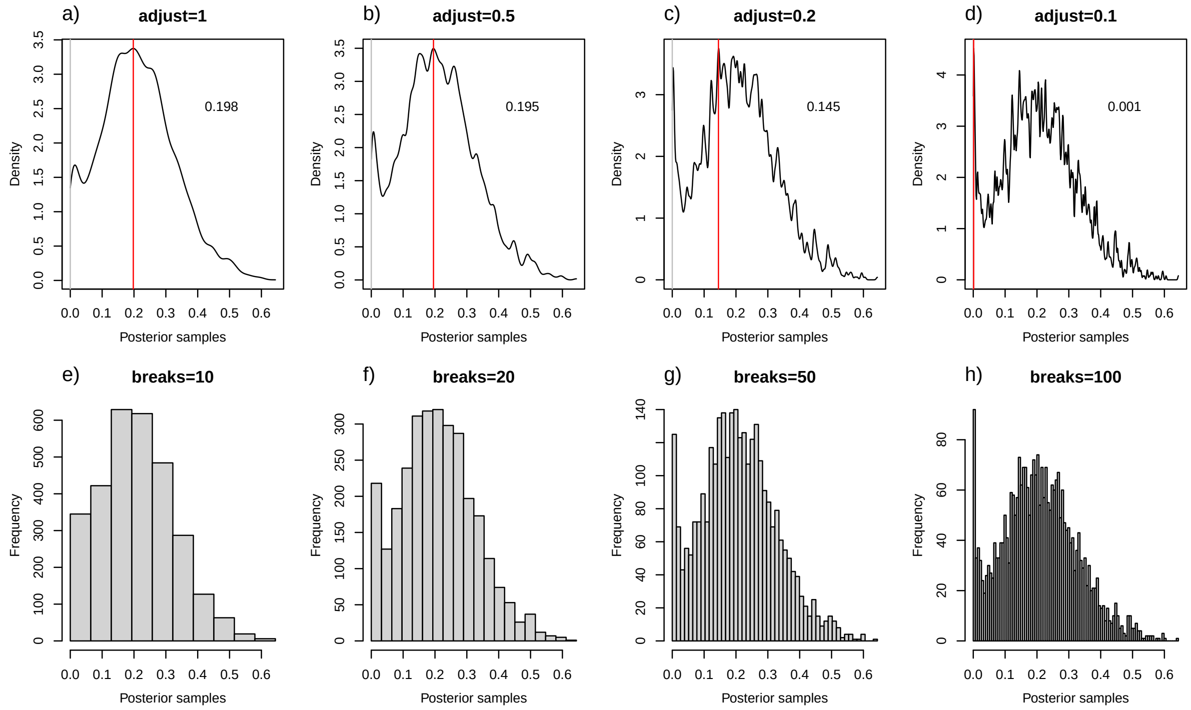

Figure 2: The effect of bandwidth choice on the estimation of the posterior mode. Top row shows kernel densities of the same posterior distribution, estimated with different bandwidth scalings, from 1 in a) to 0.1 in d). Red lines shows the posterior modes estimated from that scaling. Bottom row shows the equivalent histograms for comparison.

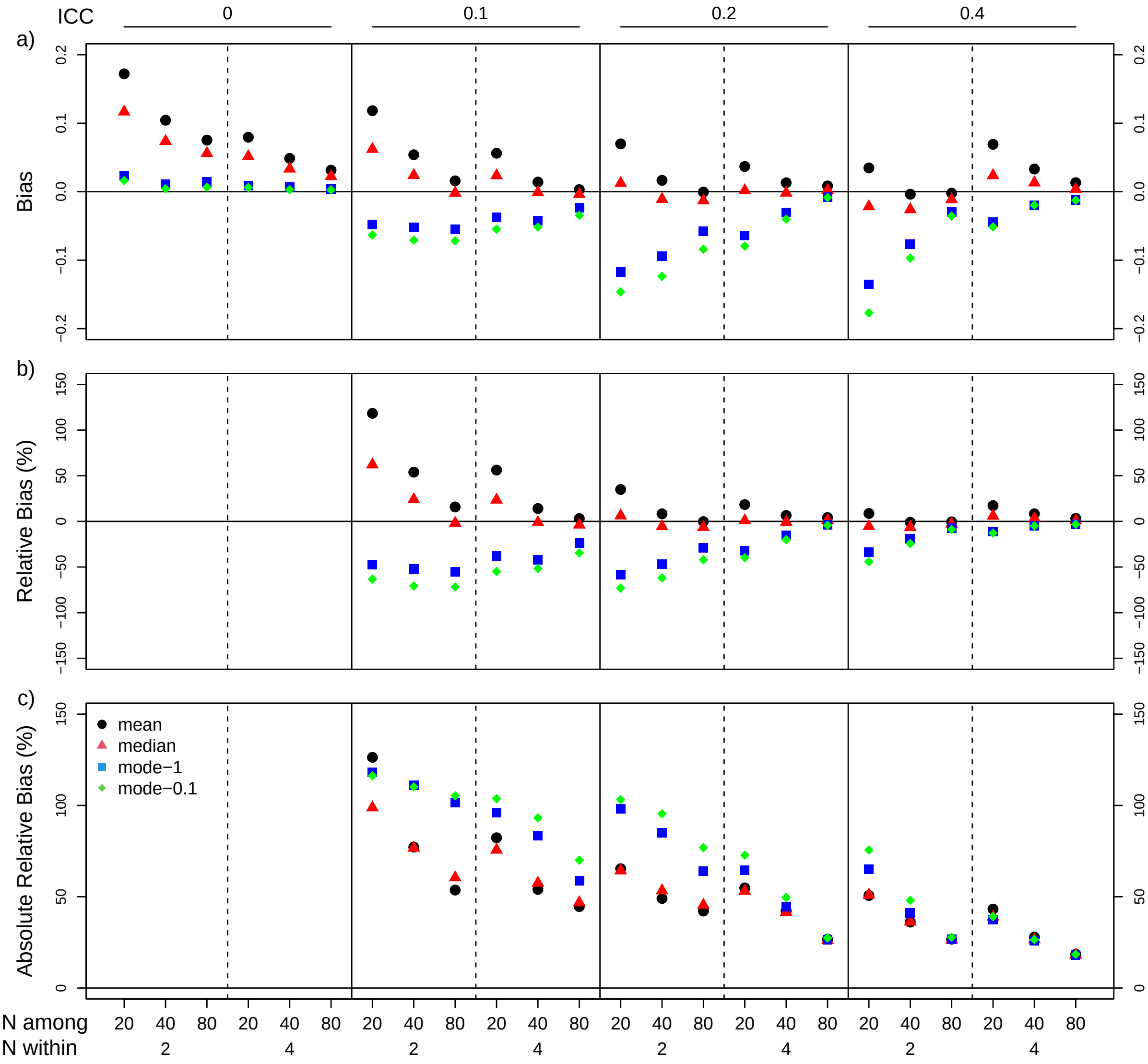

Figure 3: Bias (a), relative bias (b) and absolute relative bias (c) of posterior mean, median and mode of variance components from simulations varying in among group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Two posterior modes were estimated: mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details).

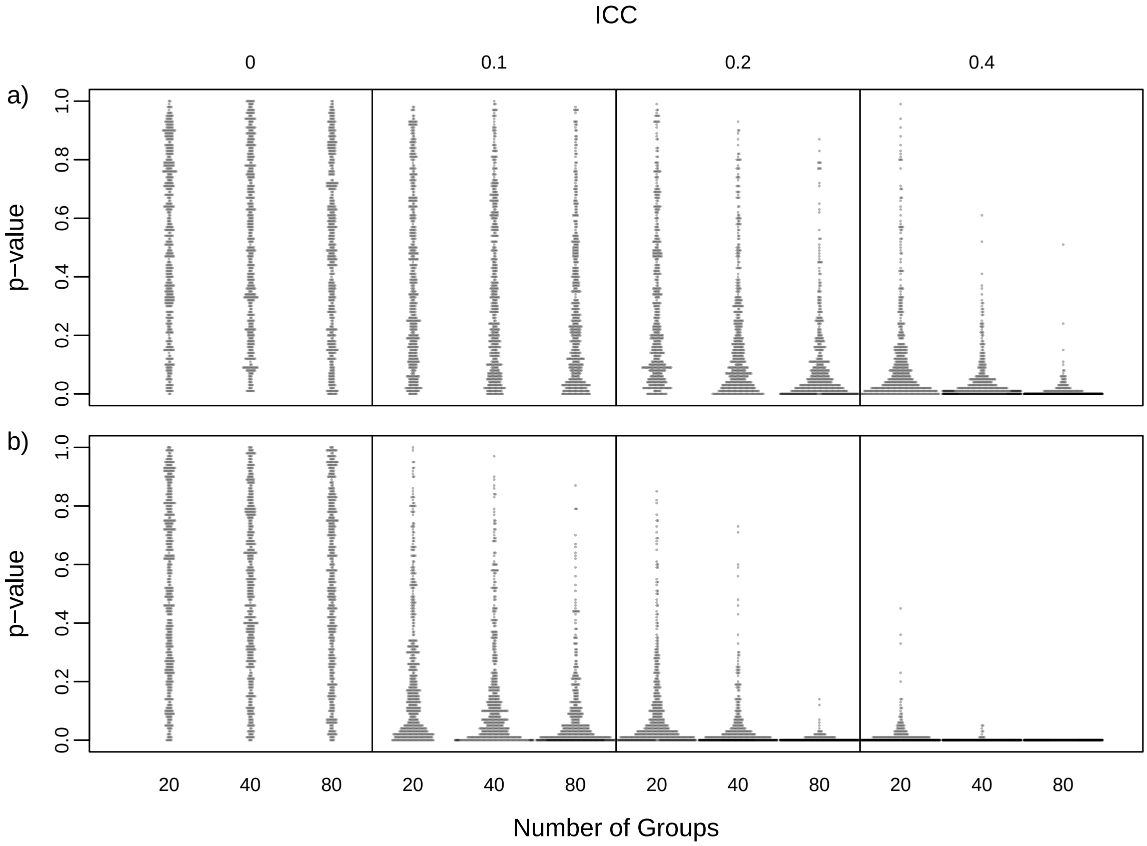

Figure 4: Distribution of p-values estimated using the posterior median and null distributions generated through simulations for datasets varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size among (20, 40, 80) groups. Example a) shows a within group sample size of 2, and b) a within group sample size of 4.

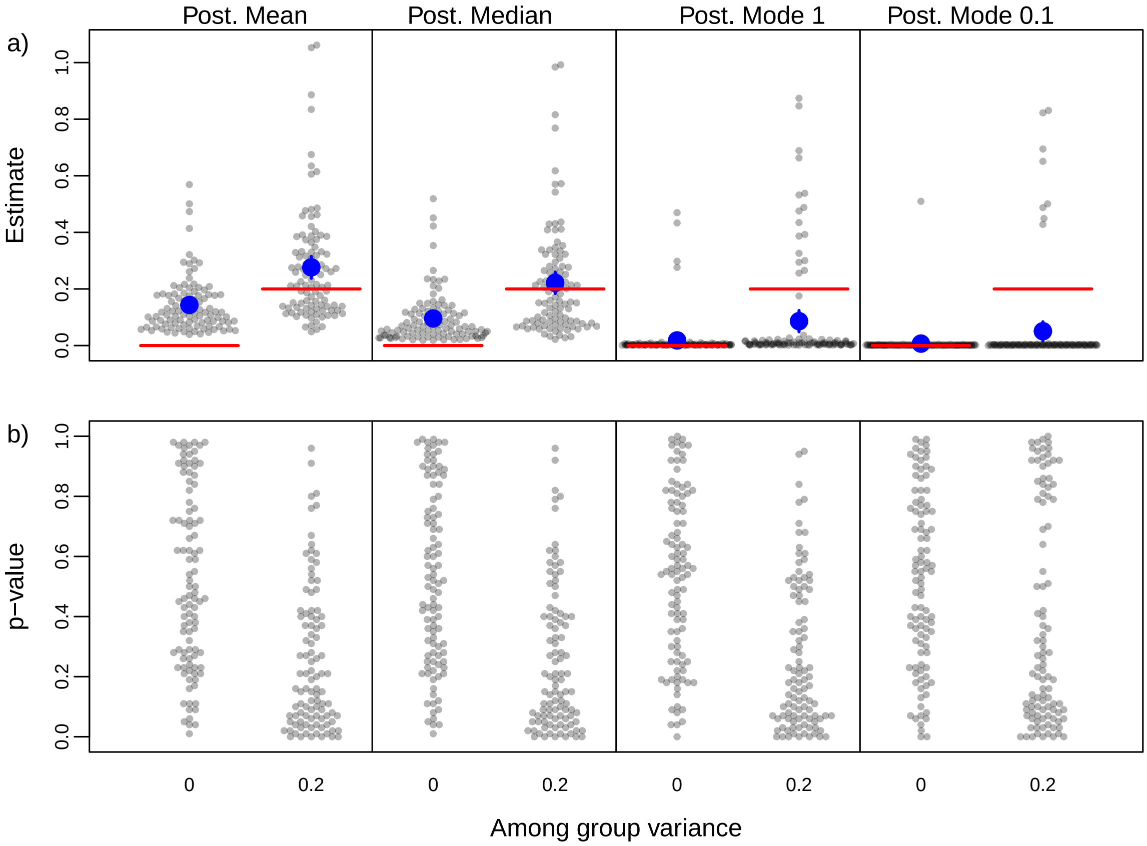

Figure 5: Sampling distributions of parameter estimates (a) and p-values (b) from GLMMs using different measures of central tendency. Two posterior modes were estimated: mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details). In a) red lines show simulated values, and blue points and error bars show mean and standard error of the sampling distributions. The p-values were generated using null distributions generated through simulation.

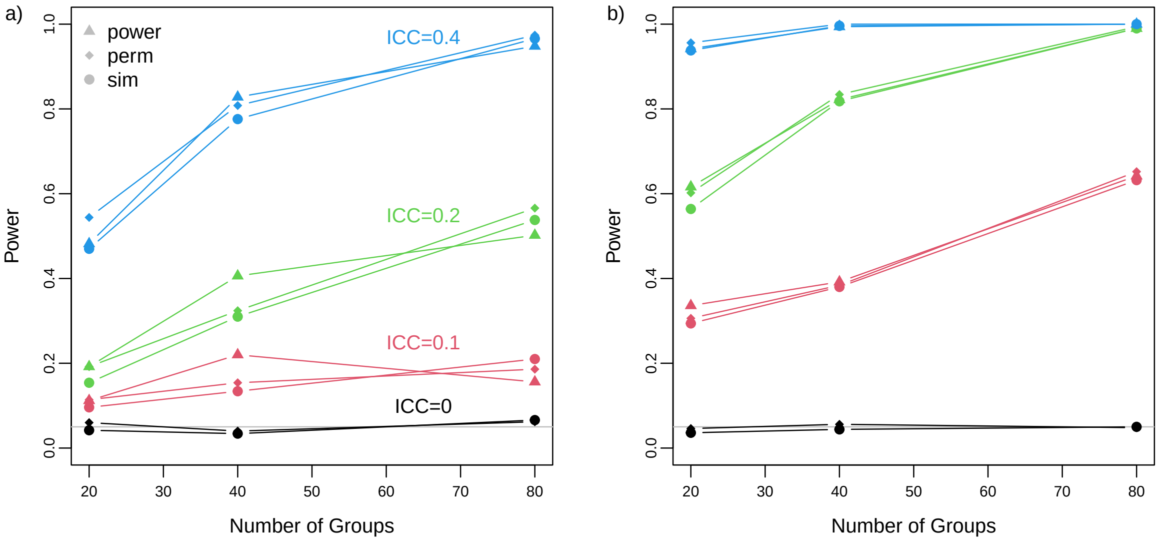

Figure 6: Comparisons of power calculated using permutation (perm), simulation (sim) or a global null distribution (power). For each within-group sample size of a) 2 and b) 4, we show results for four among-group variances (0, 0.1 ,0.2 and 0.4) and three among-group sample sizes (20, 40 and 80). Power was calculated using posterior medians.

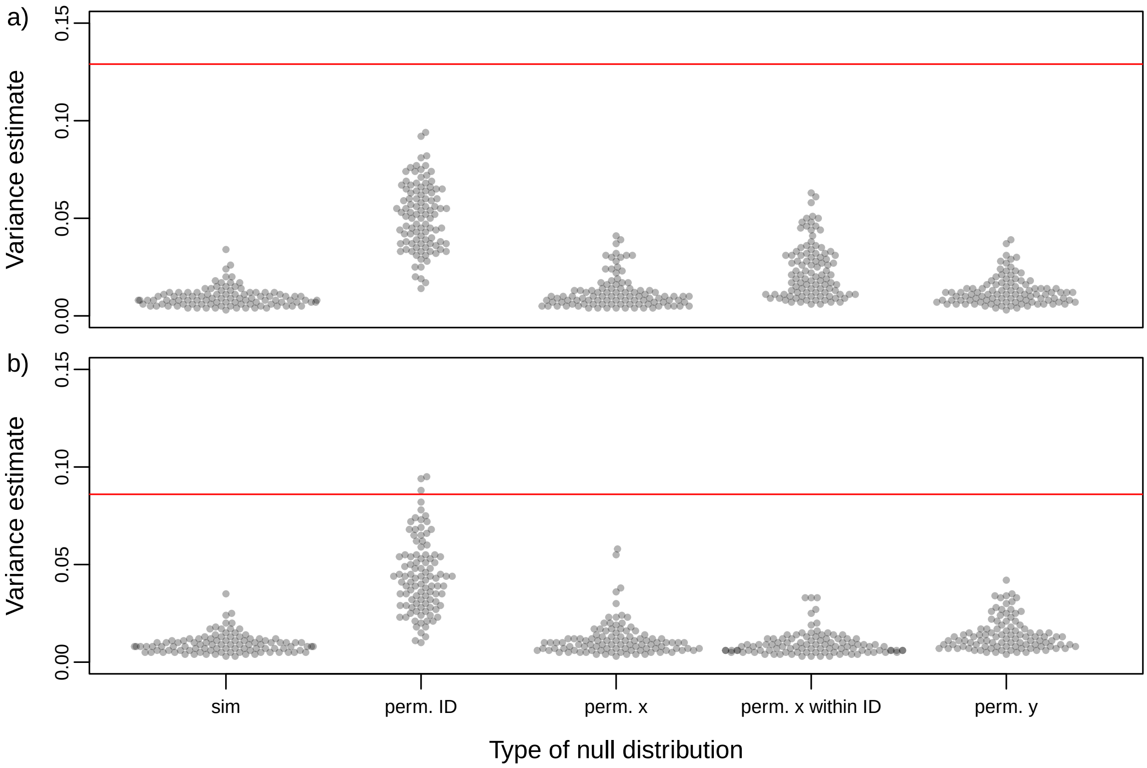

Figure 7: Null distributions of posterior medians generated with five different methods (see main text), from a) a simulated dataset, and b) a real dataset on aggressiveness in great tits. Red line represents posterior median estimated from original dataset.

Supplementary Materials

Supplementary Methods

Simulations based on Fay *et al.* (2022)

We simulated datasets based on Fay *et al.* (2022), but ran simplified models (univariate instead of bivariate), as the purpose was simply to demonstrate the effect of different measures of central tendency on the bias in these models. We simulated data with the same parameters of one set of simulation in Fay *et al.* (2022) - fast life history and low heterogeneity. We simulated the probability of survival as 0.5 and probability of reproduction as 0.7, standard deviations on the latent scale of 0.2 for both survival and reproduction and a correlation of 0.6 between the two. We simulated 100 datasets from sample sizes of 250, 500, 1000, 2000, 4000 individuals. For each simulated dataset we ran a binomial GLMM, with random effects of individual identity using Stan with the rstan package (version 2.21.3 Stan Development Team, 2022a). We specified weakly informative priors on the among-group standard deviations (half-Cauchy distribution with scale 2), and ran one chain for each model with 7500 iterations and a warm-up period of 2000 iterations. We then estimated the posterior mean, median and 2 modes as in the main text.

897 # Supplementary Figures

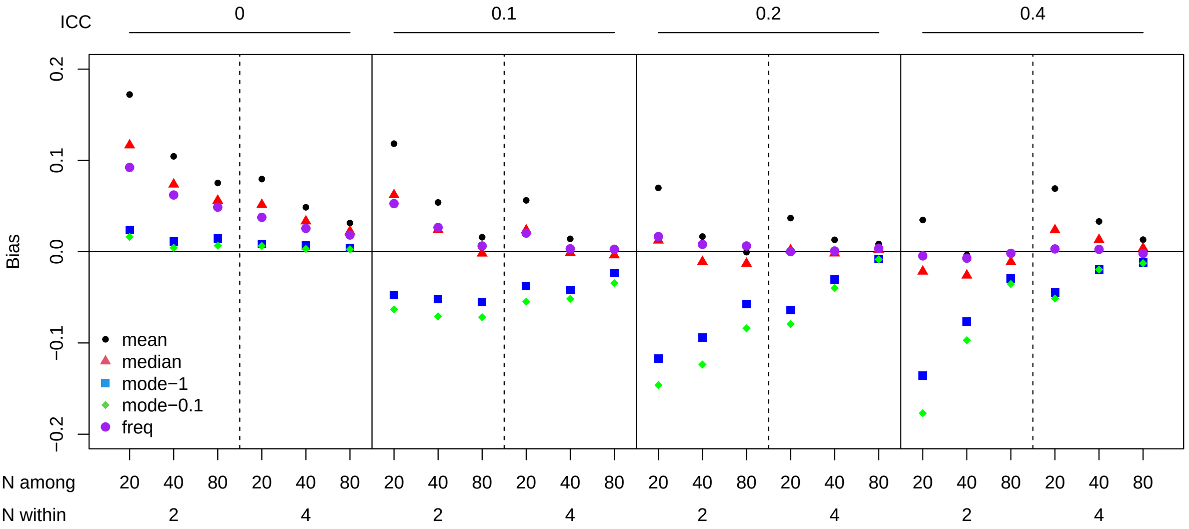

Figure S1: Bias of Frequentist estimates alongside posterior mean, median and mode of variance components, from simulations varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Two posterior modes were estimated; mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details).

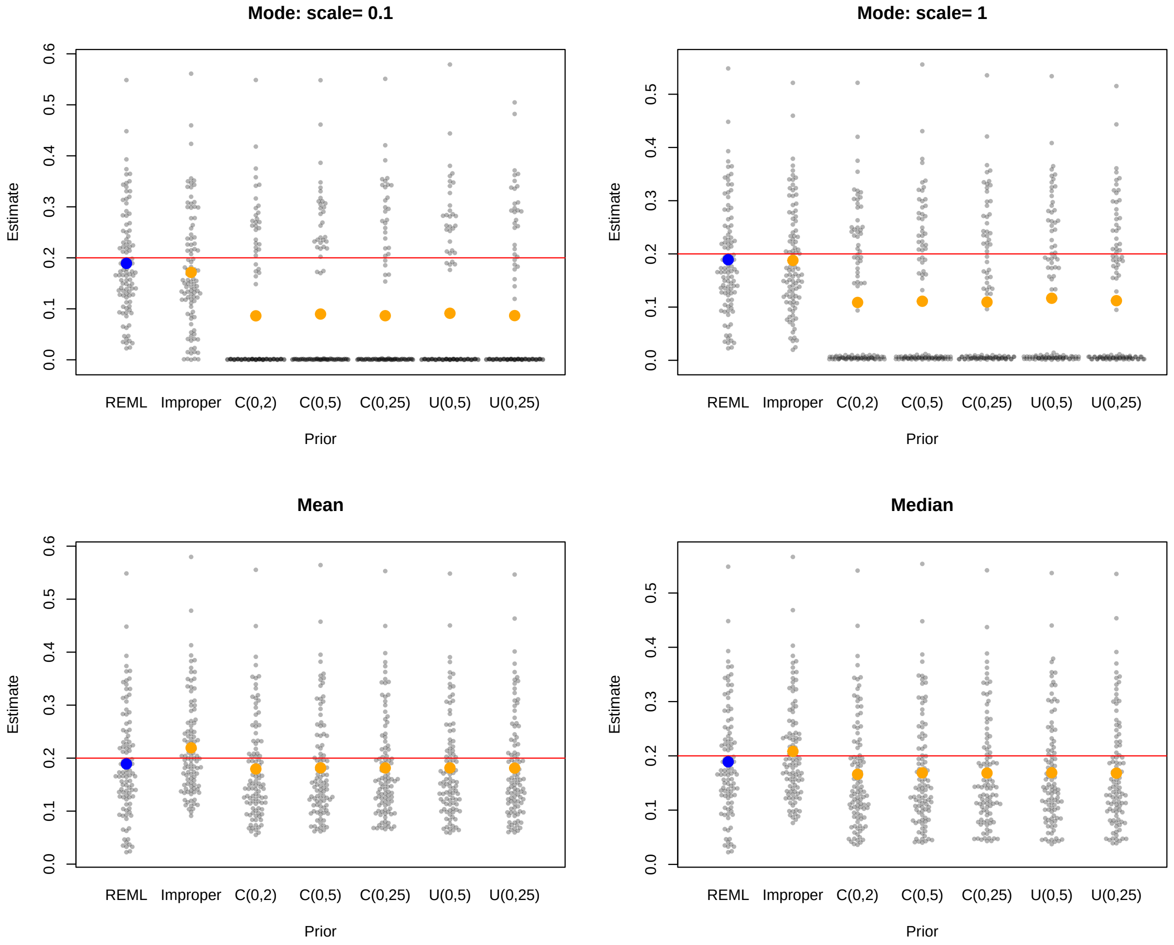

Figure S2: Impact of prior choice on measures of central tendency. 'C' represents half Cauchy priors, 'U' uniform priors, and 'Improper' uninformative improper prior. Red lines shows simulated values, and orange points shows means of different point estimates from across simulations, and blue points show the mean of the REML estimates across simulations.

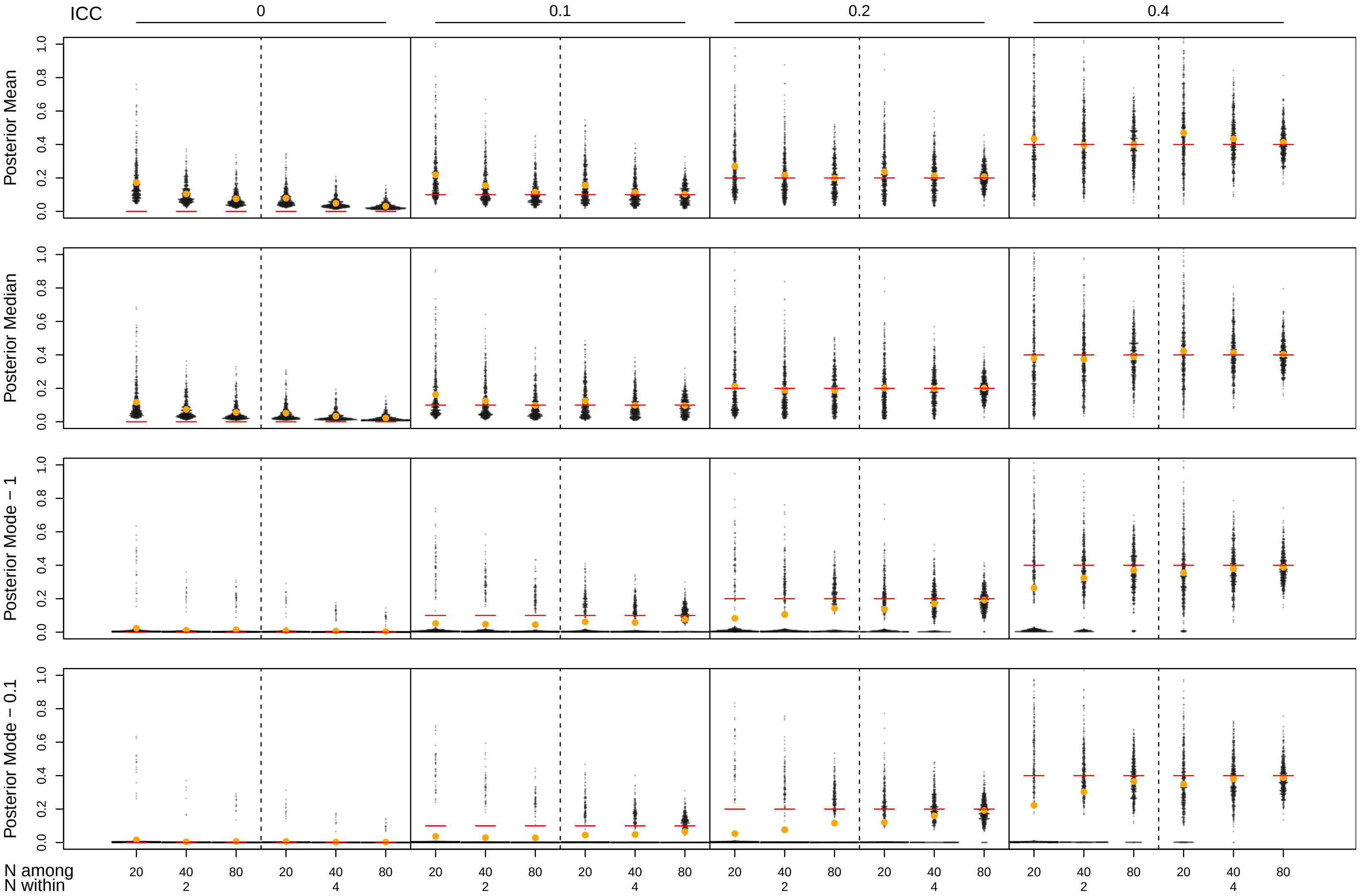

Figure S3: Sampling distributions of posterior mean, median and mode from simulations varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Red lines show the simulated value and orange points the mean of the sampling distributions.

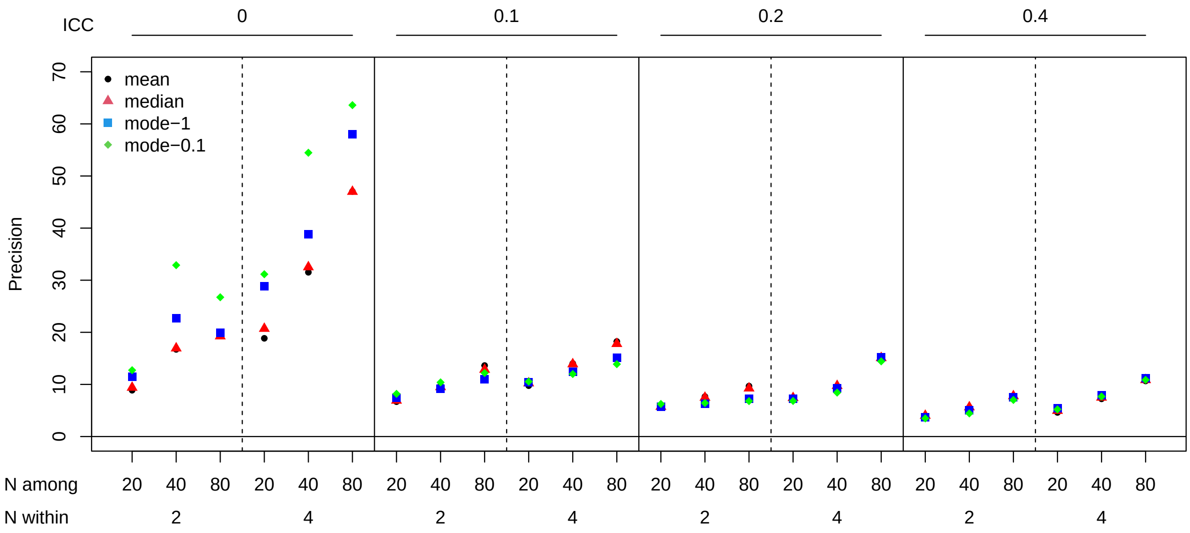

Figure S4: Precision increases with sample size, but decreases with effect size. The different panels show the precision of posterior mean, median and mode of variance components from simulations varying in among-group variance (ICC - 0, 0.1, 0.2, and 0.4) and sample size within (2 or 4) and among (20, 40, 80) groups. Two posterior modes were estimated; mode-1 and mode-0.1 with more and less smoothing, respectively (see text for more details).

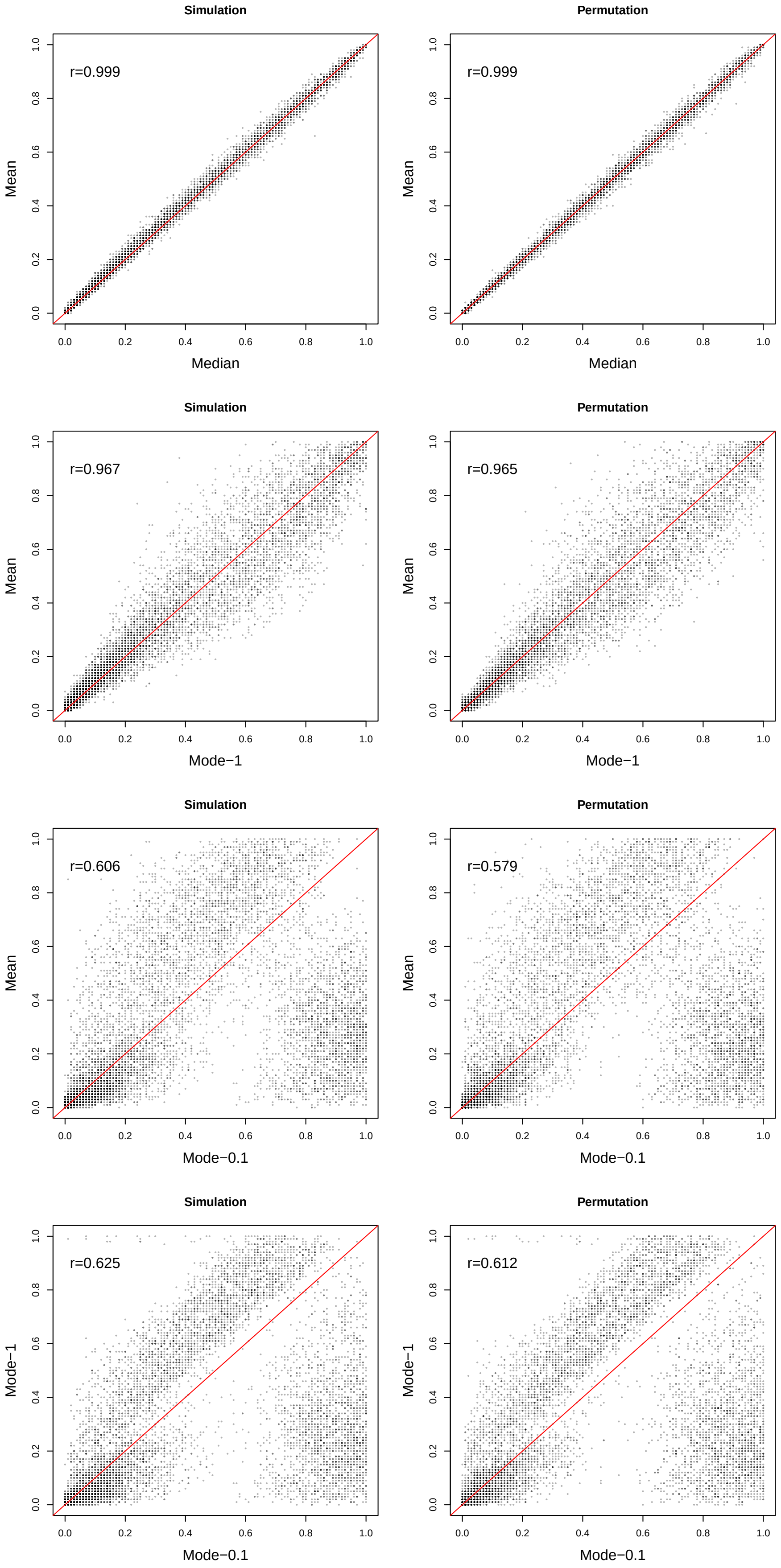

Figure S5: Comparison of p-values generated with different measures of central tendency using both simulations and permutations.

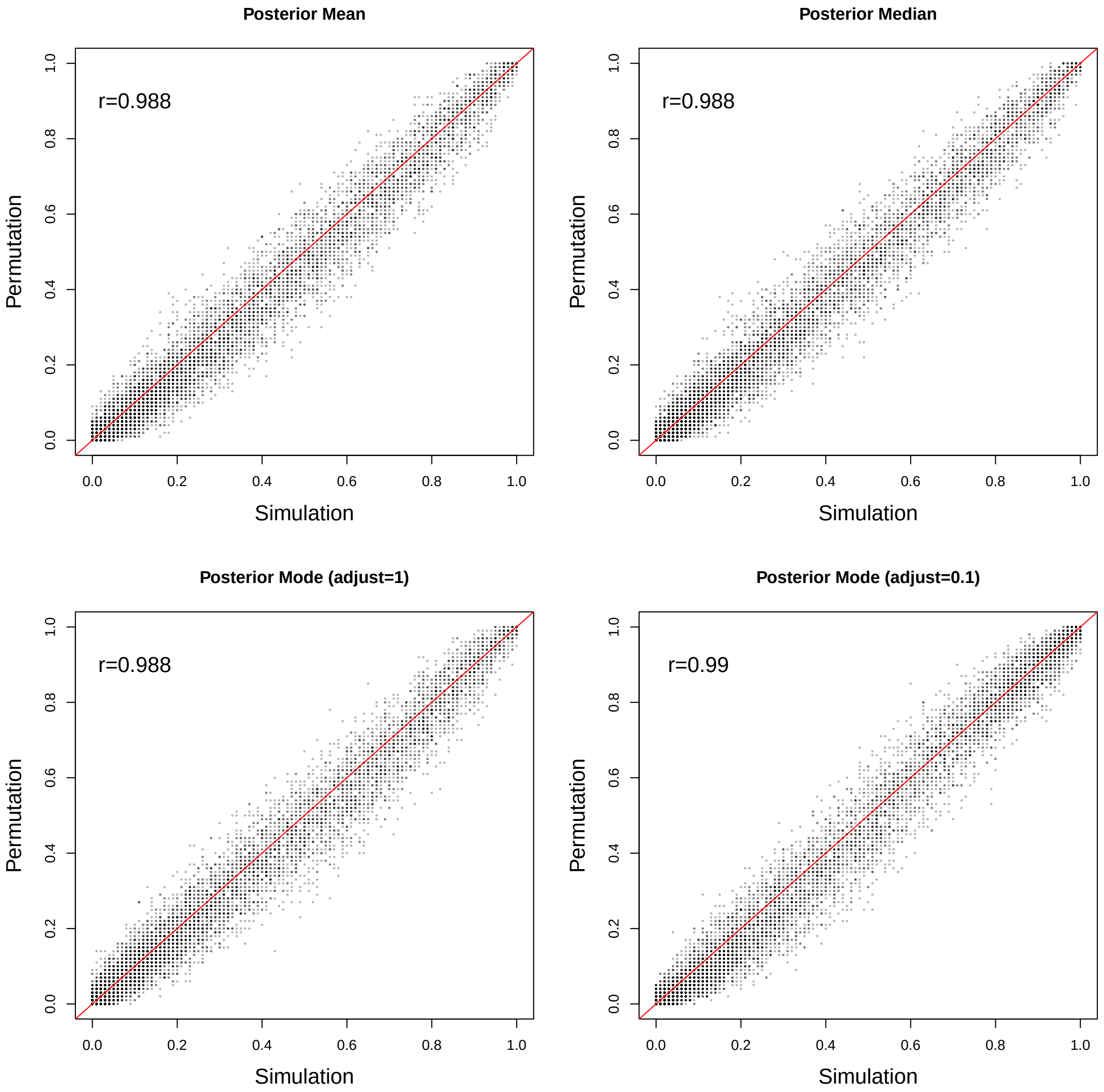

Figure S6: Comparison of p-values generated using permutation and simulation methods across all measures of central tendency.

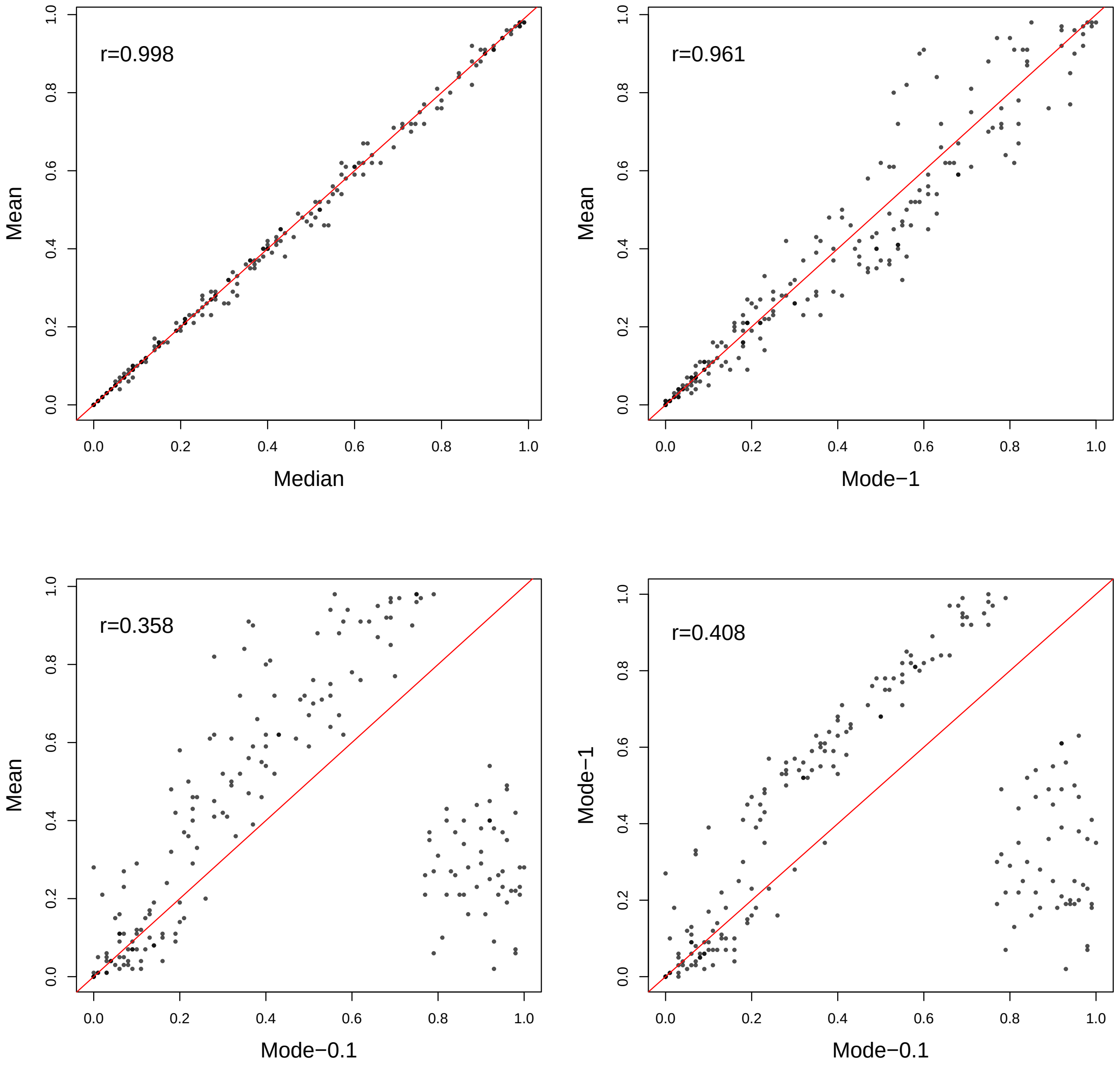

Figure S7: Comparison of p-values generated with different measures of central tendency from GLMMs using null distributions generated by simulation.

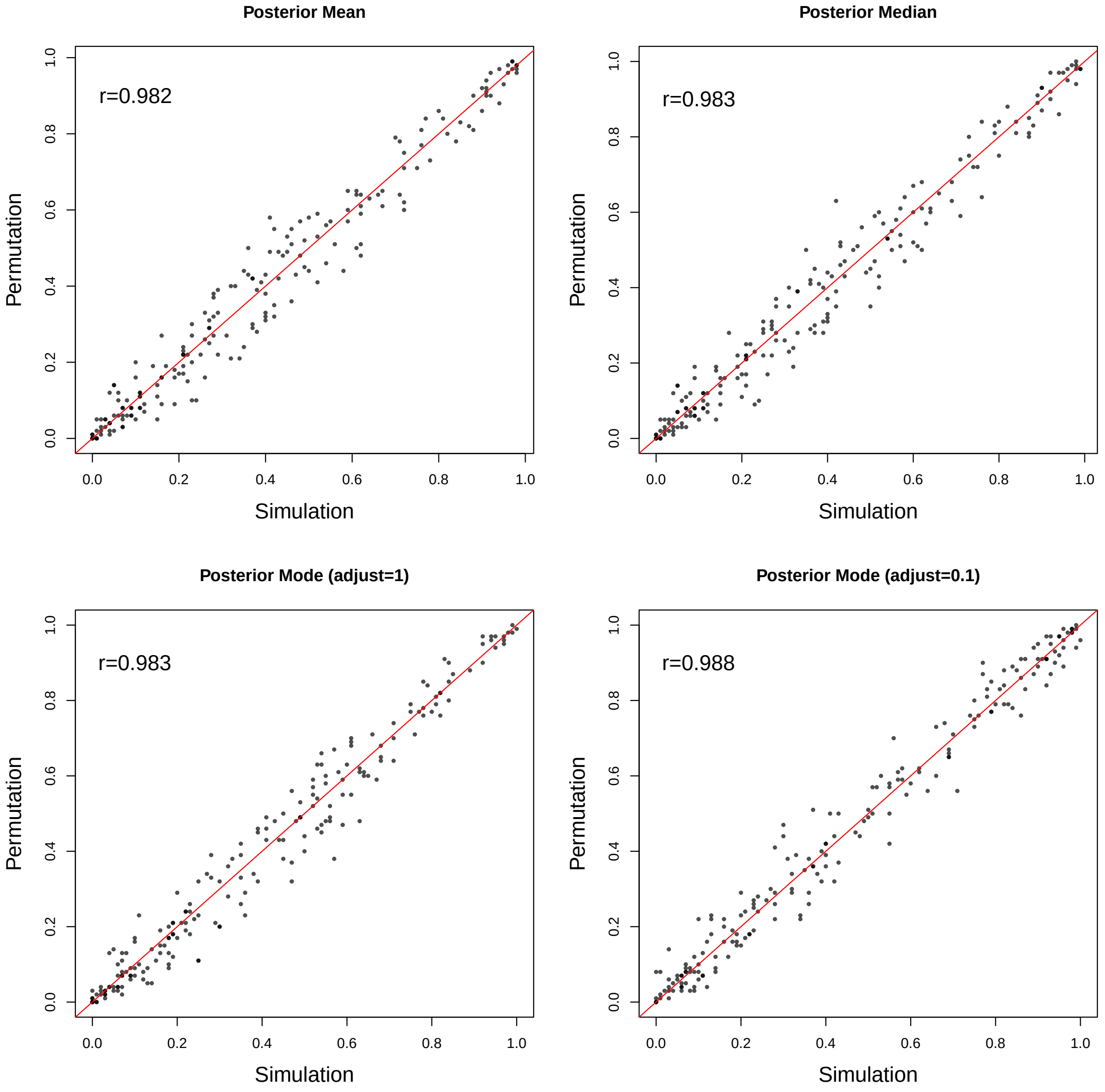

Figure S8: Comparison of p-values from GLMMs generated using permutation and simulation methods across all measures of central tendency.

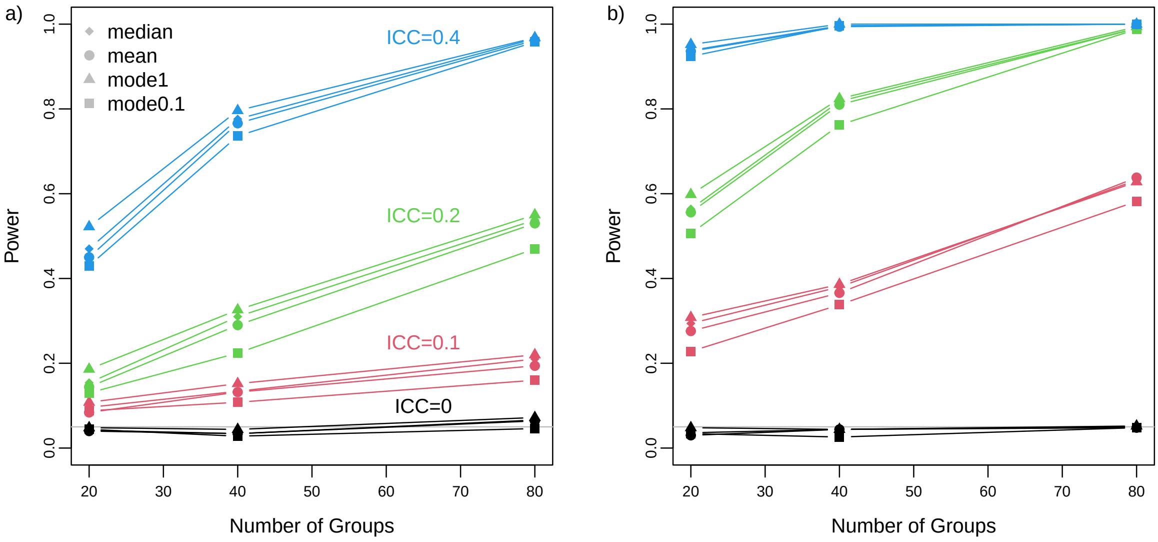

Figure S9: Comparison of power among different measures of central tendency

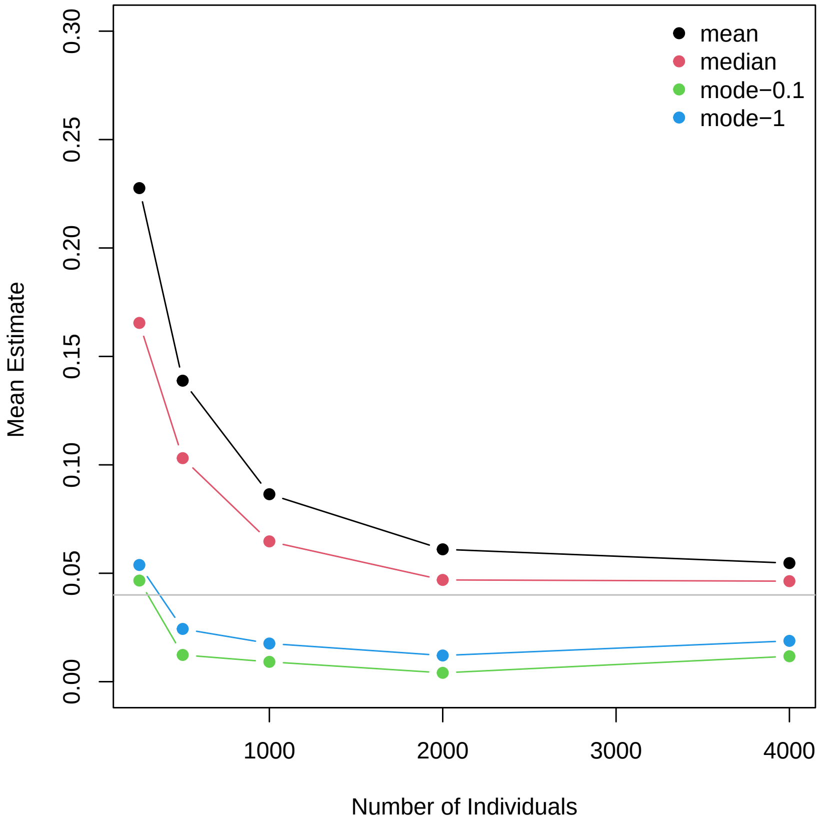

Figure S10: Mean posterior mean, median and mode of variance components from simulations based upon Fay et al. *(2022).*

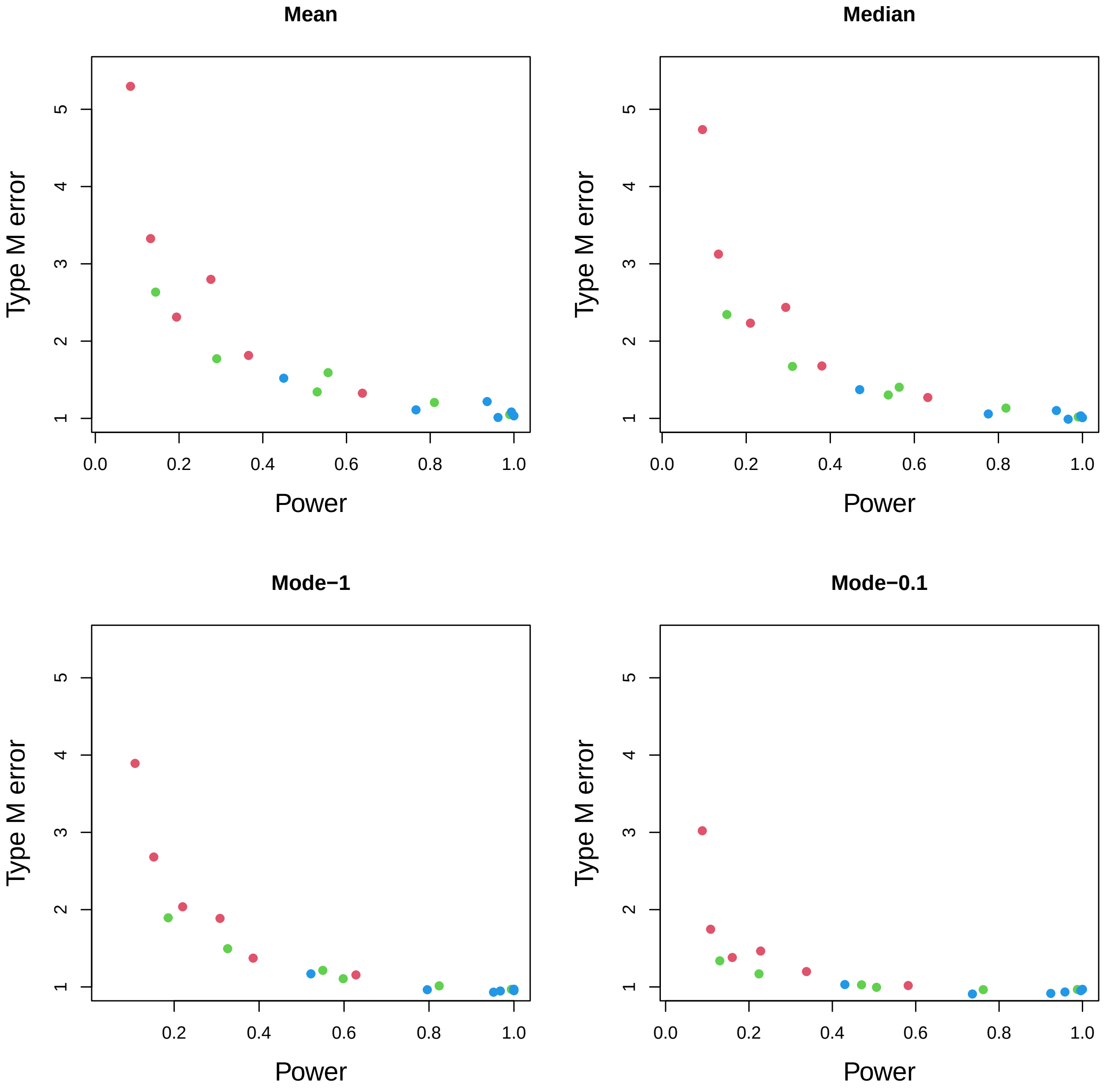

Figure S11: Type M error and power from posterior mean, median and mode calculated using null distribution generated through simulation. Colours represent simulated ICCs, red - 0.1, green - 0.2, and blue - 0.4.