Estimating Bayes Factors for Linear Models with Random Slopes on Continuous Predictors

Mirko Thalmann[ab]*, Marcel Niklaus[a], & Klaus Oberauer[a]

[a] University of Zurich & [b] University of New South Wales

Author Note

This work was partly supported by a Swiss National Science Foundation Doc.Mobility grant to the first author (#168257). All authors confirm that there are no conflicts of interest in relation to the current manuscript.

* Corresponding author:
Mirko Thalmann
Binzmuehlestrasse 14/22
CH – 8050 Zürich
mirkothalmann@hotmail.com
+41 44 63 57461

Abstract

Using mixed-effects models and Bayesian statistics has been advocated by statisticians in recent years. Mixed-effects models allow researchers to adequately account for the structure in the data. Bayesian statistics – in contrast to frequentist statistics – can state the evidence in favor of or against an effect of interest. For frequentist statistical methods, it is known that mixed models can lead to serious over-estimation of evidence in favor of an effect (i.e., inflated Type-I error rate) when models fail to include individual differences in the effect sizes of predictors ("random slopes") that are actually present in the data. Here, we show through simulation that the same problem exists for Bayesian mixed models. Yet, at present there is no easy-to-use application that allows for the estimation of Bayes Factors for mixed models with random slopes on continuous predictors. Here, we close this gap by introducing a new R package called BayesRS. We tested its functionality in four simulation studies. They show that BayesRS offers a reliable and valid tool to compute Bayes Factors. BayesRS also allows users to account for correlations between random effects. In a fifth simulation study we show, however, that doing so leads to slight underestimation of the evidence in favor of an actually present effect. We only recommend modeling correlations between random effects when they are of primary interest and when sample size is large enough. BayesRS is available under https://CRAN.R-project.org/package=BayesRS.

Introduction

Researcher X is planning a study to test whether a new drug has a beneficial effect on reading scores. To that end, X applies the drug on several days with different dosages to 40 adult participants. She also wants to be able to assess the strength of evidence for the null hypothesis because it is possible that the drug does not have an effect at all. Although X's research question seems straightforward, answering it adequately has been made possible only recently with the development of tools to analyze data in the Bayesian framework (e.g., Carpenter et al., 2016; Lunn, Thomas, Best, & Spiegelhalter, 2000; Plummer, 2003) and with the development of tools to analyze nested data (e.g., Pinheiro, Bates, DebRoy, & Sarkar, 2014). In order to adequately test her hypothesis, X has to consider two points in her statistical analysis.

First, because X wants to generalize her results to the general population of adults she has to use a statistical model that accounts for systematic variabilities between the units of observation in the sample. People will likely differ in their average reading ability. Hence, X has to add a random intercept (i.e., individual differences in the intercept) to the model. People may also differ in their sensitivity to the dosage manipulation of the drug. Hence, X has to add a random slope (i.e., individual differences in the size of a predictor's effect) on the dosage predictor. Second, she wants to use Bayesian statistics because then she can obtain evidence for the Null and test whether the Null is more likely than the Alternative.

Unfortunately, there is no easy-to-use statistical tool available at the moment that can tell X how strong the evidence in favor of or against a relationship between dosage of the drug and reading scores is. Currently available statistics packages do not allow us to compute the evidence for continuous predictors (such as dosage) with associated random slopes within a Bayesian framework. This is problematic because it has been shown that not accounting for true random slopes leads to massive over-reporting of effects when they are actually not

present (i.e., Type 1 errors; Barr, Levy, Scheepers, & Tily, 2013) in the framework of null hypothesis significance testing (NHST). Here, we show that the same problem exists also within the Bayesian framework. To help researchers address this problem, we introduce a new package in the R statistical environment (R Core Team, 2017) that is able to compute Bayes Factors (BFs) for continuous variables with associated random slopes.

In the following, we will introduce the basic principles of linear mixed-effects models and of the Bayesian statistical framework. This will clarify why the combination of these two represents a great asset for psychological research. After that, we will introduce the new R package, BayesRS, and test its functionality in five simulation studies. To foreshadow, BayesRS provides a viable new tool for researchers to analyze their data.

Bayesian Statistics

At the core of the Bayesian framework is the idea to express the believability of a proposition as a probability distribution (Kruschke, 2014). That is, probability is not defined as the expected frequency of an event, resulting from an imaginary infinite number of samples from the reference set (Neyman, 1977), but rather reflects a subjective belief that a proposition is true. Therefore, we can assign probabilities not only to events but also to hypotheses and theories. After having observed some data speaking to a proposition, we update our subjective belief in it based on these data according to Bayes' rule. Statistical inference in the Bayesian framework proceeds through three steps: First, specification of a prior probability distribution (hereafter prior) of a hypothesis; second, calculating the likelihood of the hypothesis given the data; and third, computing a posterior probability distribution (hereafter posterior) by updating the prior with the likelihood. Using the posterior for inference about the value of a parameter in a statistical model uses all of the available information (i.e., relevant prior knowledge and the data). For unimodal, symmetric posteriors it is convenient to summarize them with their mean and an interval, for example the interval

covering the most credible 95 % of its density (95 % highest density interval, HDI, e.g. Kruschke, 2014).

A major advantage of Bayesian statistics over classical NHST is that it allows for the comparison of the relative plausibility of competing hypotheses (H) given some data (D). This comparison is tightly linked to Bayes' Theorem (Bayes, Price, & Canton, 1763)

$$P(H|D) \propto P(D|H) * P(H)$$

It states that the posterior P(H|D) is proportional to the likelihood P(D|H) multiplied by the prior P(H). Bayes Theorem can also be expressed in terms of odds, that is, ratios of probabilities pertaining to the two competing hypotheses:

$$\frac{P(D|H_1)}{P(D|H_2)}.$$

This equation shows that we should update the prior odds by the likelihood ratio to obtain the posterior odds. The likelihood ratio is also known as the Bayes Factor (e.g., Jeffreys, 1935; Kass & Raftery, 1995). The inverse of the BF reflects the BF in favor of the competing hypothesis. Kass and Raftery (1995) suggest some loose guidelines how BFs might be interpreted. A BF in the range of 1-3.2 is considered to be "not worth more than a bare mention", a BF ranging from 3.2-10 as "substantial", a BF from 10-100 as "strong", and a BF > 100 as "decisive" evidence in favor of a hypothesis.

The Bayesian framework offers solutions to problems that cannot be addressed easily within the framework of NHST (Wagenmakers, 2007; Hubbard, 2004) and thus represents a powerful alternative to conventional inference based on *p* values.

Linear Mixed-Effects Models

In a seminal paper, Clark (1973) pointed out a problem he termed "The Language-as-Fixed-Effect Fallacy". He argued that experimental researchers often have a small pool of

items in their experiments but nevertheless would like to generalize their results to a class of items in general. By neglecting this fact in their statistical analyses they are likely to commit Type 1 errors. That is, they report an effect to be true in the broader class of items, even though there is actually no effect in the population of items but only in the specific sample of items. Hence, accounting for the fact that items are a random effect, and not a fixed effect, is of great importance.

To say that an effect is fixed means that the true value of the effect is the same across all observation units of a set (e.g., a set of subjects, items, studies) and differences between values across observations units are only due to sampling noise (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2010). In contrast, to say that an effect is random means that the true values of the effect differ across observation units. Specifically, it is assumed that the true values across units reflect a random draw from a larger population that is described by a distribution of values, often a normal distribution. When the dependent variable is modeled as a linear combination of independent variables and measurement noise, and includes fixed as well as random effects, the resulting class of models is called linear mixed-effects models. Mixed-effects models have only become popular in psychological research in the last few years, because statistical software that allows their computation has become available only recently (e.g., Pinheiro et al., 2014).

Although mixed-effects models represent a powerful tool, the increased complexity in model structure is accompanied by increased degrees of freedom in the model building process. For instance, the question how one should set up the random-effects structure in a given set of data is still a matter of debate. One view is that analysts should use a fit criterion and apply the random-effects structure that best fits the specific set of data (Baayen, Davidson, & Bates, 2008). Another view is that the maximal random-effects structure should be used that is justified by the design of the experiment (Barr et al., 2013). That is, even

though there may be no substantial differences in an experimental effect across the different observation units, researchers should nevertheless include a random slope for this effect in their model. Barr and colleagues (2013) showed via simulation that a model including random slopes generalizes better to an underlying population in terms of Type 1 and Type 2 errors.

Random effects for continuous predictors

Although the debate regarding the appropriate random-effects structure continues (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), the point put forward by Clark (1973) – that models that do not include a random effect that is actually present are prone to increased Type 1 errors – is still valid. Including random slopes in a model can therefore be an important step when analyzing experimental data. The BayesFactor package (Morey & Rouder, 2014), which computes BFs for mixed models, allows users to specify random intercepts and random slopes for categorical predictors (i.e., variables on a nominal scale), as are often used in experimental research. However, at present, there is no statistical software that allows users to readily compute BFs for models that include random slopes on continuous predictors, that is, predictors with multiple levels on an ordinal or even an interval scale of measurement. This is problematic as experimental researchers often include variables in their experiments that are continuous – as in the example of researcher X who was interested in modeling the effect of the dosage of a drug on reading ability. Here, we present a new R package, BayesRS, that closes this gap. We test the reliability and validity of the BFs computed with the package (Simulations 1 to 3), and investigate the consequences of including or not including random slopes, and their correlations, in mixed-effects models (Simulations 4 and 5).

Method

The BayesRS package allows the specification of linear mixed-effects models in the Bayesian framework with maximally two levels in the hierarchy. Here, we describe the model structure for the full hierarchical model, in which the coefficients of all predictors are estimated for all observation units of a set. The dependent variable and the continuous predictors are entered as z-standardized values into the model. Categorical predictors with *n* levels are entered as *n*-1 simple coded contrasts. Therefore, fixed effects or group distributions of the mean effects reflect standardized regression coefficients. An important feature is the use of default priors that we put on mean effects size, as will be seen in the following. Every single z-standardized data point yz_i is modeled according to

$$yz_i \sim \mathcal{N}(M_i, T) \quad (1)$$

where $\mathcal{N}$ is short for the probability density function of the normal distribution. The linear regression structure is on

$$M_i = \sum_{p=0}^{P} \sum_{j=1}^{J_p} x_{ij}^{(p)} \beta_j^{(p)}, \quad (2)$$

and precision (the inverse of the variance) is defined as

$$T \sim \Gamma(a,b). \quad (3)$$

We follow Gelman, Carlin, Stern, and Rubin (2014) regarding the notation in batches. The first sum in Equation (2) runs over batches; each batch stands for one predictor. Every p^{th} batch represents a set of J_p random regression coefficients belonging to a specific predictor (e.g., a categorical predictor, a continuous predictor, or an interaction). Because the model includes random slopes, each observation unit has its own beta coefficient for each predictor; hence each batch represents *J* beta coefficients. Across all batches, the $\beta_j^{(p)}$ coefficients are entries of a P x J matrix, where P and J reflect the number of predictors and the number of

observation units in all sets, respectively. The intercepts are represented in row number "0" of the matrix. That is, each observation unit has its own intercept reflecting the mean of that specific observation unit across all levels of the predictors. We assume the individual by-subjects beta coefficients to be normally distributed according to a group distribution with mean μ_p and condition precision $\frac{1}{\sigma^2 p}$

$$\beta_j^{(p)} \sim \mathcal{N}\ (\mu_p, \frac{1}{\sigma^2 p}) \quad (4)$$

and further

$$\mu_p \sim T(0, df = 1) \times \text{Scale}_j, \quad (5)$$

$$\sigma_p \sim \Gamma(1, 0.04). \quad (6)$$

Here, T stands for the Student's *t*-distribution that is scaled according to a scaling factor Scale_j. We follow previous work by placing a Cauchy distribution (i.e., a Student's *t*-distribution with 1 degree of freedom) as a default prior on μ_p (Rouder, Morey, Speckman, & Province, 2012). The logic of placing a zero-centered Cauchy distribution on the standardized effects is to place somewhat more prior probability to larger effects than, for example, a normal distribution would. We use scaling factors of $\sqrt{2/4}$ and ½ for continuous and categorical predictors, respectively. They reflect default priors that are called "medium" in other software (Rouder et al., 2012), in contrast to a "wide" prior that puts even more prior probability on larger effects.

The package also allows users to estimate only one single β coefficient for a predictor *p*, in which case the p^{th} row of the P x J matrix drops out and only μ_p is estimated for predictor *p*. In the following, we will refer to these mean parameters as fixed effects to facilitate comparison with the classical mixed-effects parlance.

Overview of the package

The BayesRS package is built in the R statistical environment (R Core Team, 2017). We use JAGS (Plummer, 2003) to sample from the posteriors of model parameters. JAGS itself is accessed from R via the rjags package (Plummer, 2013). For processing and plotting of the data we make use of several other R packages that are mentioned in Appendix A.

The package takes as input dependent variables that are normally distributed. The dependent variable and all the continuous independent variables are automatically z-standardized. This means that the modeling is done in standardized space, and therefore all regression coefficients of continuous variables reflect standardized β coefficients. Hence, a β posterior has to be retransformed into the original scales when the effect size in the original scale (i.e., the b parameter) is of interest. Categorical predictors are entered into the model via simple coding.

The user accesses the package with its interface function called `modelrun`. The `modelrun` function does all the processing before and after the model is run in JAGS. It transforms the input data into the required format and calls the translation function `modeltext` that takes a description of the data structure as an input and writes it as a text file in JAGS language. After that, the `modelrun` function hands over the data and the written text file to JAGS, which then samples from the posteriors. After that, the `modelrun` function computes the 95% HDIs and the BFs of the parameters of interest. These are finally plotted in a figure, and the BF is returned as a vector. The `modelrun` function requires three obligatory arguments and has nine additional arguments that can be changed, but are otherwise set to default values.

The first three obligatory arguments reflect the data – which have to be handed over as a data frame in the long format –, the dependent variable with its name written as a string,

and the data structure. The latter is a separate data frame that lists each predictor in a row, and defines for each whether it is continuous or categorical, and for what observation unit a random slope is required or not. That is, the package allows also for several, crossed-random effects. An additional optional argument allows users to put random slopes on interactions of predictors. Three further optional arguments allow for the variation of the number of MCMC steps for adaptation and burn-in, and the number of MCMC steps that are actually saved. Further, it is possible to obtain convergence statistics of the main parameters in the model. Convergence of the model is a prerequisite for the BFs to be interpretable.

Another argument allows the user to specify which random effects within an observation unit should be modeled as correlated. The package can model correlations between pairs of variables, and correlations between more than two variables. When both is done at once, the variables within the two correlation structures cannot overlap (e.g., one can specify a pairwise correlation between predictors A and B, and additionally, the full correlation matrix between predictors C, D, and E).

One optional argument allows users to save the chains of the random intercepts and random slopes. Two final options are to determine (a) whether a figure of the 95% HDIs and the BFs of the main parameters should be plotted and (b) whether the deviance information criterion of a model (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002) should be returned.

Computation of Bayes Factors

A main feature of BayesRS is its ability to compute BFs for models including continuous predictors. We compute BFs with the Savage-Dickey density ratio (e.g., Dickey & Lientz, 1970). The Savage-Dickey density ratio is an approximation of the BF for nested model comparison. Following Wagenmakers, Lodewyckx, Kuriyal, and Grasman, (2010) we

assume the null hypothesis to be a point null hypothesis, stating that the effect of a predictor in question is exactly zero. The alternative hypothesis states that the effect is non-zero. We can obtain the BF in favor of the alternative hypothesis by comparing a model that estimates the effect in question as a free parameter to an otherwise identical model in which it is fixed at zero. The BF is the ratio of the density of the free parameter's prior to the density of its posterior in the alternative model, evaluated at the parameter value to which it is set in the null model (i.e., at zero).

Let us assume that researcher X has collected all data from her study and tests whether application of the drug increases reading test scores. Her null hypothesis is that the drug has no effect on reading scores. The alternative hypothesis allows effect size to vary freely in the model. In the alternative model, she has the default Cauchy prior on standardized effect size. After running the model, she observes a posterior with a mean of 0.5 and a standard deviation of 0.2. She obtains the BF for the alternative hypothesis by dividing the height of the prior by the height of the posterior at zero. An illustration is shown in Figure 1.

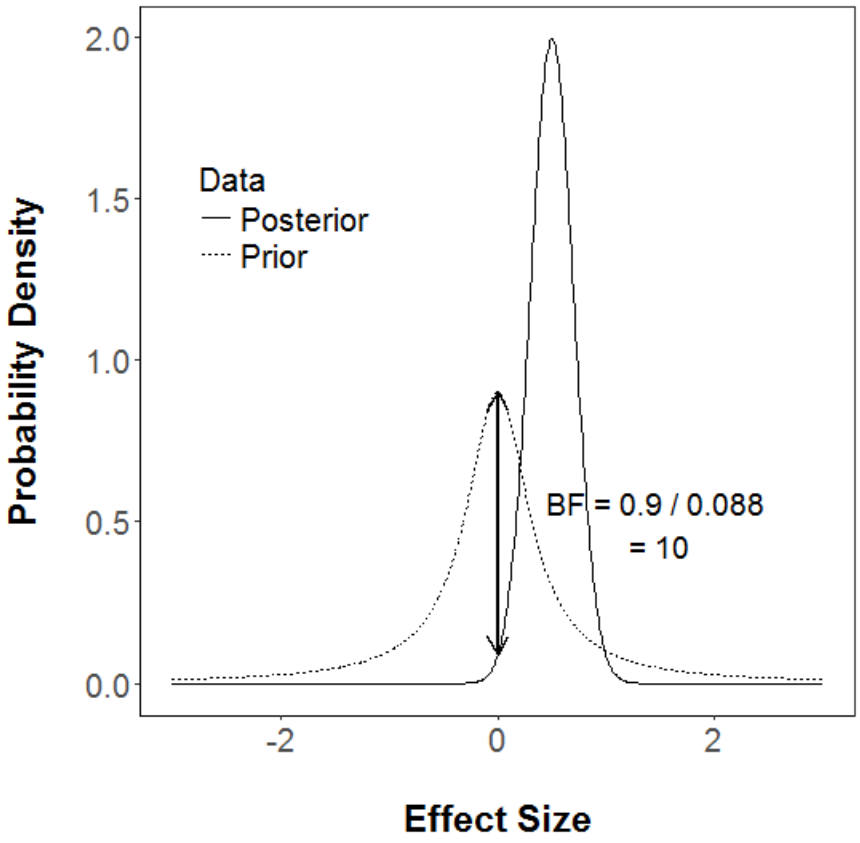

Figure 1. Demonstration of how the BF is calculated with the Savage-Dickey density ratio test. The null hypothesis is that effect size = 0. Therefore, the BF is the ratio of the density of the prior and the density of the posterior at zero.

Computationally, we use the approach by Wetzels, Raaijmakers, Jakab, and Wagenmakers, (2009) to estimate the density of the posterior at zero. That is, we compute the mean and the standard deviation from the MCMC chain of an effect of interest. Then, we estimate the density of the posterior at zero by computing the density of a normal distribution at zero with the computed mean and standard deviation. This approach takes advantage of the Bayesian central limit theorem that states that under general regularity conditions posterior distributions tend to be normally distributed when the number of observations becomes large (Wetzels et al., 2009).

Results and Discussion

In the following, we present five simulation studies that tested the functionality of the BayesRS package. We asked (1) whether the instantiation of the package leads to a reliable estimation of a BF given one set of data, (2) whether the computed BF on a fixed-effect model is comparable to an analytic solution provided by Gaussian quadrature, (3) whether the BF for the fixed effect of a continuous predictor approaches the BF of a Bayesian *t*-test on the true individual *b* values in the sample, (4) whether not accounting for random slopes when they exist in the data leads to increased potential Type 1 errors, and finally (5) whether not accounting for correlations between random slopes when they exist in the data leads to increased potential Type 1 errors.

Simulation Study 1

The first study aimed to test the reliability of the instantiation of the Savage-Dickey density ratio in the BayesRS package. Data were simulated from an experiment with one continuous independent variable with five values, and one random variable (e.g., subject). We independently varied the mean effect size of the continuous variable (0, 0.2, 0.5, 0.8), the number of subjects in the experiment (20, 40), and the number of saved MCMC steps (50’000, 100’000). We sampled ten observations per design cell of the simulated experiment for each subject. The individual by-subject intercepts were generated according to a normal distribution with mean 0 and standard deviation 1. The by-subject slopes of the continuous variable were generated according to a normal distribution with the before-mentioned mean effect sizes, and standard deviation 1. Because we only generated one set of data for every design cell of the design of the simulation study, we forced the simulated by-subject slopes to exactly reflect the desired properties in this first study. On every generated data set we computed the BF 50 times with the `modelrun` function.

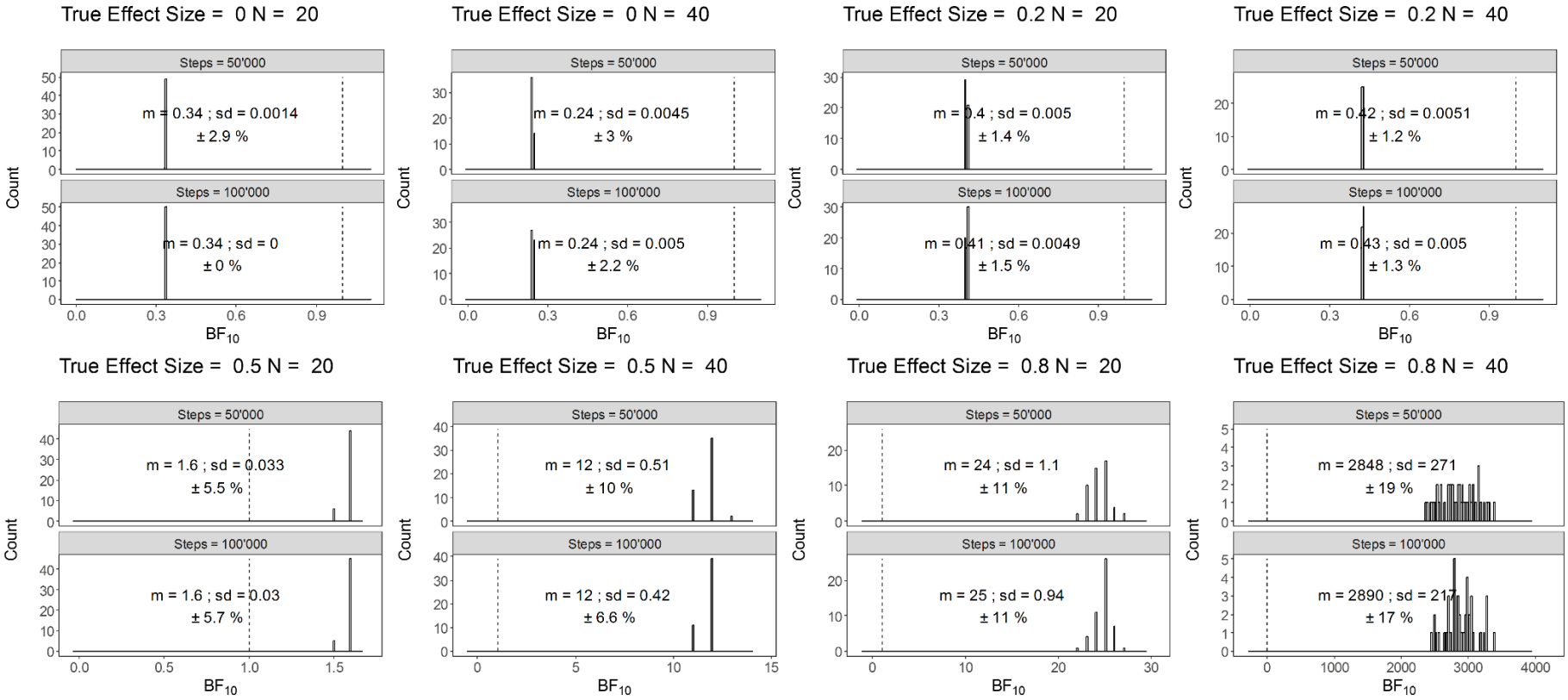

Figure 2. Distributions of BFs within the design cells of Study 1. We also show mean (m), standard deviation (sd), and the deviation of the most distant BF compared to the mean BF (deviation/mean BF*100) out of the 50 computed BFs per cell. The dashed line shows a BF of 1 representing a situation in which H_0 and H_1 are equally likely.

It can be seen in the results displayed in Figure 2 that BFs below 1 are estimated with high reliability. For example, a BF with a mean of 0.34, as in the first panel from the left in the upper row, can be estimated with perfect reliability with 100'000 MCMC samples, and a standard deviation of 0.0014 with 50'000 MCMC samples. When the BF grows larger, as with a mean of 12 in the second panel from the left of the lower row in Figure 2, it is sometimes estimated as 11 with 100'000 samples, and occasionally as 11 or (twice) as 13 with 50'000 samples. Although this represents a standard deviation of about 0.5, it is only a small, negligible loss in precision and would not change the interpretation of the results, i.e., that there is strong evidence in favor of the alternative hypothesis. When BFs grow even larger, they become more variable. In the most extreme case in the lower row in the fourth panel, the maximal and the minimal BFs were estimated to be 3392 and 2360, respectively, when 50'000 MCMC steps were saved. The variation with 100'000 samples was somewhat smaller. Although this shows substantial variation in the BF, it would not change the interpretation of the results. That is, in all cases, there is decisive evidence in favor of the alternative hypothesis.

To summarize, the estimation of densities that are further away from the mean of the posterior distribution is more variable. This is a consequence of the Savage-Dickey estimation of the BF, which relies on precise measurement of the posterior density at zero. When the mean of the posterior distribution grows larger, its density estimation at 0, being further out in the tail of the distribution, becomes less precise. Because the density is already very low, even absolutely small changes in its estimation have a relatively large influence on the estimation of the BF. Using more MCMC samples tended to attenuate the imprecision slightly. However, in none of the presented cases this variation would have led to changes in the interpretation of the data. Hence, we conclude that the ability to discriminate between the Null and the Alternative of the new R package is good enough.

Simulation Study 2

The aim of the second study was to test the validity of BayesRS, that is, whether the instantiation of how the BF is estimated yields comparable results to an existing benchmark. Specifically, we generated data with a continuous predictor without a random effect, and compared the BF when computed with the new package to the BF computed with the BayesFactor package (Morey & Rouder, 2015). The latter analytically computes the BF via Gaussian quadrature, and therefore we will call it the "true" BF in the following. We generated data from an experiment as in Study 1, but without random slopes on the continuous variable. Effect size was varied from .05 to .1 to .3, and data were randomly sampled without any constraints. The results are pooled across the different effect sizes and plotted in Figure 3 on a natural logarithm (ln) scale. Values to the right and to the left of zero on the x axis show BFs in favor of the Alternative and in favor of the Null, respectively. Values above the red line show cases in which the BF for the Alternative is larger when computed with BayesRS compared to the true BF, and values below the red line show cases in which the BF for the Null is larger when computed with BayesRS compared to the true BF.

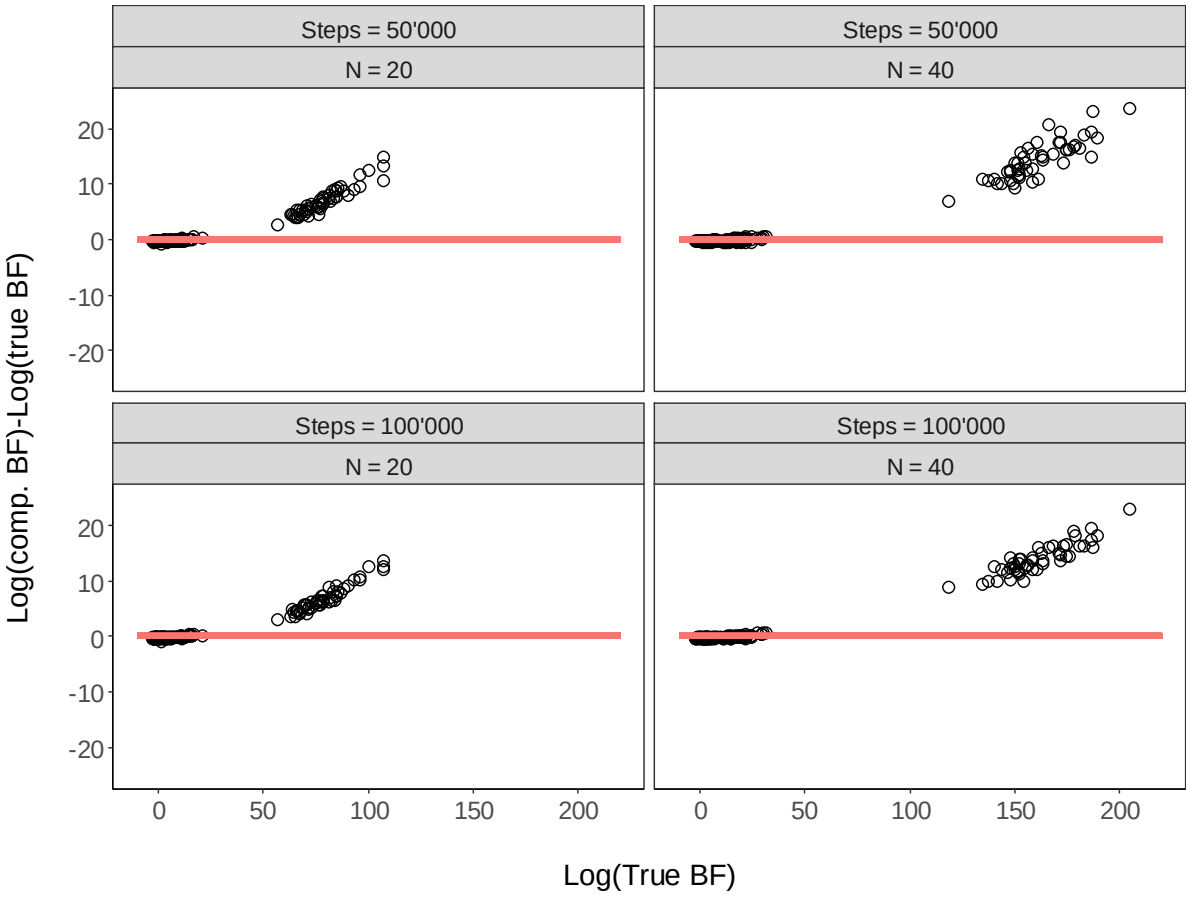

Figure 3. Deviation of the BF estimated with BayesRS from the true BF computed with the BayesFactor package, plotted against the true BF. Data are pooled across effect size and units are in Log space. A value on the red line indicates that the BF computed with BayesRS is the same as the true BF.

Figure 3 shows that the BF calculated with the Savage-Dickey density ratio represents the true BF acceptably. The approximation of the true BF is slightly better and slightly less variable with 100’000 MCMC samples than with 50’000 MCMC samples. Nevertheless, even with 100’000 MCMC samples, the estimation of the BF does not exactly match the true BF. The distortions are however in a desirable way. For true BFs in favor of the Alternative the computed BFs systematically overestimate the evidence in favor of the Alternative. The

reverse is true as well: For true BFs in favor of the Null, the computed BFs systematically overestimate the evidence in favor of the Null. When we zoom in to log(true BFs) between -2.3 and 2.3 – values that do not express strong evidence in favor of either hypothesis –, Figure 4 shows that the Savage-Dickey density ratio slightly overestimates the evidence in favor of the Null in all of these cases. In the most extreme case (see the outlier in Figure 4), our implementation underestimates a true BF of about 4.7 in favor of the Alternative as only about 1.9 in favor of the Alternative. Arguably, this underestimation is not severe. Overall, BayesRS yields comparable results to an analytical solution and therefore passes the second test.

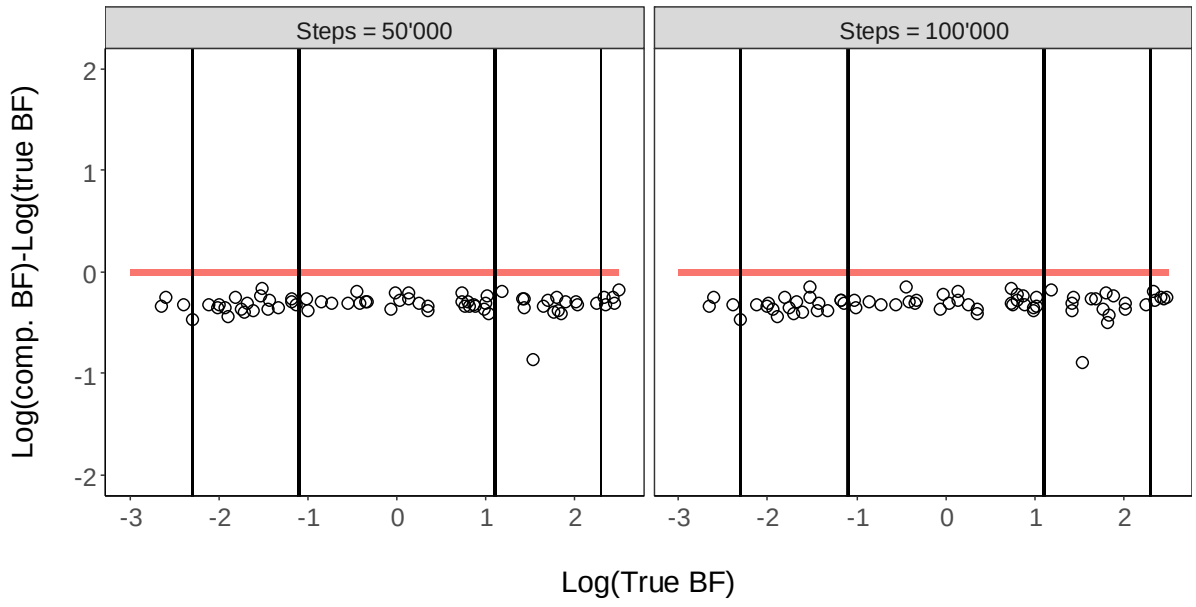

Figure 4. Deviation of the BF computed with the BayesRS package from the true BF, plotted against the true BF in Log space, zooming in on BFs that are within a range of relatively high uncertainty, where small errors in estimating the BF would matter most according to the classification of Kass and Raftery (1995). Note: The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin

vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

Simulation Study 3

The third study tested whether the BF of a fixed effect of a continuous predictor with random slopes asymptotically approaches the BF of a Bayesian *t*-test from the BayesFactor package on the known, true *b* values of individual observation units that were used to generate the data set. In the following, we refer to the latter again as the "true" BF. The data were again simulated from a similar experiment as in Study 1, except that we varied the number of subjects between 10, 25, 50, and 100, and varied the number of observations per design cell between 3, 6, and 9. With these values we aimed to get more BFs that are closer to 1, which are crucial according to the classification by Kass and Raftery (1995) to decide whether an effect is not worth mentioning ($0.3125 < BF < 3.2$) or considered to be substantial ($BF > 3.2$ or $BF < 0.3125$) or strong (BFs > 10 or $BF < 1/10$). In addition, we fixed the number of MCMC steps to 100'000 because of slightly higher reliability compared to 50'000 steps. The results are depicted in Figure 5 and show a similar pattern as in Study 2. That is, BFs between -2.3 and 2.3 do not differ substantially between the two methods. However, when the true BFs grow larger, those computed with BayesRS tend to overestimate the evidence in favor of the Alternative. The overestimation seems to be slightly more pronounced for the smaller sample size. As in Study 2, we argue that the distortions are not problematic because in neither case they change the interpretation of the results of a given data set. Hence, the BayesRS package is able to discriminate between the Null and the Alternative of a fixed effect with associated random slopes. Although there is a substantial overestimation of the BF in some cases, this applies only when the BF in favor of the Alternative computed with the *t*-test is already large.

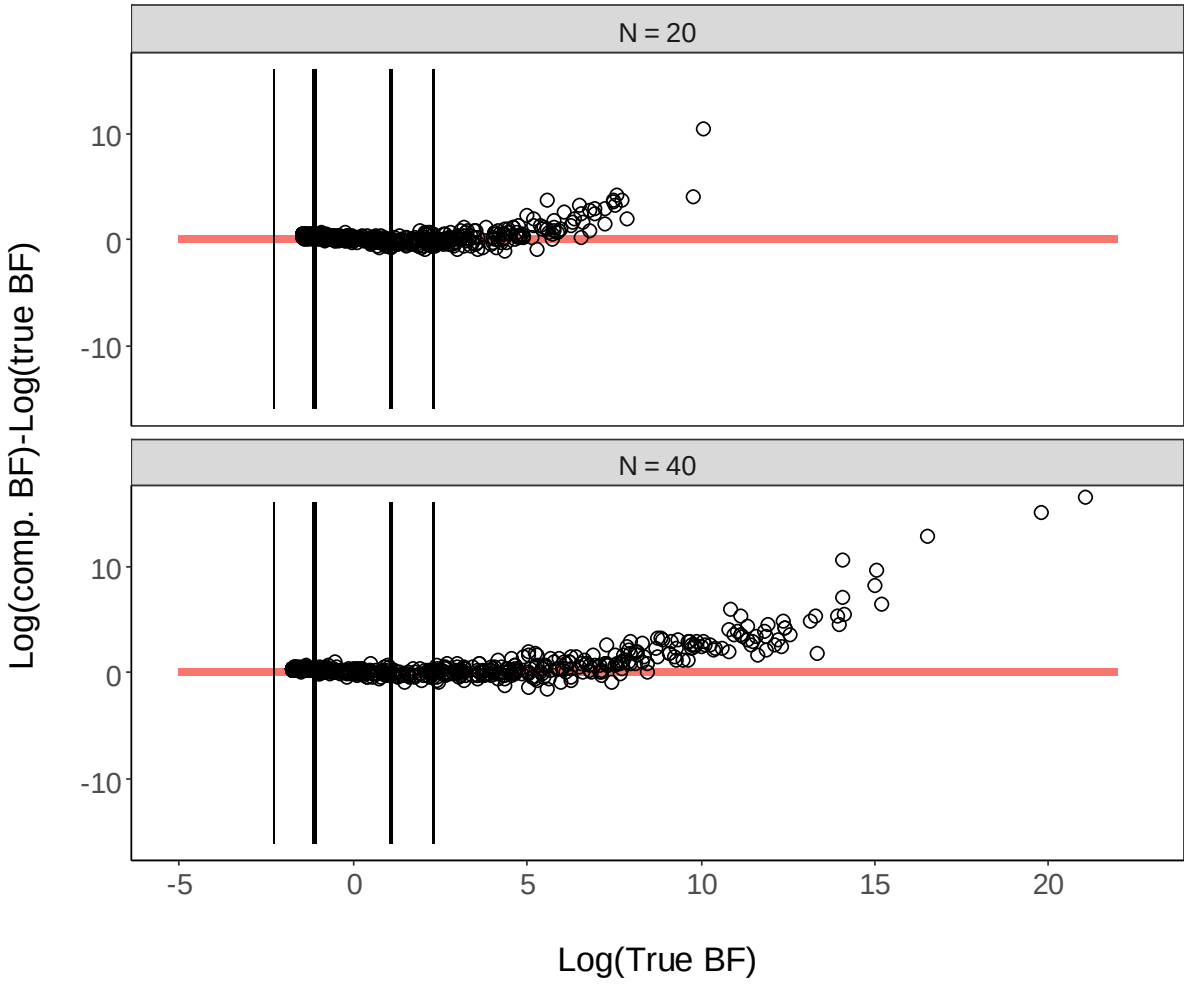

Figure 5. Deviation of the BF computed with BayesRS from the "true" BF from the Bayesian *t*-test computed with the BayesFactor package, plotted in Log space. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values). A value on the red line indicates that the BF computed with BayesRS is the same as the true BF.

Simulation Study 4

Work by Barr and colleagues (2013) has shown that not accounting for true random slopes heavily increases Type 1 errors when *p*-values are used for inference. This may be problematic as there is no currently available tool that computes BFs and adequately accounts for the random-effects structure of continuous predictors. That is, using a Bayesian model with only a fixed effect on the continuous predictor might overestimate the evidence in favor of the Alternative.

The concept of Type 1 and Type 2 errors does not fit well into the Bayesian framework. Bayesian inference – whether it is based on posterior distributions or based on BFs – does not involve accepting or rejecting a hypothesis, but rather assigns hypotheses a posterior degree of credibility. However, if the posterior degree of credibility of a hypothesis is overestimated, it leads our belief about the hypothesis in a wrong direction. This problem may even be aggravated when a decision criterion to decide between competing hypotheses is applied (e.g., does the 95 % HDI of a posterior exclude a region of practical equivalence, Kruschke & Liddell, 2017, or, do we stop data collection when the BF in favor of a hypothesis exceeds a threshold, Rouder, 2014). Therefore, we tested whether not accounting for actually present random slopes overestimates the evidence in favor of the Alternative when the Null is actually true (i.e., effect size on the fixed effect is zero). In the following, we refer to that overestimation as a "potential Type 1 error".

To this end, we again generated data from an experiment as in Study 1 with one continuous predictor and by-subject random slopes. We varied the number of subjects between 20, 40, and 60, the number of observations per design cell between 3, 6, and 9, and the effect size of the continuous predictor between 0, .15, .3, .45, .6, and .75. We fit the data once with only a by-subject random-intercepts model using the BayesFactor package that does not allow the specification of random slopes on continuous predictors, and once with a

by-subject random-intercepts and random-slopes model with BayesRS. Specifically, we were interested to investigate what the models infer when the mean effect of the continuous variable is zero, as this will be informative for potential Type 1 errors. For illustrative purposes, we classified the computed BFs into the categories suggested by Kass and Raftery (1995): Strong Null (BF < 0.1), Substantial Null (0.1 < BF < 0.3125), Ambiguous (0.3125 < BF < 3.2), Substantial Alternative (3.2 < BF < 10), and Strong Alternative (BF > 10). The results of Study 4 are shown in Figure 6 and Figure 7.

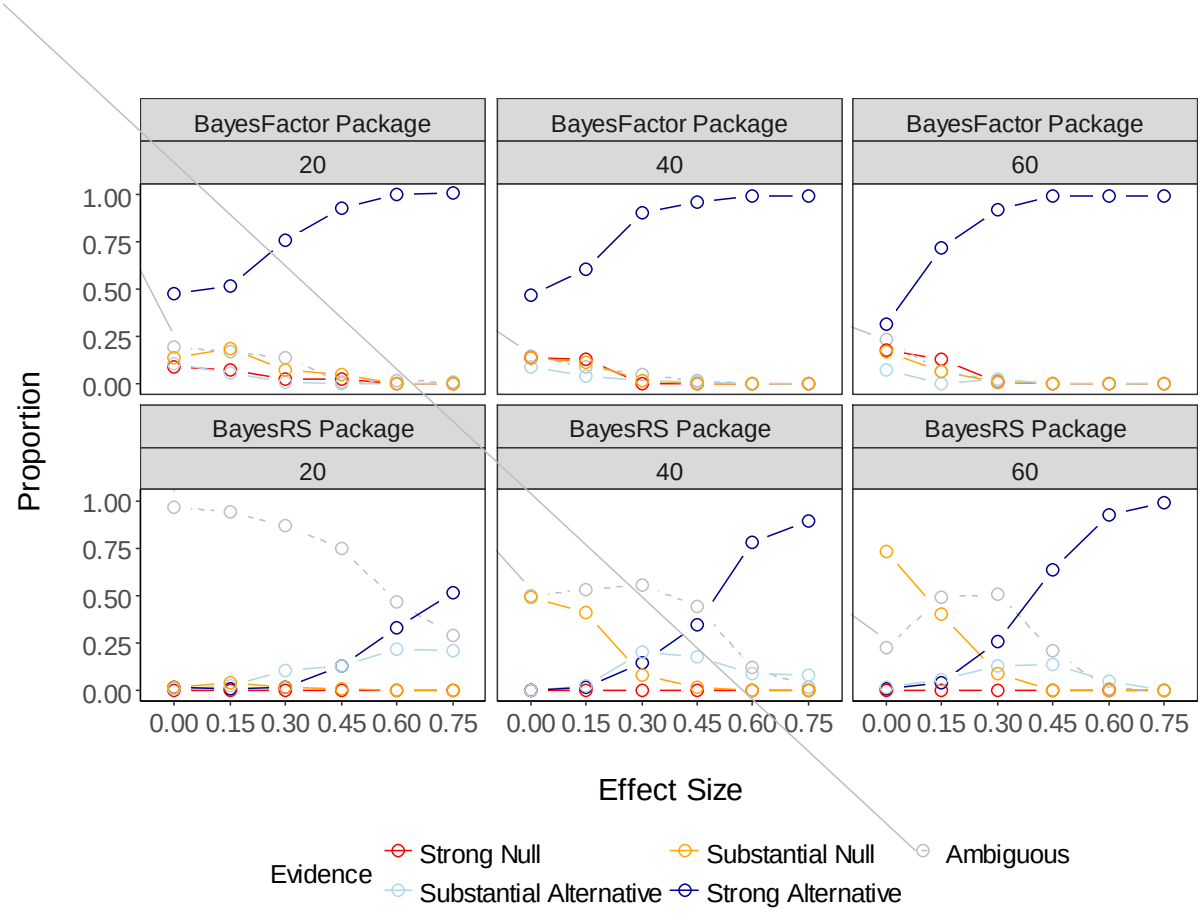

Figure 6. Proportion of BFs falling in one of the five categories suggested by Kass and Raftery (1995) plotted against the true effect size within the different design cells.

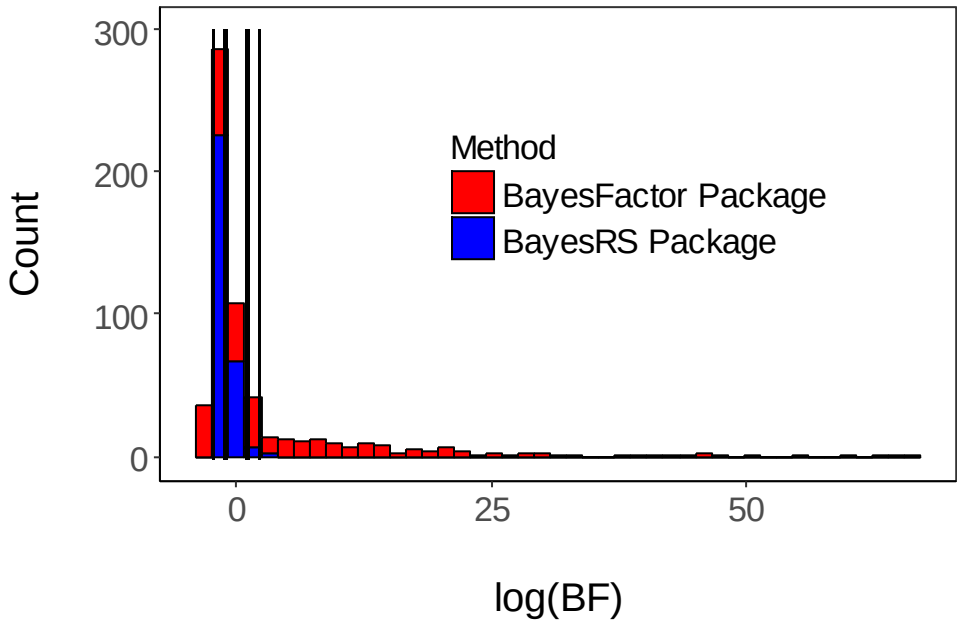

Figure 7. Distribution of log (BFs) when the effect size is 0. Note. The range of the x-axis differs between the two panels. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

First, let us focus on the model without a random slope that was run with the BayesFactor package and is shown in the upper row of Figure 6. When effect size is zero, the model massively overestimates the evidence in favor of the alternative hypothesis. Although in this scenario the BF should favor the null hypothesis, in 42 % of the data sets the BF is regarded as strong evidence, and in a further 9 % as substantial evidence, in favor of the Alternative. In 20 % the BF is ambiguous, and in only 14 % and 15 % it shows strong and substantial evidence in favor of the Null, respectively. The red histogram in Figure 7 shows

the distribution of BFs from the BayesFactor package when the true effect is zero. The long tail to the right reflects strong but wrong evidence in favor of the Alternative.

Now, let us focus on the bottom row in Figure 6 that shows the results of BayesRS. When focusing on effects of size zero, in only 1 % of the data sets the BF reflects strong evidence, and in a further 1 % substantial evidence in favor of the Alternative hypothesis. In 41 % of the data sets the BF reflects substantial evidence in favor of the Null, and in the remaining 57 % the evidence is ambiguous. The blue histogram in Figure 7 confirms that the log(BF) values from BayesRS are predominantly in the normative region < 0, and rarely stray much into the positive region. Hence, accounting for the random slopes decreases potential Type 1 errors when using BFs for inference, analogous to the conclusion of Barr et al. (2013) for frequentist mixed-effects models.

As can be seen in Figure 6, fewer potential Type 1 errors comes at the expense of less sensitivity to detect an effect in the model with a random slope. It takes more subjects and larger effect sizes for the BF to grow larger than 10 for the alternative hypothesis when it is true. This is, however, not an undesirable property because it essentially means that more data are needed to decide about the existence of small than of large effects. Although there is a chance of about 28 % for the BF to show substantial evidence in favor of the Null when the actual effect is .15, the BF is ambiguous in most of the cases, and it never shows strong evidence in favor of the Null. We argue that the increased potential Type 1 error rate in the model without random slopes is more problematic than the conservative nature of the model with random slopes. That is, in 42 % of the cases when there are true random slopes on the continuous predictor but the fixed effect is zero, researchers omitting random slopes from the model would conclude that there is strong evidence in favor of the effect. This argument weighs even more when it is considered that committing Type 1 errors has been regarded to be worse than committing Type 2 errors (Neyman, 1950). The results of this study therefore

buttress the claim that random slopes have to be accounted for, also for continuous predictors, in Bayesian as much as in frequentist mixed-effects models.

Simulation Study 5

In a last simulation study, we further explored a point put forward by Barr et al. (2013). Barr and colleagues observed that models that do not account for actually present correlations between random effects lead to slightly increased Type 1 errors in some circumstances, but also to slightly increased power. We incorporated a feature in BayesRS that allows users to model correlations between random effects. Here, we tested whether not accounting for actually present correlations between random slopes leads to increased potential Type 1 error rates, and to increased potential power when the BF is used for inference.

We simulated data from an experiment that slightly differed from the previous ones. In addition to a continuous predictor we now also included a categorical predictor. The by-subject slopes of the continuous variable and the categorical variable were drawn from a multivariate normal distribution. Thereby, (a) the mean effect of the continuous variable was fixed at zero to explore overestimation of the evidence (potential Type 1 error), but the mean effect of the categorical variable was either 0.2 or 1, also to explore a potential boost of the evidence for a true effect (i.e., potential power), (b) the correlation between the by-subject slopes of the two predictors was either 0.2, 0.5, or 0.8, and (c) the standard deviations of the by-subject slopes of both predictors were fixed to 1. This time, we simulated 100 data sets with 8 observations for every design cell of the simulated experiment and saved 100’000 MCMC steps for two experiments with 20 and 100 subjects, respectively.

First, let us focus on the continuous predictor for which there was no true mean effect, which is diagnostic for potential Type 1 errors. The distribution of BFs is plotted in Figure 8

for the two different model types (i.e., models with/without modeled correlations). As the results did not vary systematically, neither with the size of the correlation between the random slopes nor with the size of the categorical effect, we pooled across these variables. Neither model leads to more than a negligible number of potential Type 1 errors. Across all generated data sets, the model without correlations led to 0.1 % of BFs larger than 10 (2 cases). The model with correlations led to 0.06 % of BFs larger than 10 across all generated data sets (1 case). Hence, accounting for the correlations only has a negligible effect on potential Type 1 error rate. It is clear that inference regarding the fixed effect of the continuous variable does not differ between the two models when the sample size is large (lower row). However, when the sample size is small (upper row) we can observe that the model without correlations is actually slightly more sensitive to detect a Null effect. In other words, the proportion of BFs that provide substantial evidence in favor of the Null is larger in the model without correlations than in the model with correlations, and this is at the expense of fewer BFs categorized as ambiguous in the model without correlations, corroborating the findings by Barr et al. (2013). Considering these results, we conclude that not accounting for actually present correlations between random slopes is not problematic, and may even be desirable when only data from a few subjects are available.

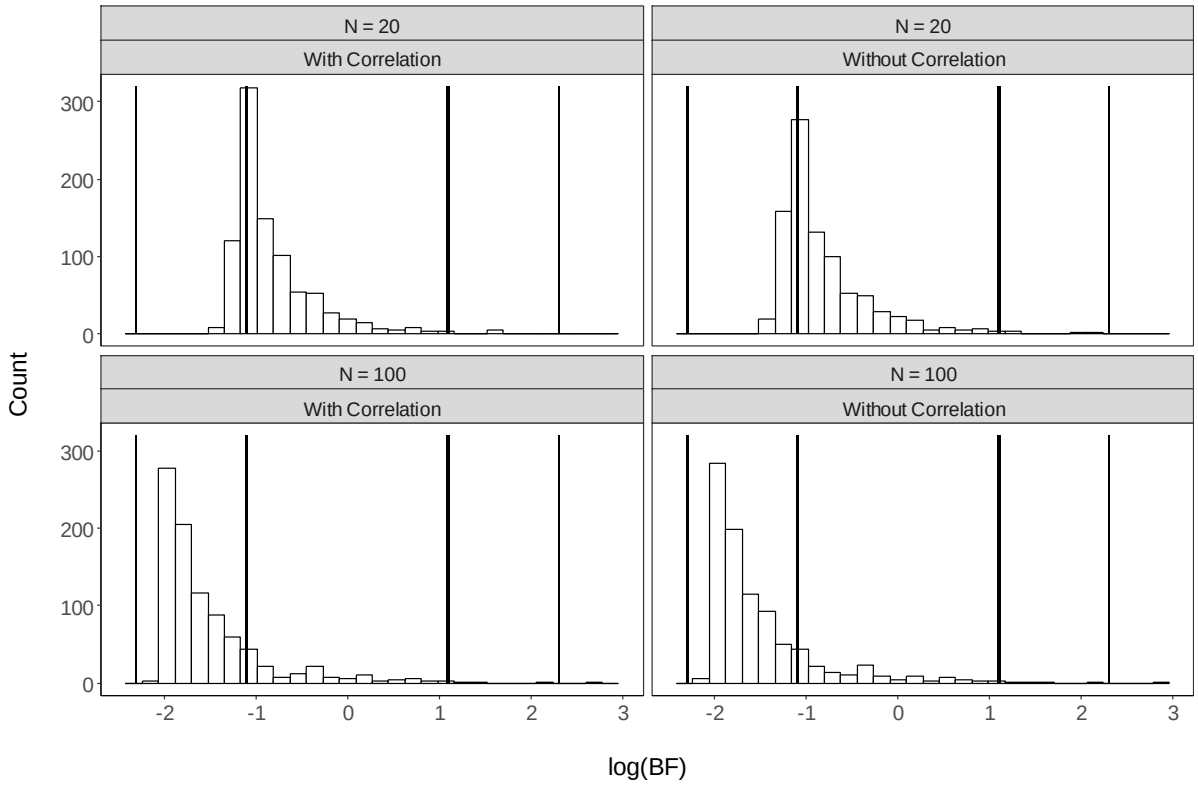

Figure 8. Distribution of log (BFs) for continuous predictor with true effect of 0 when accounting for the correlation (upper row) and when not accounting for the correlation (lower row). Note. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

Second, let us focus on the categorical predictor whose effect size was varied between 0, .2, and 1, which is diagnostic for potential Type 2 errors. Density plots of the BFs are depicted in Figure 9. We again pooled across correlation size. Here, the differences between the two model types are slightly larger. The model without correlations leads to slightly decreased potential Type 2 error rates when the effect size of the categorical variable is small (i.e., .2) and when there are only 20 subjects in the experiment. This can be seen in the

middle plot of the upper row: The proportion of BFs that are classified as ambiguous/substantial Null is smaller in the model without correlations than in the model with correlations, and instead the model without correlations yielded more BFs reflecting substantial or strong evidence in favor of the alternative hypothesis. When there are 100 subjects in the experiment, this tendency is reduced but still present. Considering both potential Type 1 and potential Type 2 error rates, it appears unproblematic to ignore actually present correlations between random slopes. In contrast, the present results even suggest that a model that does not account for the correlation may be the better choice for determining whether or not a fixed effect is present. That is, while potential Type 1 error rates are largely the same, the model omitting correlations has a higher sensitivity to detect small effects in small samples, confirming for the Bayesian framework what Barr et al. observed in the frequentist framework.

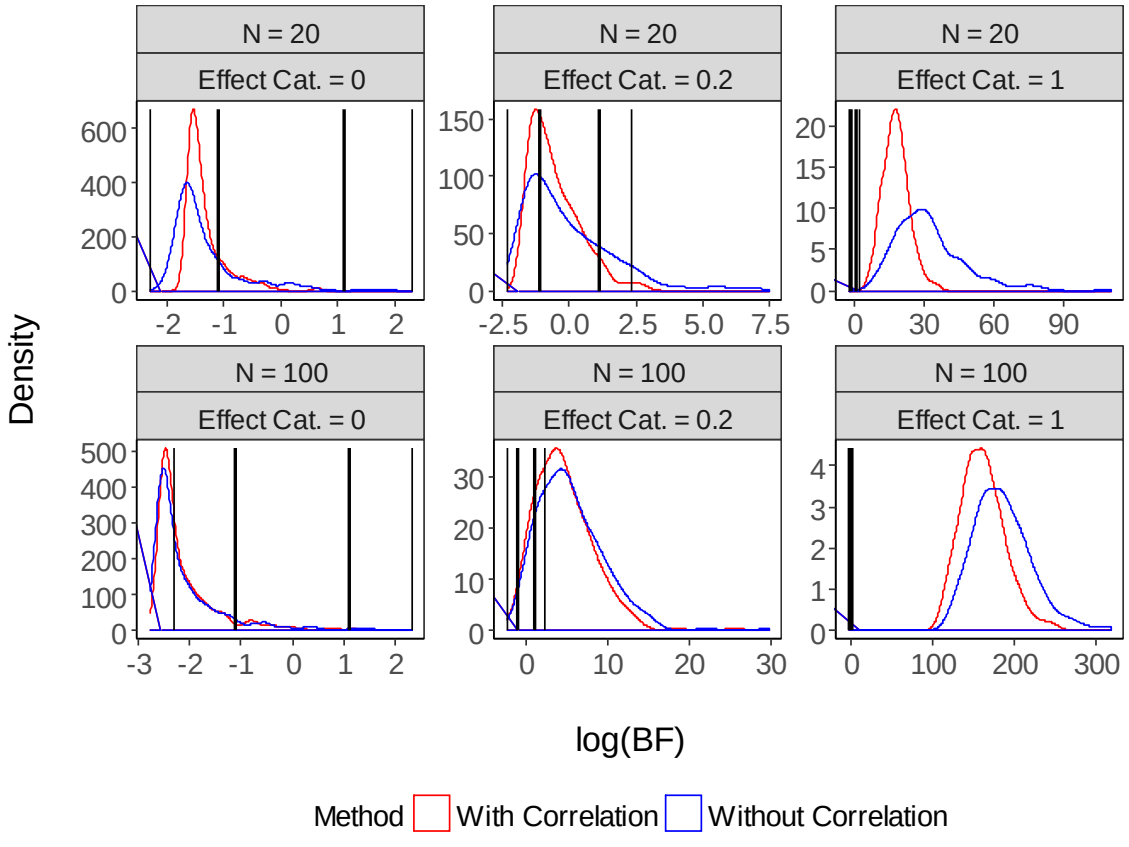

Figure 9. Density plots of log (BFs) within the different design cells. Note that the ranges of the y- and x-axes between panels differ. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

General Discussion

Statisticians have advocated the use of Bayesian statistics for inference instead of *p* values, and the use of mixed-effects models or, more generally, hierarchical models. Mixed-effects models tend to overestimate the evidence in favor of the alternative hypothesis when they erroneously fail to include random slopes (Barr et al., 2013). As our Simulations 4 and 5 show, this is not only true for frequentist but also for Bayesian methods of inference on mixed-effects models. This raises a problem, because so far, there was no easy-to-use statistical tool that computes BFs for models with random slopes on continuous predictors. Here, we closed this gap by introducing a new R package, BayesRS. The new package is also able to model correlations between random effects. We tested its functionality in five simulation studies.

The first study showed that the BF in favor of a fixed effect with associated random slopes can be estimated with good reliability. Especially BFs that are critical to decide whether the evidence in favor of an effect is not worth to be mentioned, substantial, or strong (Kass & Raftery, 1995) can be estimated with high reliability. When BFs grow larger their estimation becomes more variable. A BF in the region of several thousand may vary as much as about 20 %. However, usually it does not matter whether a BF is 3000 or 2400. Either value reflects decisive evidence in favor of a hypothesis. The second study showed that the BF of a fixed effect of a continuous predictor without associated random slopes is a good estimate of the true BF when it is small, but it slightly overestimates BFs when they grow

large. In the third study, the BF of a fixed effect with associated random slopes was a good approximation of the BF of a Bayesian *t*-test on the true *b*-values. Again, BayesRS overestimated the BF when the BF of the *t*-test grew large. To summarize, the first three studies provide evidence that BayesRS estimates BFs that are reliable and valid (i.e., a good approximation of the true BF).

In two further studies (Simulations 4 and 5), we showed that not accounting for true random slopes overestimates the evidence in favor of the alternative hypothesis, and often signals strong evidence for it when it is actually not true, which echoes previous findings using *p* values for inference (Barr et al., 2013). Although there are clear differences between frequentist and Bayesian statistics, we argue that accounting adequately for the structure in the data is necessary regardless of the statistical framework. This means that if there are true differences in an effect between observation units but the differences are not accounted for in the structure of the statistical model, a simplification of the structure will lead to biased inference. Simulation 5 also showed that modeling correlations between random effects leads to a slightly increased potential Type 2 error rate when both effect size and sample size are small. Therefore, when correlations between random effects are not of primary interest in a study, not accounting for them in the model may actually slightly increase the sensitivity to detect a true effect.

Whereas we strongly advocate including random slopes in a model, we do not advise to do so thoughtlessly, but rather we recommend to carefully consider how the data are structured. For example, the debate about how to set up an appropriate random-effects structure (Barr et al., 2013; Matuschek et al., 2017) highlights that researchers should carefully consider how their data are structured and inform their statistical models based on this consideration. Matuschek and colleagues suggest to test in a first step whether there is evidence for the random slopes in the data by using a fit criterion. If there is not, researchers

could proceed omitting the random slopes. Further, the assumption that effects are distributed according to a normal distribution across units of observation is not going to be appropriate in all circumstances. For example, some subjects may be sensitive to an experimental manipulation, whereas others are not, which essentially leads to a bimodal distribution. To test such a hypothesis we incorporated an option in BayesRS that allows users to output the individual β values of an effect for all units of observation, so they can investigate their distribution. If the assumption of a normal distribution is not justified, researchers are encouraged to use a different model to analyze the data.

References

Auguie, B. (2016). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.2.1. https://CRAN.R-project.org/package=gridExtra

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Special Issue: Emerging Data Analysis*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bayes, T., Price, R., & Canton, J. (1763). *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Chapman & Hall/CRC Boca Raton, FL, USA.

Hubbard, R. (2004). Alphabet Soup. *Theory & Psychology*, *14*(3), 295–327. https://doi.org/10.1177/0959354304043638

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 1–29. https://doi.org/10.3758/s13423-016-1221-4

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.7). Retrieved from http://cran.at.r-project.org/web/packages/BayesFactor/index.html

Neyman, J. (1950). First course in probability and statistics. *First Course in Probability and Statistics, by J. Neyman. Published by Henry Holt, 1950.*

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2014). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. *Available at H ttp://CRAN. R-Project. Org/package= Nlme*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling (Vol. 124, p. 125). Presented at the Proceedings of the 3rd international workshop on distributed statistical computing, Vienna.

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC. *R Package Version*, *3*.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583.

Team, R. C. (2017). R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2014.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752–760.

Appendix

Appendix A – Used R packages

The MASS package to sample from multivariate distributions and from univariate distributions with known sample statistics

The BayesFactor package to compute BFs

The ggplot2, grid, and gridExtra packages for graphing purposes

The metRology package to compute densities of scaled Cauchy distributions

The reshape package for data processing

The rjags and coda packages to run Bayesian models via JAGS