AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2025, Vol. 154, No. 1, 294–304
https://doi.org/10.1037/xge0001666

A Cautionary Note Against Selective Applications of the Bayes Factor

Marcel R. Schreiner and Wilfried Kunde
Department of Psychology, Julius-Maximilians-Universität Würzburg

Bayes factor analysis becomes increasingly popular, among other reasons, because it allows to provide evidence for the null hypothesis, which is not easily possible with the traditional frequentist approach. A conceivable strategy that apparently takes favorable aspects of both approaches on board is to use traditional frequentist analyses first and to support theoretically interesting nil effects by Bayesian analyses thereafter. Here, we asked whether such a selective application of Bayesian analyses to only nonsignificant effects of foregoing frequentist analyses creates bias. In two simulation studies, we observed that such selective application of Bayesian analyses, in fact, severely overestimates evidence in favor of the null hypotheses, when a true population effect exists. While this bias can be attenuated by using more informative priors in the Bayesian analyses, we recommend to not apply such selective combination of analytical approaches, but instead to use either frequentist or Bayesian analyses consistently.

Public Significance Statement
This study suggests that selectively computing Bayes factors after a nonsignificant frequentist test result leads to an overestimation of evidence in favor of the null hypothesis if a true effect exists in the population. This bias can be reduced, but not fully avoided, by using more informative priors. The results of this study suggest that frequentist and Bayesian analysis approaches should be applied consistently to derive unbiased statistical inference.

Keywords: Bayes factor, Bayesian *t* test, Bayesian analysis of variance, null-hypothesis testing

Supplemental materials: https://doi.org/10.1037/xge0001666.supp

The use of Bayes factors has become increasingly common in evaluating hypotheses, particularly because the use of *p* values and classical null-hypothesis testing has received repeated criticism (e.g., Krueger, 2001; Rozeboom, 1960; Wagenmakers, 2007). The Bayes factor quantifies the evidence in the data for two competing statistical models (or hypotheses translated into statistical models; e.g., Dienes, 2011; Edwards et al., 1963; Heck et al., 2023; Kass & Raftery, 1995). Given a model reflecting the null hypothesis (M_0) and a model reflecting the alternative hypothesis (M_1), the Bayes factor in favor of the alternative hypothesis (BF_{10}) can be computed as the ratio of the probability of the observed data (D) under the two models (see Equation 1):

$$BF_{10} = \frac{(D|M_1)}{(D|M_0)}. \quad (1)$$

This measure can be interpreted such that the obtained data are BF_{10} as likely under the alternative hypothesis than under the null hypothesis. Thus, a Bayes factor larger than 1 indicates evidence in favor of the alternative hypothesis, a Bayes factor smaller than 1 indicates evidence in favor of the null hypothesis, and a Bayes factor of 1 indicates the data not favoring either hypothesis over the other. Typically, Bayes factors between 0.33 and 3 are considered weak or anecdotal evidence, whereas Bayes factors smaller than 0.33 or larger than 3 are considered substantial evidence (Jeffreys, 1961; see also Lee & Wagenmakers, 2014). Notably, Bayesian inference requires the specification of a prior distribution for the model parameters, specifying which parameters are more or less likely before considering the observed data. These prior distributions can either be subjective, reflecting the expectations of the researcher, or one can use default prior distributions that are chosen to receive

This article was published Online First October 7, 2024.
Joseph Toscano served as action editor.
Marcel R. Schreiner https://orcid.org/0000-0001-7719-6281
Wilfried Kunde https://orcid.org/0000-0001-6256-8011
All data, analysis code, and research materials have been made publicly available on the Open Science Framework and are accessible at https://osf.io/jnt6w/. This study was not preregistered. This article has been published as a preprint on PsyArXiv at https://doi.org/10.31234/osf.io/kucvh, and a reproducible version of the article has been made available on the Open Science Framework at https://osf.io/jnt6w/. The authors have no conflicts of interest to declare.

Marcel R. Schreiner played a lead role in data curation, formal analysis, investigation, methodology, software, visualization, and writing–original draft and an equal role in conceptualization and writing–review and editing. Wilfried Kunde played a lead role in project administration, a supporting role in writing–original draft, and an equal role in conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Marcel R. Schreiner, Department of Psychology, Julius-Maximilians-Universität Würzburg, Röntgenring 11, 97070 Würzburg, Germany. Email: marcel.schreiner@uni-wuerzburg.de

well-calibrated Bayes factors (but do not reflect prior beliefs of a researcher; Heck et al., 2023).[1] Whereas the *p* value can only provide evidence in favor of the alternative hypothesis (in case of a significant result), but remains impartial in terms of evidence for the null hypothesis (in case of a nonsignificant result; Fisher, 1935), the Bayes factor allows to quantify both evidence in favor of the alternative hypothesis and evidence in favor of the null hypothesis. Therefore, it has been recommended (e.g., Dienes, 2014; Dienes et al., 2018) and has become increasingly common (e.g., Janczyk et al., 2019; Qiao et al., 2023; Tsuji & Imaizumi, 2022) to compute Bayes factors to gain additional insights into nonsignificant results yielded by frequentist tests, that is, to determine whether the nonsignificant result is due to the data providing evidence for the null hypothesis (and evidence against the alternative hypothesis) or whether the evidence provided by the data is insensitive to distinguish between the two hypotheses. It is our informal observation that the practice to compute Bayes factors selectively for nonsignificant effects is recommended quite often by reviewers and editors, and we have done so ourselves (Klaffehn et al., 2018; Weller et al., 2020; Wirth & Kunde, 2020).

Here we asked whether such selective applications of the Bayes factor in case of nonsignificant frequentist tests is problematic. Given that nonsignificant effects in a frequentist test constitute a nonrandom subsample of possible cases, selectively computing Bayes factors for these effects results in an overweighting of these nonsignificant findings, which may increase the likelihood for observing evidence in favor of the null hypothesis. While this is a problem for individual studies, as obtained evidence in favor of the null hypothesis would be biased, the continued application of this practice would lead to a misrepresentation of the evidence in the published literature. This may be a problem for meta-analyses, particularly if they specifically rely on reported Bayes factors rather than a combined assessment of results from frequentist and Bayesian testing. The problem of computing Bayes factors conditional on the outcome of a frequentist test is also highlighted in Dienes (2024).

In essence, the problem can be described as one of selective inference (Benjamini, 2010; Benjamini & Bogomolov, 2014), that is, focusing statistical inference on a subgroup of findings after having viewed the data (Benjamini, 2020). By selectively computing Bayes factors, one would essentially split frequentist test results into two families—significant and nonsignificant results—and apply an additional Bayesian analysis to only one of the two families. However, this limits the interpretability of measures of uncertainty (Benjamini, 2010). In the described case, the distribution of Bayes factors in the subsample for which they are computed is not representative for the full population of cases. When applying selective inference, one should use a more conservative testing criterion (Taylor & Tibshirani, 2015). Thus, the problem of selective applications of the Bayes factor may be further aggravated by the reliance on default prior distributions, as is likely the case for most researchers. First, the reliance on default prior distributions prevents the application of a more conservative testing criterion for the Bayes factor analysis. Second, in this two-step approach, the information provided by the previous nonsignificant test is ignored, even though this test was used to inform the decision for conducting the second analysis. Ignoring this additional information from the frequentist test would lead to a relatively less informative prior distribution, reflecting a vaguer assumption about the effect in question. This relatively increased vagueness increases the chance of getting evidence for the null hypothesis using Bayesian inference (see Dienes, 2014). However, the fact that an initial frequentist test yielded a nonsignificant result can be used to inform the construction of a more conservative testing criterion for the secondary Bayes factor analysis. This draws parallels to the empirical Bayes method (e.g., Casella, 1985; Efron, 2008), in which empirical data are used to estimate prior distributions. While we do not argue that the default prior distributions should be updated with the empirical data, using the information from the nonsignificant frequentist test to inform the selection of a narrower prior distribution may be beneficial. Using a more "informative" prior distribution centered around the parameters in the model reflecting the null hypothesis would take the additionally available information from the initial frequentist test into account and lead to a more conservative test for the null hypothesis (cf. Dienes, 2024). Therefore, the practice of selectively applying Bayes factors upon obtaining nonsignificant results may lead to an overestimation of the evidence in favor of the null hypothesis, but this may be, at least partly, mitigated by the use of prior distributions being more informative than the default implementations. Note that with a more informative prior distribution, we refer to the decreased vagueness of the prior or reduced width of the distribution but not to the prior being directly informed by scientific evidence.

We demonstrate the problem of selectively applying Bayes factors upon obtaining nonsignificant results from a frequentist test in two simulation studies examining two widespread effects of interest: mean differences between two groups (using an independent-samples *t* test) and interactions between two categorical variables in a 2 × 2 within-subjects design (using a repeated measures analysis of variance [ANOVA]). For both effects, we assume point-null hypotheses, as are typically assumed in classical null-hypothesis testing, but the problem applies to equivalence testing as well (Dienes, 2024). We show that selectively applying Bayes factors indeed leads to an overestimation of evidence in favor of the null hypothesis if the alternative hypothesis is actually true (i.e., there is a population effect) but not if the null hypothesis is actually true (i.e., there is no population effect), compared to nonselectively computing Bayes factors. We further show that this problem can at least be partially mitigated by using more informative prior distributions when selectively computing Bayes factors after receiving a nonsignificant result with a frequentist test.

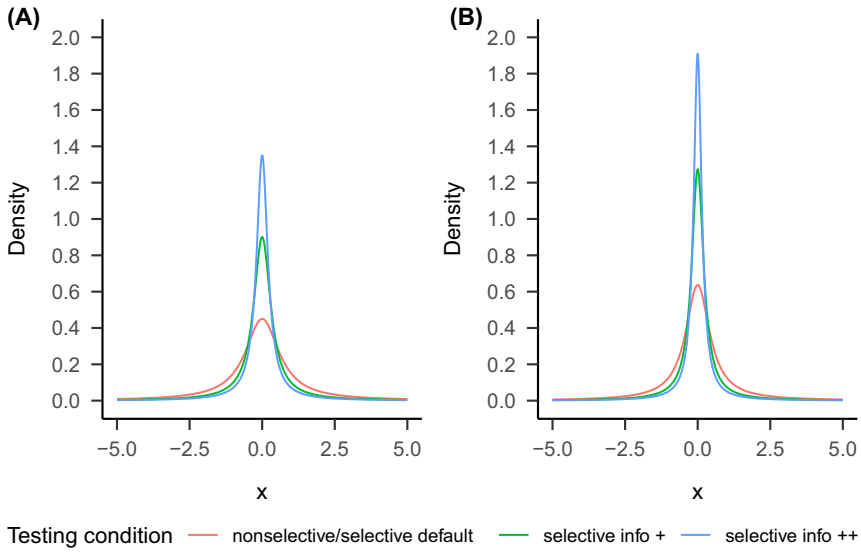
Simulation Study 1: Independent-Samples *t* Test

In the first simulation study, we considered the case of comparing the means of two groups or conditions, for example, when evaluating the effectiveness of a treatment by comparing an intervention and a control group. Here, the null hypothesis would state that there is no mean difference between the two groups, whereas the (undirected) alternative hypothesis would state that there is a mean difference between the two groups. To distinguish between these hypotheses, an independent-samples *t* test is commonly applied.

[1] Note that this distinction for the selection of prior distributions is not the only possible one, but distributions may also be derived from the specified predictions of a theory (see Dienes, 2019).

Figure 1

Probability Density Functions of the Prior Distributions Used in the Different Testing Conditions in Simulation Study 1 (A) and Simulation Study 2 (B)



Note. See the online article for the color version of this figure.

Method

For the simulation, we assumed a one-factorial between-subjects design with two groups or conditions. Responses were sampled from two normal distributions. For the first group, responses were sampled from a standard normal distribution with $M = 0$ and $SD = 1$. For the second group, we varied the mean of the normal distribution, being either 0, 0.2, or 0.5 (i.e., the mean of the first group was the same or smaller than the one of the second group), while the standard deviation was kept at $SD = 1$. This corresponds to effect sizes of $d = 0$, $d = 0.2$, and $d = 0.5$ for the mean difference between the two groups, reflecting no, a small, or a medium effect, respectively, according to Cohen (1988).[2] We varied the sample size, ranging from $n = 10$ to 70 in increments of 10 per group, and additionally included sample sizes of $n = 100$ and $n = 150$. Bayesian independent-samples t tests were conducted using the R package BayesFactor (Morey & Rouder, 2022). Note that other packages (e.g., Jeffreys's Amazing Statistics Program, JASP) use comparable algorithms, such that the results we report are not restricted to specific implementations in the BayesFactor package (Rouder et al., 2017). In a nonselective testing condition, a Bayesian independent-samples t test was always conducted, independent of the outcome of the frequentist independent-samples t test. In this condition, we used the default prior distribution of the BayesFactor package, which is a Cauchy distribution with a scale parameter of $r = \sqrt{2}/2$ (cf. Morey et al., 2011; Rouder et al., 2012). In the selective testing conditions, a Bayesian independent-samples t test was only conducted when the frequentist independent-samples t test yielded a nonsignificant result at a significance level of $\alpha = 5\%$. There were three selective testing conditions. In the first condition (selective default), we used the default prior distribution; in the second condition (selective info+), we used a more informative prior distribution with a scale parameter of $r = \sqrt{2}/4$; and in the third condition (selective info++), we used an even more informative prior distribution with a scale parameter of $r = \sqrt{2}/6$. Figure 1A shows the prior distributions for the different testing conditions. Our simulation design was therefore a 9 (n) × 3 (effect size) × 4 (testing condition) design, resulting in 108 simulation conditions. For each simulation condition, we conducted 10,000 replications.

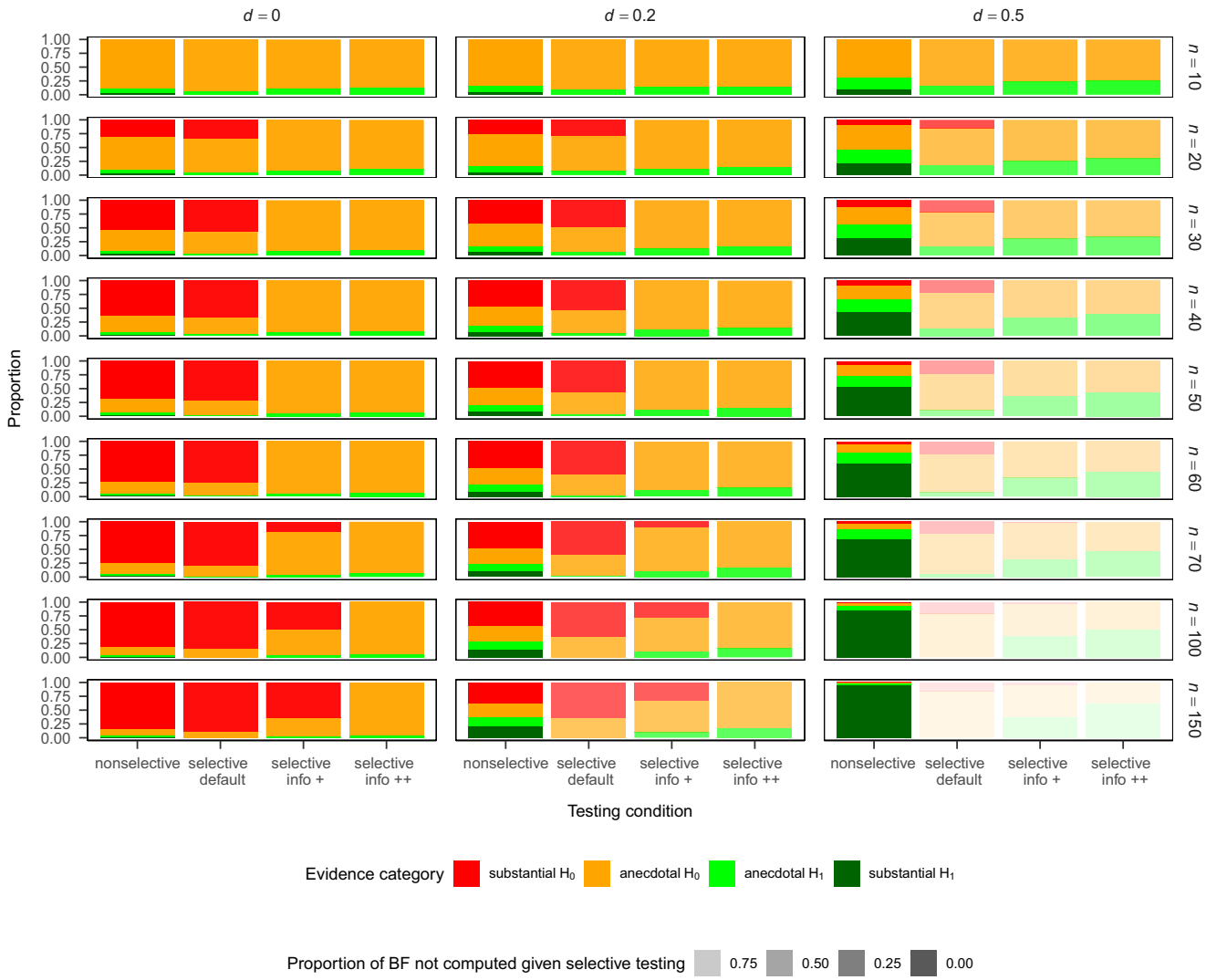
Results

The proportion of Bayes factors indicating substantial or anecdotal evidence in favor of the null hypothesis (H_0) or alternative hypothesis (H_1) across replications in the different simulation conditions is shown in Figure 2. The distribution of Bayes factors (cut off at the value of 7 for better readability) in the different testing and effect-size conditions is shown in Figure 3, exemplary for a sample size of $n = 30$. Results for different sample sizes were similar. Note that for the selective testing conditions, the number of Bayes factors that are actually computed is lower than for the nonselective testing condition, because Bayes factors are only computed given a nonsignificant result in the frequentist test. Therefore, the number of Bayes factors

[2] As a robustness check, we also used a different data-generating model in a supplemental simulation for which we sampled responses from a Cauchy distribution with a location parameter of $x_0 = 0$ and a scale parameter of $r = 1$, with varying location parameters of $x_0 = 0$, $x_0 = 2$, and $x_0 = 5$ for the second group. This roughly corresponded, on average, to effect sizes of $d = 0$, $d = 0.2$, and $d = 0.5$, respectively, for the mean difference between the two groups. The results largely mirrored the ones from the main simulation, demonstrating the robustness of the results to different data-generating models and, in this case, also to nonnormality of the underlying data. We do not report the results here, but the code and results for the additional online materials are available on the Open Science Framework at https://osf.io/jnt6w/.

Figure 2

Proportion of BFs Indicating Substantial or Anecdotal Evidence for the Null Hypothesis (H_0) or the Alternative Hypothesis (H_1) in Simulation Study 1



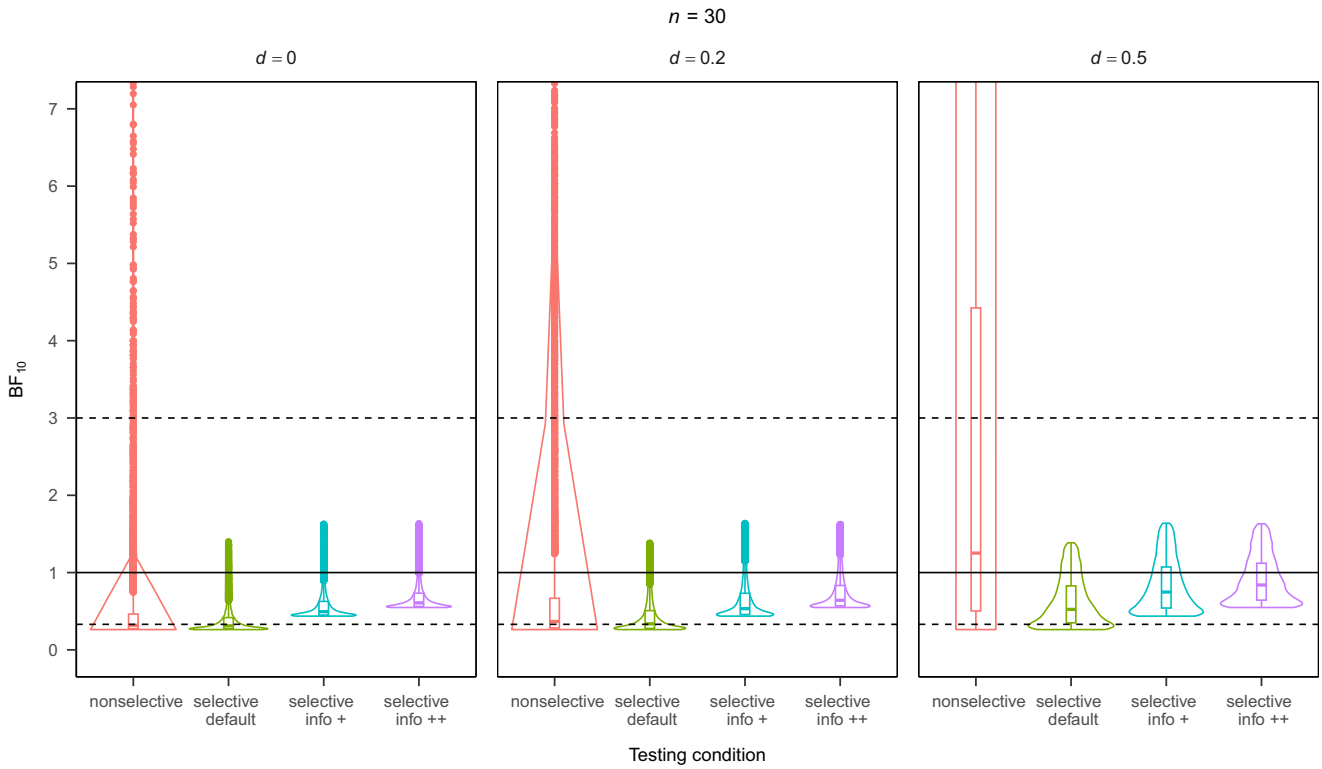
Note. The proportion of BFs that were not computed is only relevant for the selective testing conditions and equals the proportion of frequentist tests reaching significance, therefore being inversely related to the statistical power of the frequentist test. BF = Bayes factor; substantial H_0 = substantial evidence in favor of the null hypothesis ($BF_{10} \leq 0.33$); anecdotal H_0 = anecdotal evidence in favor of the null hypothesis ($0.33 < BF_{10} < 1$); anecdotal H_1 = anecdotal evidence in favor of the alternative hypothesis ($1 < BF_{10} < 3$); substantial H_1 = substantial evidence in favor of the alternative hypothesis ($BF_{10} \geq 3$). See the online article for the color version of this figure.

computed in the selective testing conditions (i.e., the number of cases where Bayes factors were selectively applied) is inversely related to the statistical power of the frequentist test. This is represented by the opacity of the bars in Figure 2.

Given that H_0 is true (i.e., there is no difference in means between the two groups on the population level), in the nonselective testing condition, 93.27% of Bayes factors indicated substantial or anecdotal evidence in favor of H_0, collapsed across the different sample size conditions. In the selective testing conditions, this was the case for 97.95% (selective default), 94.31% (selective info+), and 92.21% (selective info++) of Bayes factors. Compared to the nonselective testing condition, these proportions are increased when using the default prior but very similar when using more informative priors, which is also reflected in an upward shift in the Bayes factor distributions when using more informative priors. However, whereas in the nonselective testing condition, 1.51% of Bayes factors indicated substantial evidence for H_1, this was not the case for any of the selective testing conditions. This is also reflected in the Bayes factor distributions being much more compressed in the selective testing conditions than in the nonselective testing condition. Using more informative priors resulted in most Bayes factors indicating anecdotal evidence in favor of H_0. However, there was still no Bayes factor indicating substantial evidence for H_1.

Figure 3

Distribution of BFs (Violin and Box Plots) for a Sample Size of n = 30 in Simulation Study 1



Note. The *y* axis is cut off at the value of 7 for better readability. The maximum BFs were 236 ($d = 0$), 4,457 ($d = 0.2$), and 51,388 ($d = 0.5$), which all occurred in the nonselective testing condition. The solid horizontal black line indicates a BF of 1, whereas the dashed horizontal lines indicate BFs of 0.33 and 3, which are the values below or above which BFs are typically considered to reflect substantial evidence (Jeffreys, 1961). BF = Bayes factor. See the online article for the color version of this figure.

These results suggest that, given that H_0 is true, selective applications of the Bayes factor are rather unproblematic, as the results of the selective default testing condition largely mirror the ones of the nonselective testing condition. Using more informative priors leads to more ambiguous evidence, which is more in line with the frequentist interpretation of nonsignificant results. However, if H_0 is true, using the default prior leads to a better representation of the absent population effect.

Yet, the picture looks different if H_1 is true (i.e., there is a difference in means between the two groups on the population level). Here, in the nonselective testing condition, 78.45% (when $d = 0.2$) and 30.30% (when $d = 0.5$) of Bayes factors indicated substantial or anecdotal evidence in favor of H_0, collapsed across the different sample size conditions, which would indicate evidence that is inconsistent with the actual population effect. For a visual inspection of this bias, compare the size of the red and orange bars between nonselective and selective conditions with $d = 0.2$ (and $d = 0.5$) in Figure 2. The distribution of Bayes factors is also very wide. In the selective default testing condition, 96.20% (when $d = 0.2$) and 86.25% (when $d = 0.5$) of Bayes factors indicated substantial or anecdotal evidence in favor of H_0. These values are severely increased compared to the nonselective testing conditions, even more so for the larger effect ($d = 0.5$), which is also reflected in the Bayes factor distributions being much more compressed compared to the nonselective testing condition. In the selective info+ testing condition, the respective proportions are 88.59% (when $d = 0.2$) and 71.08% (when $d = 0.5$) of Bayes factors, and in the selective info++ testing condition, the respective proportions are 84.43% (when $d = 0.2$) and 65.46% (when $d = 0.5$) of Bayes factors. These values are still increased compared to the nonselective testing condition, but less so than in the selective default testing condition. This is again also reflected in an upward shift in the Bayes factor distributions when using more informative priors.

These results suggest that, given that H_1 is true, selective applications of the Bayes factor lead to Bayes factors that reflect a gross overestimation of the evidence in favor of H_0. Again, using more informative priors leads to more ambiguous evidence, which is more in line with the frequentist interpretation of nonsignificant results. Therefore, the inferences drawn from Bayes factors computed with more informative priors are more appropriate than the ones drawn from Bayes factors computed with the default prior, given that H_1 is true and Bayes factors are computed selectively, but the evidence in favor of H_0 is still inflated compared to applying the Bayes factor nonselectively.

Simulation Study 2: 2 × 2 Within-Subjects ANOVA

In the second simulation study, we considered the case of investigating an interaction between two categorical variables in a 2 × 2 within-subjects design using a repeated measures ANOVA. Such a design is, for example, often applied in the investigation of serial process assumptions, where the interaction is of particular interest. Here, the additive factor method (Sternberg, 1969) is commonly applied. For example, one may assume that a stage of perception is followed by a decision stage. One may then manipulate two factors, such as stimulus intensity and the size of the choice set, and investigate the effect of these manipulations on response times. A conceivable model would hold that the manipulation of stimulus intensity affects the perception stage, but not the decision stage, and vice versa for the manipulation of choice set size. Therefore, the two factors are assumed to selectively influence one of the stages, and given the assumption that response times are the sum of the duration of the two stages, the combined effect on response times should be the sum of their individual effects (i.e., the effects of the two factors are additive). In a repeated measures ANOVA, this would be reflected in significant main effects of the two factors but a nonsignificant interaction. A significant interaction would indicate that the two factors influence the same stage. For demonstrating that factors selectively influence different stages, one would want to find a nonsignificant interaction, and it may be particularly tempting to selectively apply Bayes factors in this case (e.g., Aschenbrenner & Balota, 2019). The same theoretical interest to demonstrate the absence of an interaction applies to many other situations such that, for example, a factor impacting response times in a psychological refractory period paradigm does so at all possible intervals between tasks (e.g., Janczyk et al., 2019).

Method

For the simulation study, we assumed a 2 × 2 within-subjects design. Responses were sampled from a mixed linear model (see e.g., Hoffman & Rovine, 2007) with random person intercepts. Independent variables were effect coded. One value was sampled per participant and experimental condition, making the sampling procedure equivalent to a repeated measures ANOVA design. Fixed effects were set to $b = 0$ for the intercept and $b = 0.1$ for the two main effects. We varied the fixed effect for the interaction, being either $b = 0$, $b = 0.15$, or $b = 0.35$. This roughly corresponded to effect sizes of η^2_G (generalized eta squared) = 0, $\eta^2_G = .02$, and $\eta^2_G = .08$, reflecting no, a small, or a medium effect, respectively, according to Cohen (1988). The variance of the random intercept was set to 0.5, and the residual standard deviation was set to 1.[3] We again varied the sample size ranging from $N = 10$ to 70 in increments of 10 and additionally included sample sizes of $N = 100$ and $N = 150$. The remainder of the method was the same as in Simulation Study 1, except that we conducted (Bayesian) repeated measures ANOVAs and used different priors. In the BayesFactor package (Morey & Rouder, 2022), the default prior follows a Cauchy distribution with a scale parameter of $r = 1/2$ for fixed effects and $r = 1$ for random effects (cf. Rouder et al., 2012). We kept the default for random effects for all simulation conditions, but we varied the scale parameter for fixed effects, using either $r = 1/2$ (nonselective and selective default testing conditions), $r = 1/4$ (selective info+ testing condition), and $r = 1/6$ (selective info++ testing condition). Figure 1B shows the prior distributions in the different testing conditions. Our simulation design was again a 9 (N) × 3 (effect size) × 4 (testing condition) design, resulting in 108 simulation conditions. For each simulation condition, we conducted 10,000 replications. We computed Bayes factors comparing a model with the interaction with a model without the interaction to test for specifically an interaction effect.[4] Frequentist ANOVAs were conducted using the package afex (Singmann et al., 2023).[5]

Results

The proportion of Bayes factors indicating substantial or anecdotal evidence in favor of H_0 or H_1 across replications in the different simulation conditions is shown in Figure 4. The distribution of Bayes factors (cut off at the value of 7 for better readability) in the different testing and effect-size conditions is shown in Figure 5, exemplary for a sample size of $N = 30$. Results for different sample sizes were similar.

Given that H_0 is true (i.e., there is no interaction effect on the population level), in the nonselective testing condition, 93.98% of Bayes factors indicated substantial or anecdotal evidence in favor of H_0, collapsed across the different sample size conditions. In the selective testing conditions, this was the case for 99.41% (selective default), 98.59% (selective info+), and 97.83% (selective info++) of Bayes factors. Compared to the nonselective testing condition, these proportions are increased. This is particularly the case when using the default prior and slightly less so when using more informative priors, which is also reflected in an upward shift in the Bayes factor distributions when using more informative priors. Whereas in the nonselective testing condition, 2.78% of Bayes factors indicated substantial evidence for H_1, this was only the case for 0.02%, 0.02%, and 0.02% of Bayes factors in the selective default, selective info+, and selective info++ testing conditions, respectively. This is also reflected in the Bayes factor distributions being much more compressed in the selective testing conditions than in the nonselective testing condition. Using more informative priors resulted in most Bayes factors indicating anecdotal evidence in favor of H_0.

These results again suggest that, given that H_0 is true, selective applications of the Bayes factor are not problematic, as the results of

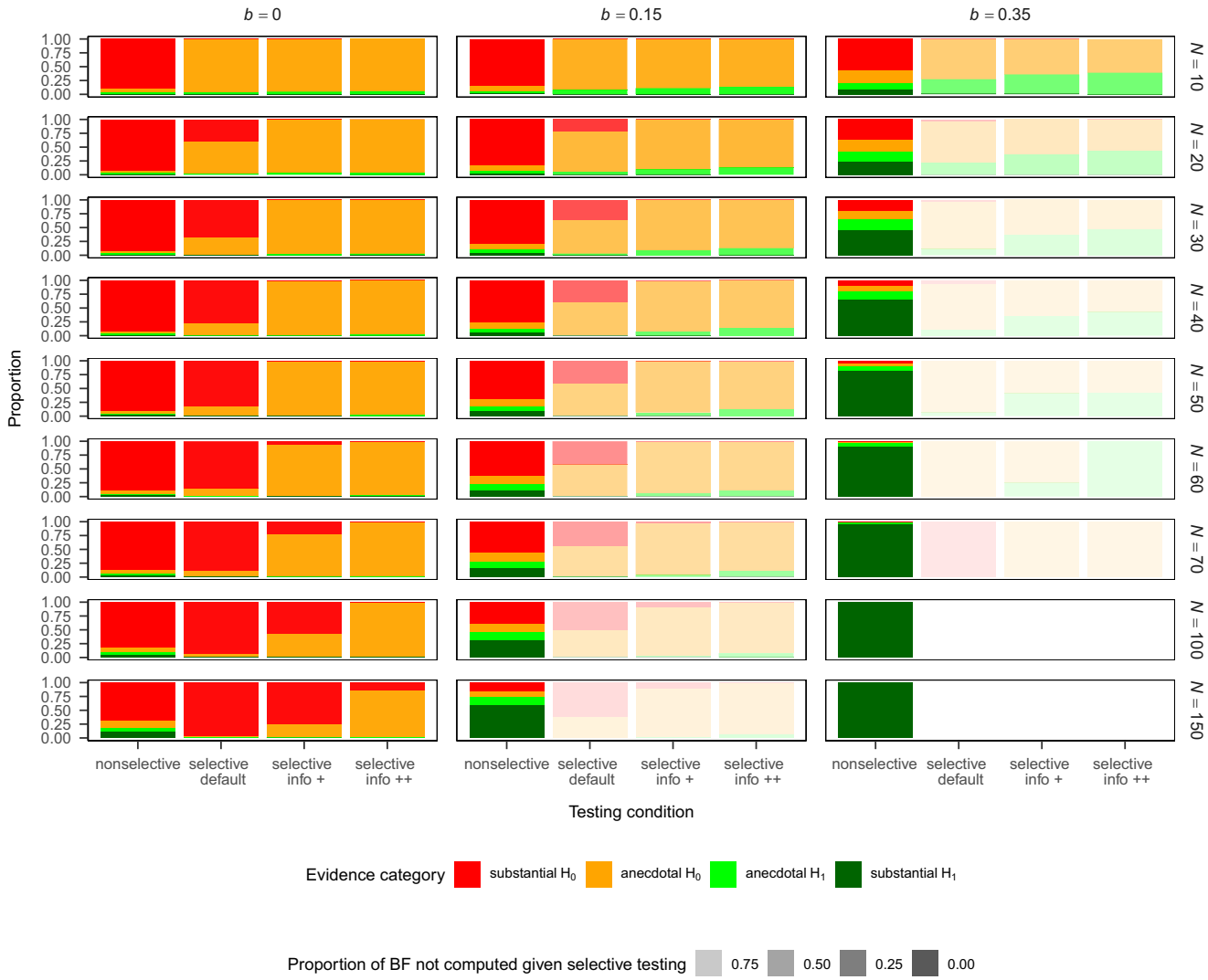
[3] As a robustness check, we also used a different data-generating model in a supplemental simulation. Here, we sampled responses for each cell of the design from a multivariate normal distribution with means of zero, variances of 1, and covariances of 0.5. In simulation conditions with a true population effect, we added data sampled from a normal distribution with $SD = 1$ and $M = 0.5$ or $M = 1.25$, thus implementing an interaction effect (cf. Pfister, 2021). This also roughly corresponded to effect sizes of $\eta^2_G = 0$, $\eta^2_G = .02$, and $\eta^2_G = .08$, respectively. The results largely mirrored the ones from the main simulation, demonstrating the robustness of the results to different data-generating models. We do not report the results here, but the code and results for the additional online materials are available on the Open Science Framework at https://osf.io/jnt6w/.

[4] Note that this is not the default implementation in the BayesFactor package, which is testing each combination of effects against an intercept-only model, with the constraint that an effect including the interaction must include the main effects. Therefore, in the default implementation, the BayesFactor involving the interaction effect would actually compare the full model with an intercept-only model, but would not reflect the specific interaction effect.

[5] Note that the afex package, by default, estimates Type III sums of squares.

Figure 4

Proportion of BFs Indicating Substantial or Anecdotal Evidence for the Null Hypothesis (H_0) or the Alternative Hypothesis (H_1) in Simulation Study 2



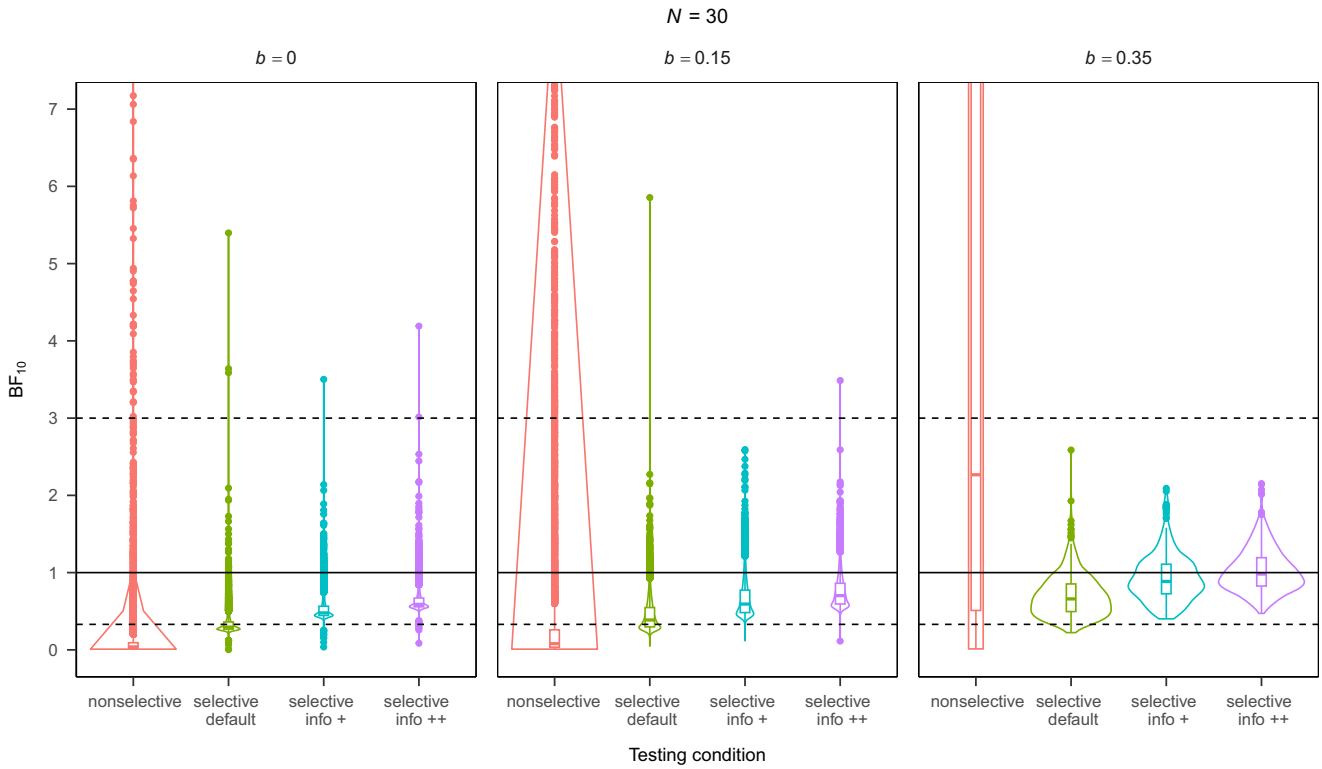
Note. The proportion of BFs that were not computed is only relevant for the selective testing conditions and equals the proportion of frequentist tests reaching significance, therefore being inversely related to the statistical power of the frequentist test. BF = Bayes factor; substantial H_0 = substantial evidence in favor of the null hypothesis ($BF_{10} \leq 0.33$); anecdotal H_0 = anecdotal evidence in favor of the null hypothesis ($0.33 < BF_{10} < 1$); anecdotal H_1 = anecdotal evidence in favor of the alternative hypothesis ($1 < BF_{10} < 3$); substantial H_1 = substantial evidence in favor of the alternative hypothesis ($BF_{10} \geq 3$). See the online article for the color version of this figure.

the selective default testing condition largely mirror the ones of the nonselective testing condition. Using more informative priors leads to more ambiguous evidence, which is more in line with the frequentist interpretation of nonsignificant results. However, if H_0 is true, using the default prior leads to a better representation of the absent population effect.

If H_1 is true (i.e., there is an interaction effect on the population level), in the nonselective testing condition, 75.81% (when $b = 0.15$) and 23.25% (when $b = 0.35$) of Bayes factors indicated substantial or anecdotal evidence in favor of H_0, collapsed across the different sample size conditions, which would indicate evidence that is inconsistent with the actual population effect. The distribution of Bayes factors is also very wide. In the selective default testing condition, 96.64% (when $b = 0.15$) and 75.09% (when $b = 0.35$) of Bayes factors indicated substantial or anecdotal evidence in favor of H_0. These values are severely increased compared to the nonselective testing conditions, even more so for the larger effect ($b = 0.35$), which is also reflected in the Bayes factor distributions being much more compressed compared to the nonselective testing condition. In the selective info+ testing condition, the respective proportions are 91.78% (when $b = 0.15$) and 63.73% (when $b = 0.35$) of Bayes factors, and in the selective info++ testing condition, the respective proportions are 87.43% (when $b = 0.15$) and 59.72% (when $b = 0.35$) of Bayes factors. These values are still increased compared to the

Figure 5

Distribution of BFs (Violin and Box Plots) for a Sample Size of N = 30 in Simulation Study 2



Note. The *y* axis is cut off at the value of 7 for better readability. The maximum BFs were 253 ($b = 0$, occurred in the selective default condition), 2,805 ($b = 0.15$), and 2,170,537 ($b = 0.35$), which all occurred in the nonselective testing condition. The solid horizontal black line indicates a BF of 1, whereas the dashed horizontal lines indicate BFs of 0.33 and 3, which are the values below or above which BFs are typically considered to reflect substantial evidence (Jeffreys, 1961). BF = Bayes factor. See the online article for the color version of this figure.

nonselective testing condition, but less so than in the selective default testing condition. This is again also reflected in an upward shift in the Bayes factor distributions when using more informative priors.

These results again suggest that, given that H_1 is true, selective applications of the Bayes factor lead to Bayes factors that reflect a gross overestimation of the evidence in favor of H_0. Using more informative priors leads to more ambiguous evidence, which is more in line with the frequentist interpretation of nonsignificant results. Therefore, the inferences drawn from Bayes factors computed with more informative priors are more appropriate than the ones drawn from Bayes factors computed with the default prior, given that H_0 is true and Bayes factors are computed selectively, but the evidence in favor of H_0 is still inflated compared to applying the Bayes factor nonselectively. The results of Simulation Study 2 are therefore very similar to the ones of Simulation Study 1.

Transparency and Openness

We follow JARS (Kazak, 2018). All data, analysis code, and research materials have been made publicly available on the Open Science Framework and are accessible at https://osf.io/jnt6w/. Simulations were conducted in R 4.2.2 (R Core Team, 2022) using the package SimDesign (Version 2.14, Chalmers & Adkins, 2020) and (in Simulation Study 2) the package simr (Version 1.0.7, Green & MacLeod, 2016). Bayesian analyses were conducted using the package BayesFactor (Version 0.9.12-4.4, Morey & Rouder, 2022). Frequentist ANOVAs (in Simulation Study 2) were conducted using the package afex (Version 1.3-0, Singmann et al., 2023). Further, the package collection tidyverse (Version 2.0.0, Wickham et al., 2019) and the package ggpubr (Version 0.6.0, Kassambra, 2023) were used for analyses and visualization. We used the packages papaja (Version 0.1.2, Aust & Barth, 2023) and tinylabels (Version 0.2.4, Barth, 2023) for reporting. The simulation studies were not preregistered.

Discussion

Bayes factors are increasingly applied to evaluate hypotheses, particularly because of their ability to quantify evidence in favor of the null hypothesis, which is not possible under a classical null-hypothesis testing framework. However, Bayes factors appear to be also increasingly selectively applied to *specifically* quantify evidence in favor of the null hypothesis given a nonsignificant frequentist test. Here, we showed in two simulation studies, with the examples of independent-samples *t* tests and repeated measures ANOVAs, that this selective application of the Bayes factor can be problematic, leading to a gross overestimation of evidence in favor of the null hypothesis if the alternative hypothesis is actually true, compared to a nonselective application of the Bayes factor

independent of the outcome of a frequentist test. This was particularly the case when relying on default priors. Thus, the distribution of Bayes factors clearly differed between the nonselective testing condition, in which Bayes factors were computed for the full population of cases, and the selective testing condition, in which Bayes factors were only computed for a nonrandom subset of cases. However, given the prevalence of applying the Bayes factor selectively after the outcome of a nonsignificant frequentist test, many researchers appear to (implicitly) assume that the distributions do not differ. The prevalence of problematic situations depends on the power of the frequentist test. A high power reduces the probability of a nonsignificant result given a true effect and, therefore, reduces opportunities for selectively applying the Bayes factor. However, also considering that effects in psychology tend to be small (Szucs & Ioannidis, 2017; see also Götz et al., 2022), even relatively high-powered studies do not fully avoid the problem. A practice of selectively applying Bayes factors given nonsignificant frequentist tests therefore leads to a misrepresentation of the available evidence.

For individual studies, this means that, if Bayes factors are selectively applied, the obtained Bayes factors are biased toward evidence for the null hypothesis, and the obtained evidence appears thus stronger than would be warranted. Therefore, researchers should be less confident in obtained evidence in favor of the null hypothesis and not assume that a Bayes factor provides strong evidence in favor of the null hypothesis if the evidence was obtained as a result of a selective application of the Bayes factor.

Selectively applying Bayes factors may also pose problems for meta-analyses (or drawing inferences across multiple studies more generally), although the magnitude of the problem likely depends on how the meta-analysis is conducted. A meta-analysis that solely draws inferences in a frequentist framework and takes all frequentist tests, be they significant or not, into account, would of course not pose a problem for drawing inferences. If, however, Bayes factors are selectively computed for nonsignificant frequentist test results, and somehow affect the drawn inferences, a problem emerges, as the Bayes factors computed for the nonsignificant frequentist test results would be strongly biased toward the null hypothesis.

For illustration, consider the following example: 10 studies were conducted to test a specific effect. Let us assume that the effect truly exists in the population but may be somewhat smaller than assumed in the power analyses for the studies. In Scenario A, the effect was consistently tested in a frequentist framework with five studies yielding a significant result and five studies yielding a nonsignificant result. In Scenario B, the outcomes of the frequentist testing is the same, but additionally, Bayes factors were selectively computed for the five studies with nonsignificant results, all suggesting decisive evidence *against* the effect (i.e., for the null hypothesis). Researchers' inference regarding the existence of the effect might differ between the two scenarios, such that they may be less inclined to conclude that the effect exists in Scenario B than in Scenario A. In that sense, the accumulation of studies selectively applying the Bayes factor in the literature would result in a kind of reverse file drawer effect. The traditional file drawer effect describes the problem that nonsignificant findings receive zero weight in inferences drawn from the literature, because they are never published. The selective application of the Bayes factor may result in the opposite, a "file exhibition" effect, that is, an increased exposition of evidence in favor of the null hypothesis for nonsignificant findings. Because of such a file exposition effect, these findings may receive increased weight in inferences drawn from the literature. As said before, if a meta-analysis rests on all (non)significant results and inferences are drawn in the frequentist framework alone, there would be no reason to expect biased inferences. But to the extent that Bayes factors somehow infiltrate inferences from such a meta-analysis as well, these inferences might be biased if the Bayes factors themselves were biased.

To avoid these problems, researchers should either specify a priori whether they will test under a frequentist or a Bayesian framework (see also Dienes, 2024) and accept possible drawbacks of this decision (e.g., the inability to tell whether a nonsignificant result reflects the absence of evidence or evidence of absence for the alternative hypothesis under a frequentist framework) or to report Bayes factors and consistently base inferences on Bayes factors, irrespective of the outcomes of frequentist tests (see e.g., Burton et al., 2019; Keysers et al., 2020; Lakens et al., 2020; Pickering et al., 2021). If, for whatever reasons, Bayes factors are selectively applied nevertheless, the use of more informative priors mitigates the problem, as they yield a more conservative test for the null hypothesis.[6] The use of more informative priors leads to more ambiguous evidence, therefore being more in line with the frequentist interpretation of a nonsignificant result (i.e., absence of evidence). As selective applications of the Bayes factor reflect a two-step procedure, the use of more informative priors can also be viewed as taking into account additional information obtained from the first step (i.e., the nonsignificant frequentist test), which would otherwise be ignored. Provided that the more informative prior distribution more accurately represents real effects, the misrepresentation of evidence as a result of selective applications of the Bayes factor is reduced. This is likely, as psychological effects tend to be small (Szucs & Ioannidis, 2017; see also Götz et al., 2022). However, the use of more informative priors still yielded Bayes factor distributions reflecting an overestimation of evidence in favor of the null hypothesis compared to distributions obtained when applying the Bayes factor nonselectively given an actual population effect in the simulation studies. In addition, the use of more informative priors comes at the cost of also receiving Bayes factors reflecting more ambiguous evidence if the null hypothesis is actually true, leading to a worse representation of the (absent) population effect than when using default priors. Thus, the use of more informative priors may mitigate the problem, but it is not a remedy. Therefore, researchers should critically reflect on decisions to selectively apply Bayesian inference after nonsignificant frequentist findings. We recommend that researchers should decide on a testing framework a priori or otherwise report Bayes factors for all tests and consistently use them as the inferential tool. If Bayes factors have to be selectively applied for some reason, the use of more informative priors may mitigate the resulting problems to a certain extent.

Constraints on Generality

While the problem of selectively applying Bayes factors conditionally on a nonsignificant result from a frequentist test should generalize to other situations and models, we limited our

[6] The same applies to the case of nonselective testing. Using a more informative prior distribution would reduce the evidence for the null hypothesis (thus yielding a more conservative test for the null hypothesis) and increase the evidence for the alternative hypothesis (see Dienes, 2024).

simulation studies to the examples of a between-group comparison using an independent-samples *t* test and the test of a 2 × 2 within-subjects interaction using a repeated measures ANOVA, as these are two widespread effects of interest and two analysis methods commonly applied. In addition, we demonstrated the robustness of the simulation results to different kinds of data-generating models with two supplemental simulations we conducted (see Supplemental Materials). However, exploring whether the problem may differ in magnitude across different effects and analysis methods and to what extent the problem persists in more complex models would be valuable. Finally, strategic analysis decisions, such as the application of the Bonferroni correction or other post hoc adjustments, can also increase the likelihood of finding nil (nonsignificant) effects, which may also impact decisions to selectively apply Bayes factors. This is, however, beyond the scope of the current article.

References

Aschenbrenner, A. J., & Balota, D. A. (2019). Additive effects of item-specific and congruency sequence effects in the vocal Stroop task. *Frontiers in Psychology*, *10*, Article 860. https://doi.org/10.3389/fpsyg.2019.00860

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown* (R package Version 0.1.2) [Computer software]. https://github.com/crsh/papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels* (R package Version 0.2.4) [Computer software]. https://cran.r-project.org/package=tinylabels

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, *52*(6), 708–721. https://doi.org/10.1002/bimj.200900299

Benjamini, Y. (2020). Selective inference: The silent killer of replicability. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.fc62b261

Benjamini, Y., & Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 297–318. https://doi.org/10.1111/rssb.12028

Burton, R. L., Lek, I., Dixon, R. A., & Caplan, J. B. (2019). Associative interference in older and younger adults. *Psychology and Aging*, *34*(4), 558–571. https://doi.org/10.1037/pag0000361

Casella, G. (1985). An introduction to Empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87. https://doi.org/10.1080/00031305.1985.10479400

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, Article 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dienes, Z. (2024). Use one system for all results to avoid contradiction: Advice for using significance tests, equivalence tests, and Bayes factors. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(5), 531–534. https://doi.org/10.1037/xhp0001202

Dienes, Z., Coulton, S., & Heather, N. (2018). Using Bayes factors to evaluate evidence for no effect: Examples from the SIPS project. *Addiction*, *113*(2), 240–246. https://doi.org/10.1111/add.14002

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Efron, B. (2008). Microarrays, Empirical Bayes and the two-groups model. *Statistical Science*, *23*(1), 1–22. https://doi.org/10.1214/07-STS236

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., … Hoijtink, H. (2023). A review of applications of the Bayes factor in psychological research. *Psychological Methods*, *28*(3), 558–579. https://doi.org/10.1037/met0000454

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Janczyk, M., Humphreys, G. W., & Sui, J. (2019). The central locus of self-prioritisation. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *72*(5), 1068–1083. https://doi.org/10.1177/1747021818778970

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Clarendon Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kassambra, A. (2023). *ggpubr: 'ggplot2' based publication ready plots* (R package Version 0.6.0) [Computer software]. https://cran.r-project.org/package=ggpubr

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. https://doi.org/10.1038/s41593-020-0660-4

Klaffehn, A. L., Schwarz, K. A., Kunde, W., & Pfister, R. (2018). Similar task-switching performance of real-time strategy and first-person shooter players: Implications for cognitive training. *Journal of Cognitive Enhancement*, *2*(3), 240–258. https://doi.org/10.1007/s41465-018-0066-3

Krueger, J. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist*, *56*(1), 16–26. https://doi.org/10.1037/0003-066X.56.1.16

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, *75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes factors for common designs* (R package Version 0.9.12-4.4) [Computer software]. https://cran.r-project.org/package=BayesFactor

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*(5), 368–378. https://doi.org/10.1016/j.jmp.2011.06.004

Pfister, R. (2021). Variability of Bayes factor estimates in Bayesian analysis of variance. *The Quantitative Methods for Psychology*, *17*(1), 40–45. https://doi.org/10.20982/tqmp.17.1.p040

Pickering, J. S., Henderson, L. M., & Horner, A. J. (2021). Retrieval practice transfer effects for multielement event triplets. *Royal Society Open Science*, *8*(11), Article 201456. https://doi.org/10.1098/rsos.201456

Qiao, L., Zhang, L., Li, H., & Chen, A. (2023). Control transition between cued and voluntary choice tasks: Effects on cognitive flexibility. *Current Psychology*, *42*(17), 14812–14822. https://doi.org/10.1007/s12144-021-02680-w

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org

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. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*(2), 304–321. https://doi.org/10.1037/met0000057

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). *afex: Analysis of factorial experiments* (R package Version 1.2-1) [Computer software]. https://cran.r-project.org/package=afex

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. https://doi.org/10.1016/0001-6918(69)90055-9

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), Article e2000797. https://doi.org/10.1371/journal.pbio.2000797

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(25), 7629–7634. https://doi.org/10.1073/pnas.1507583112

Tsuji, N., & Imaizumi, S. (2022). Sense of agency may not improve recollection and familiarity in recognition memory. *Scientific Reports*, *12*(1), Article 21711. https://doi.org/10.1038/s41598-022-26210-1

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

Weller, L., Pfister, R., & Kunde, W. (2020). Anticipation in sociomotor actions: Similar effects for in- and outgroup interactions. *Acta Psychologica*, *207*, Article 103087. https://doi.org/10.1016/j.actpsy.2020.103087

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Wirth, R., & Kunde, W. (2020). Feature binding contributions to effect monitoring. *Attention, Perception, & Psychophysics*, *82*(6), 3144–3157. https://doi.org/10.3758/s13414-020-02036-9

Received August 14, 2023
Revision received July 8, 2024
Accepted July 30, 2024 ■