Can we Find Evidence for the Null in a Bayesian *t*-Test? Not Unless we Reconsider Bayes Factor Thresholds

Phylactou, P.[1,2], Chen, S.[3], Seminowicz, D. A.[4], & Schabrun, S. M.[1,2]*

[1]School of Physical Therapy, Faculty of Health Sciences, University of Western Ontario, Lodnon, Canada

[2]The Gray Centre for Mobility and Activity, Parkwood Institute, London, Canada

[3]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, USA

[4]Department of Medical Biophysics, Schulich School of Medicine & Dentistry, University of Western Ontario, London, Canada

*Corresponding author: Phivos Phylactou, School of Physical Therapy, Faculty of Health Sciences, University of Western Ontario. Email pphylact@uwo.ca

Preprint version 1, April 14, 2024.

This preprint version has not been peer-reviewed.

Abstract

Within the fields of behavioural and psychological research, the use of Bayesian statistics has gathered increased interest. A statistical test commonly employed in behavioral and psychological research is the *t*-test. For the Bayesian *t*-test, a Bayes Factor (*BF*) can be computed which reflects evidence in favor of either the alternative hypothesis (H_1) or the null hypothesis (H_0). Even though the *BF* is a continuous measure of evidence, it is common to define specific thresholds for accepting the evidence in favor of either the H_1 or the H_0. Such evidence thresholds (e.g., $BF > 3$, $BF > 6$, $BF > 10$) are adopted by related scientific journals to define minimum publication or preregistration requirements. However, exceeding these thresholds is not analogous when H_1 is true compared to when H_0 is true. In turn, this disanalogy might require scientists to invest additional time and resources when H_0 is true, as opposed to when H_1 is true. In this study, we simulated 200 million *BF*s for various effect size, sample size, and variance assumptions, to demonstrate this disanalogy. Further, we show that despite having small shifts in the sample sizes required for exceeding various *BF* thresholds when H_1 is true, when the H_0 is true the probabilities of exceeding a $BF > 6$ or a $BF > 10$ are close to chance. As such, we recommend the use of a $BF > 3$ evidence threshold for the H_0 independently of the evidence threshold set for H_1.

Keywords: Bayes Factors, Bayesian statistics, *t*-test, null hypothesis, *BF* threshold, *BF*

Can we Find Evidence for the Null in a Bayesian T-Test? Not Unless we Reconsider Bayes Factor Thresholds

Following the so-called replication crisis (see Derksen, 2019; Nelson et al., 2018), behavioural and psychological research has increased its reliance on rigour-enhancing methods. Rigour-enhancement includes adopting methods such as pre-registration and methodological transparency (Chambers & Tzavella, 2022; Lin et al., 2024; Nosek et al., 2018), confirmatory testing (Dienes, 2014; Lakens et al., 2020), and appropriate sample size recruitment (Lakens, 2013, 2022). Such rigour-enhancing methods have been shown to increase the replicability of findings from the behavioural and social sciences (Protzko et al., 2023; but see van den Akker et al., 2023)[1].

Within a similar context, many researchers have advocated for the adoption of Bayesian statistics, contrary to the traditional Neyman-Pearson frequentist approach (Neyman & Pearson, 1933), for making inferences and reaching conclusions (Dienes, 2014, 2021; Heck et al., 2023; Kruschke, 2013; Rouder et al., 2009; Schönbrodt et al., 2017; Wagenmakers et al., 2010). One of the arguments in favor of the adoption of Bayesian statistics, relates to the calculation of the Bayes Factor (*BF*). The *BF* has the advantage of quantifying evidence in favor of either of two competing hypotheses, such as the alternative hypothesis (H_1) or the null hypothesis (H_0), as opposed to the conventional approach of using a *p*-value, which can only inform about the rejection (or the failure of the rejection) of H_0 (Johansson, 2011; Wagenmakers, 2007; but see Lakens et al., 2020 for examples of frequentist tests for H_0).

One of the statistical tests frequently utilized in behavioural and psychological research is the *t*-test, which can be used to explore pairwise comparisons. The Bayesian approach for the *t*-test is thoroughly described in earlier work (Fu et al., 2021; Kruschke, 2013; Rouder et al., 2009). For the commonly used Bayesian *t*-test, H_1 is assigned a *prior distribution*, which expresses the anticipated effect size under H_1. In psychology, the prior distribution is most commonly described by a curved distribution, and is usually expressed as a Cauchy (Rouder et al., 2009). Once data are observed a *posterior distribution* can be computed, which represents the uncertainty about the statistical parameter of interest (i.e., the difference between the two

[1] At the time of writing, editors at *Nature Human Behaviour* were investigating criticisms regarding the registration, hypotheses, predictions, and analyses of the Protzko et al. (2023) paper and an editorial response was meant to follow.

samples). Considering that most model comparisons in psychological research involve nested models (i.e., the null hypothesis is a special case of the model, which contains all parameters for the *t*-test; Heck et al., 2023; Wagenmakers et al., 2010), a *BF* can be computed for the Bayesian *t*-test using the Savage-Dickey density ratio (Dickey, 1971).

Put simply, the Savage-Dickey density ratio is a convenient way of computing *BFs* for nested models, by dividing the height of the posterior distribution by the height of the prior distribution at a specific value of interest (Dickey, 1971; see also Heck et al., 2023; Wagenmakers et al., 2010). In the case of a *t*-test, this value of interest will typically concern $\delta = 0$, representing no differences between the two samples. An illustration of how the *BF* is calculated based on the Savage-Dickey density ratio is shown in Figure 1 (see Wagenmakers et al., 2010 for details and mathematical proof). This approach of calculating a *BF* is adopted by widely used software for conducting Bayesian statistics, such as JASP (for examples see Wagenmakers et al., 2018) and JAMOVI.

Figure 1. Illustration of the Savage-Dickey Density Ratio Method of Calculating a Bayes Factor.

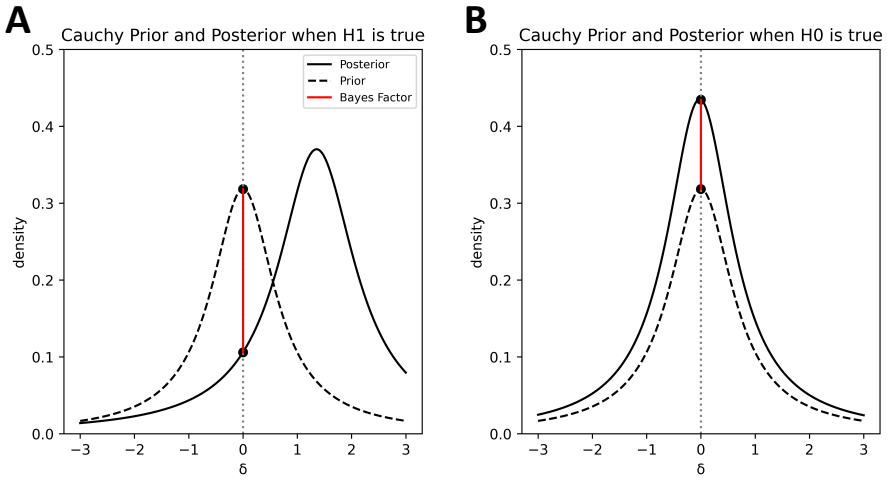


Fig 1. The Savage-Dickey density ratio estimates a Bayes Factor for two competing (nested) models (e.g., H_1 vs. H_0) by dividing the height of the posterior distribution by that of the prior distribution at a specific value of interest (or vice versa). In this example, the value of interest is $\delta = 0$ (i.e., no difference), represented by the gray dotted line. The data are illustrated to show the Bayes Factor **(A)** when H_1 is true ($\delta \neq 0$) and **(B)** when H_0 is true ($\delta = 0$).

Following the Savage-Dickey approach, the resulting *BF* reflects the marginal likelihood of a set of data under H_1 over H_0, often denoted "BF_{10}", or the likelihood of H_0 over H_1, in which

case denoted "BF_{01}". Accordingly, a $BF_{10} > 1$ indicates evidence in support of H_1, while a $BF_{01} > 1$ indicates support for H_0[2]. The *BF* can range from negative infinity to infinity, and the higher the *BF*, the more evidence we have in favor of the respective hypothesis. As such, the *BF* serves as a continuous, updatable, measure of evidence, where additional observations can be collected until the *BF* reflects what one considers adequate evidence in favor of one of the competing hypotheses with confidence (Schönbrodt et al., 2017; Schönbrodt & Wagenmakers, 2018; van Ravenzwaaij & Etz, 2021).

Despite the *BF*'s continuous property, recommendations have been proposed for setting specific evidence thresholds for the *BF.* Common evidence thresholds based on recommendations from Jeffreys (1998; see also Lee & Wagenmakers, 2014), are shown in Table 1. These recommendations are meant to serve as a heuristic to help researchers decide whether to 'accept' that there is adequate evidence to support either H_1 or H_0, once the *BF* reaches or exceeds a specific, predefined threshold. This practice opposes the nature of the *BF* as a continuous measure of evidence, however it may serve an important role in psychological science, for which some argue that dichotomous claims are crucial (see Uygun Tunç et al., 2023). Further, given the *BF*'s updatable property, it can also serve as an indicator when deciding whether adequate sample size has been included, and thus informing researchers if data collection should be continued or stopped (Fu et al., 2021; Schönbrodt et al., 2017).

Table 1. Bayes Factor evidence threshold recommendations (*adapted from Jeffreys, 1998 and Lee & Wagenmakers, 2014*).

Bayes Factor	Log (Bayes Factor)	Interpretation
>1-3	>0-1	Anecdotal evidence
>3-10	>1-2.3	Moderate evidence
>10-30	>2.3-3.4	Strong evidence
>30-100	>3.4-4.6	Very strong evidence
>100	> 4.6	Extreme evidence

Note: Log (Bayes Factor) column added for reference only and illustrates the respective Bayes Factor log transformation.

[2] *BFs* are ratios, where BF_{10} and BF_{01} simply represent the inverse of one over the other (i.e., $BF_{01} = 1/BF_{10}$). As such, both BF_{10} and BF_{01} can be used to quantify evidence for either hypothesis. For example, a $BF_{10} = .5$ indicates that the data are twice as likely to be represented under H_0 than H_1 and is the same as $BF_{01} = 2$. For the purpose of this work, and to make comparisons more intuitive, we use BF_{10} for presenting our results for a simulated true H_1 and a BF_{01} when presenting our results for a simulated true H_0.

At the time of writing, the use of evidence thresholds for *BF*s is widely adopted by many journals relevant to behavioural and psychological research. In detail, many journals define minimum *BF* thresholds, which are required for publication and/or preregistration. The minimum *BF* threshold requirements vary from journal to journal, with common thresholds set at $BF > 3$ (e.g., *Psychology of Consciousness: Theory, Research and Practice*), $BF > 6$ (e.g., *Cortex*, *NeuroImage: Reports*), and $BF > 10$ (e.g., *Nature Human Behaviour*). Notably, this minimum *BF* threshold is expected to be the same for both BF_{10} *and* BF_{01}, which can be problematic, especially in the case of the commonly used *t*-test.

The problem of designating the same threshold to both BF_{10} *and* BF_{01} relates to the representation assigned by the prior distribution to the value of interest. As mentioned previously, in the Bayesian *t*-test, the Savage-Dickey density ratio compares the heights of the posterior and prior distributions at $\delta = 0$. However, $\delta = 0$ happens to be the point where the prior distribution has its highest density. Because of this, the value 0, which also represents the H_0, is assigned a higher probability. As such, as the central tendency of the posterior distribution moves further away from 0, the highest point of the prior distribution is compared to a lower point on the tail of the posterior distribution (Figure 1A). Contrary, when the central tendency of the posterior distribution moves closer to 0, the highest point of the prior distribution is compared to a higher point towards the center of the posterior distribution (Figure 1B). Respectively, obtaining a *BF* when H_1 is true ($\delta \neq 0$) is not analogous to obtaining a *BF* when H_0 is true ($\delta = 0$), since the high density of the prior distribution at 0, will result in a smaller posterior-to-prior height ratio, hence, a smaller *BF*.

The disanalogous nature of *BF*s when H_1 is true compared to when H_0 is true has been mentioned in earlier work (Schönbrodt & Wagenmakers, 2018), and it has been demonstrated in previous simulations (Phylactou et al., 2023; Phylactou & Konstantinou, 2022). This disanalogy has led some researchers to propose the use of 'flexible' or 'negotiable' *BF* thresholds (Weiss, 1997; see also Dienes, 2016; Dienes & Mclatchie, 2018), where, for example, different thresholds are used for H_1 (e.g., $BF_{10} > 6$) and H_0 (e.g., $BF_{01} > 3$). However, to date, the de facto approach adopted by various journals is to set a single, identical, *BF* threshold for both H_1 and H_0. In the context of a *t*-test, reaching a specific threshold can be more time-consuming, resource demanding, and, in turn, more expensive for when H_0 is true, as opposed to when H_1 is true.

Considering the above, the current use of *BF* thresholds needs to be reconsidered. Therefore, the purpose of this simulation study is to demonstrate the disanalogy between reaching an evidence threshold for H_0 compared to H_1 with the Bayesian *t*-test. Further, this study attempts to define a BF_{01} threshold that could serve as comparable to a BF_{10}, to guide researchers, editors, reviewers, and other stakeholders, when designing, conducting, and evaluating research.

Methods

To illustrate the disanalogy between BF_{10} and BF_{01} we run a simulation that was designed based on previous work (Kruschke, 2013; Phylactou et al., 2023; Phylactou & Konstantinou, 2022; Schönbrodt & Wagenmakers, 2018; van Ravenzwaaij & Etz, 2021) and generated 200 million *BFs*: 100 million BF_{10} simulating a true H_1 and 100 million BF_{01} simulating a true H_0. To eradicate confusion with data analysis results, we report simulated *BFs* reflecting a true H_1 as 'BF_{H1}' and *BFs* reflecting a true H_0 as 'BF_{H0}'. Our simulations were implemented in Python (v3.11.5) and *BFs* were computed using the Pingouin package (v0.5.3; Vallat, 2018).

Simulation Procedure

In our simulation, for $i \in \{1, 2, 3, \ldots, 10000\}$ iterations, we generated a standard deviation ($sd_i \sim Uniform[\alpha = 0.3, \beta = 2]$), an effect size that also corresponded to the *t*-test prior distribution Cauchy scale ($r_i \sim Uniform[\alpha = 0.1, \beta = 2]$), and a sample size ($n_i \sim Uniform[\alpha = 5, \beta = 200]$). For the i^{th} iteration, $j \in \{1, 2, 3, \ldots, 10000\}$ iterations where further performed, and a BF_{10} was calculated for when H_1 ($\delta \neq 0$) was simulated to be true, while a BF_{01} was calculated when H_0 ($\delta \neq 0$) was simulated to be true. To simulate a true *H1*, where $\delta \neq 0$, a sample with ν_{ij} observations ($\nu_{ij} = n_i$, let ν be an integer) was drawn from a normal distribution with mean r_i and standard deviation μ_{ij}^{sd1} ($\mu_{ij}^{H1} \sim Normal[r_i, \mu_{ij}^{sd1}]$, where $\mu_{ij}^{sd1} \sim Half\text{-}Normal[sd_i, 0.001]$). A true H_0 ($\delta = 0$) was simulated by generating a sample of zeros for ν_{ij} observations. To conduct a *t*-test and generate a *BF*, a reference sample (serving as the second sample for each *t*-test) was drawn from a normal distribution centered on 0 with a standard deviation μ_{ij}^{sd2} ($\mu_{ij}^{ref} \sim Normal[0, \mu_{ij}^{sd2}]$, where $\mu_{ij}^{sd2} \sim Half\text{-}Normal[sd_i, 0.001]$). A BF_{10} was calculated by conducting a paired *t*-test between the reference distribution and the distribution representing H_1, using a Cauchy prior distribution centered on 0 with a scale r_i. The same Cauchy prior was used to calculate BF_{01} by comparing the distribution representing a true H_0 with the reference distribution. Therefore, each *ij* iteration resulted in two *BF*, one reflecting a true H_1 (from hereon reported as BF_{H1}) and one

reflecting a true H_0 (from hereon reported as BF_{H0}), resulting in 200 million *BFs* from 100 million total iterations. The *BF*s derive from various effect sizes and standard deviations ranging from 0 to 2, and for numerous possible sample sizes ranging from 5 to 200. A visualisation of our simulation is shown in Figure 2.

Figure 2. Simulation pipeline used to generate 200 million Bayes Factors.

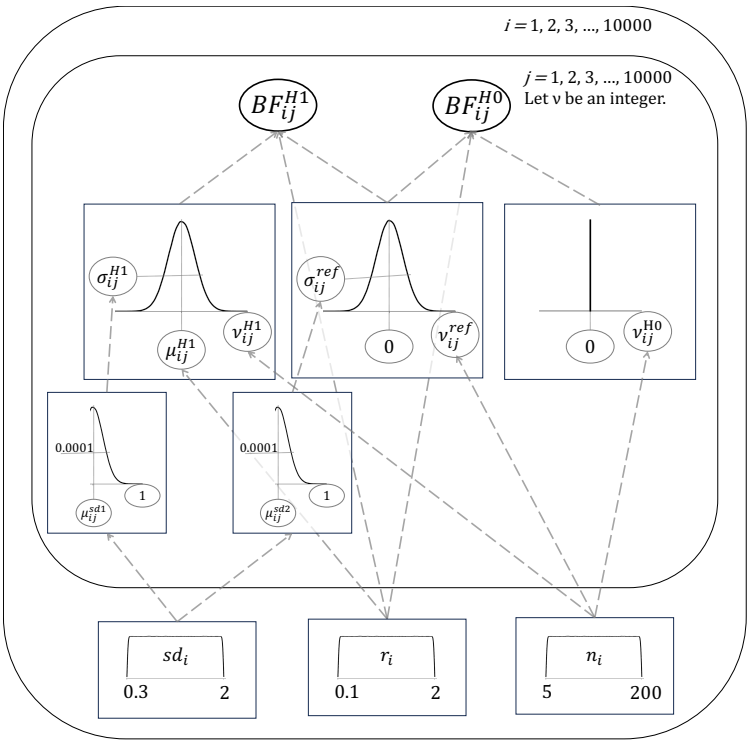


Fig 2. The simulation conducted to generate 200 million Bayes Factors and estimate the probability of surpassing specific thresholds. Specifically, we used uniform distributions to generate a standard deviation ($sd_i \sim Uniform[\alpha = 0.3, \beta = 2]$), an effect size and respectively the prior distribution scale of the *t*-test ($r_i \sim Uniform[\alpha = 0.1, \beta = 2]$), and a sample size ($\nu_i \sim Uniform[\alpha = 0.3, \beta = 2]$), for *i* iterations ($i \in \{1, 2, 3, \ldots, 10000\}$). For each iteration *i*, Bayes Factors were calculated *j* times ($j \in \{1, 2, 3, \ldots, 10000\}$). To simulate a true *H1* ($\delta \neq 0$) a sample with the expected effect size was drawn from a normal distribution ($\mu_{ij}^{H1} \sim Normal[r_i, \mu_{ij}^{sd1}]$, where $\mu_{ij}^{sd1} \sim Half\text{-}Normal[sd_i, 0.001]$), for ν_{ij} observations ($\nu_{ij} = n_i$, let ν be an integer). To simulate a true *H0* ($\delta = 0$), ν_{ij} zeros were generated. To conduct a *t*-test and generate a Bayes Factor, a reference distribution centered on 0 was simulated from a normal distribution ($\mu_{ij}^{ref} \sim Normal[0, \mu_{ij}^{sd2}]$, where $\mu_{ij}^{sd2} \sim Half\text{-}Normal[sd_i, 0.001]$). The distribution for a true H_1 was compared with the reference distribution using a Bayesian *t*-test using a Cauchy centered on 0 and scale r_i (not shown) as a prior distribution, to generate a Bayes Factor. The same Cauchy prior was used to compare the distribution for a true H_0 with the reference distribution and generate a Bayes Factor.

Data Analysis

For our analyses, Bayesian linear regressions were conducted in JASP (v0.18.3, Apple Silicon; https://jasp-stats.org), using a *JZS* prior (see Rouder et al., 2009) with scale $r = .354$ and a binomial model prior (*Model* ~ *Beta* [$\alpha = 1$, $\beta = 1$]). Bayesian ANOVA models were conducted using a *JZS* prior with scale $r = 0.5$ and a uniform model prior (Rouder et al., 2012). Differences between the probabilities were tested using paired *t*-tests, assuming a Cauchy prior distribution centered on 0 with an *r* scale set at $r = 0.5$. A narrow *r* scale was chosen to assign higher density around the null, and thus reflect a stricter criterion for assessing differences (Phylactou et al., 2022). Reported means are accompanied by their standard error (*SE*) and 95% Credible Intervals (*CI*).

For each iteration *i* we calculated the probability of exceeding a *BF* threshold of 3, 6, and 10, separately for BF_{H1} (when H_1 was true) and BF_{H0} (when H_0 was true). These thresholds ($BF > 3$, $BF > 6$, $BF > 10$) were chosen because they are commonly recommended and/or required by relevant scientific journals. Moreover, to enable analyses in a computationally feasible manner, for each iteration *i* we exported the median pooled standard deviation (*sd*), the median $logBF_{H1}$, the median $logBF_{H0}$, and the ratio between the median $logBF_{H1}$ over the median $logBF_{H0}$. We used the logarithmic transformation of the *BF*s as this allowed us to handle extreme *BF* values that derived from iterations with large samples and effect sizes.

The $logBF_{H1}/logBF_{H0}$ ratio was exported to evaluate the analogy between BF_{H1} and BF_{H0}. In detail, a $logBF_{H1}/logBF_{H0}$ ratio equal to one, illustrates that the resulting *BF* is identical for H_1 and H_0 under the same circumstances (same expected effect size, same sample size, same data variability). Moreover, a $logBF_{H1}/logBF_{H0}$ ratio greater than one illustrates that the resulting *BF* is greater for H_1 than H_0 under the same circumstances, and, respectively, a $logBF_{H1}/logBF_{H0}$ ratio smaller than one illustrates that the resulting *BF* is greater for H_0 than H_1. To assess the relationship between sample size, *sd*, and expected effect size on the $logBF_{H1}/logBF_{H0}$ ratio we conducted a Bayesian linear regression. Accordingly, if the relationship between BF_{H1} and BF_{H0} is analogous, then none of the covariates included in the model (sample size, *sd*, expected effect size) should describe the model better than a null model. Alternatively, evidence in favor of a model including any of the covariates would provide support that BF_{H1} and BF_{H0} do not share an analogous relationship.

Additional exploration was conducted after categorizing the data into bins. Specifically, sample sizes were grouped into bins of 10, and expected effect sizes (and respectively *r*) were grouped into bins of 0.2. These analyses were conducted as either one sample *t*-test testing against the value 0.5 (equivalent to a 50% percentage of reaching the given threshold) or as a paired *t*-test. These *t*-tests were described by a half-Cauchy distribution, with a wide scale $r = 1$. The choice of $r = 1$ was motivated by the greater variation that was anticipated in the data, due to the binned sample.

Data Availability

All material used in this study can be accessed through: https://doi.org/10.17605/OSF.IO/RCKF4.

Results

Following our simulations, we estimated the overall mean probability of exceeding a predefined $BF_{H1} > 3$ (*mean* = .791, SE = .003, *95% CI* = [.784, .797]), $BF_{H1} > 6$ (*mean* = .759, SE = .004, *95% CI* = [.752, .765]), and $BF_{H1} > 10$ (*mean* = .737, SE = .004, *95% CI* = [.730, .745]), when H_1 was true. Likewise, we estimated the overall mean probability of exceeding a $BF_{H0} > 3$ (*mean* = .789, SE = .003, *95% CI* = [.784, .795]), $BF_{H0} > 6$ (*mean* = .576, SE = .004, *95% CI* = [.569, .583]), and $BF_{H0} > 10$ (*mean* = .370, SE = .004, *95% CI* = [.363, .377]) for a true H_0. Bayesian paired *t*-tests revealed that the probabilities of exceeding $BF > 3$ were similar between BF_{H1} and BF_{H0} ($logBF_{10} = -3.939$), but different for $BF > 6$ ($logBF_{10} = \infty$) and $BF > 10$ ($logBF_{10} = \infty$). A summary of the overall probabilities is presented in Table 2 and illustrated in Figure 3.

Table 2. Overall probabilities of reaching a predefined Bayes Factor threshold.

Threshold	Mean probability (SE)	Lower 95% CI	Upper 95% CI	*t*-test $log(BF_{10})$
$BF_{H1} > 3$	0.791 (0.003)	0.784	0.797	$logBF_{10} = -3.939$
$BF_{H0} > 3$	0.789 (0.004)	0.784	0.795	
$BF_{H1} > 6$	0.759 (0.004)	0.752	0.765	$logBF_{10} = \infty$
$BF_{H0} > 6$	0.576 (0.003)	0.569	0.583	
$BF_{H1} > 10$	0.737 (0.004)	0.730	0.745	$logBF_{10} = \infty$
$BF_{H0} > 10$	0.370 (0.004)	0.363	0.377	

Figure 3. Overall mean probability of reaching a predefined Bayes Factor threshold.

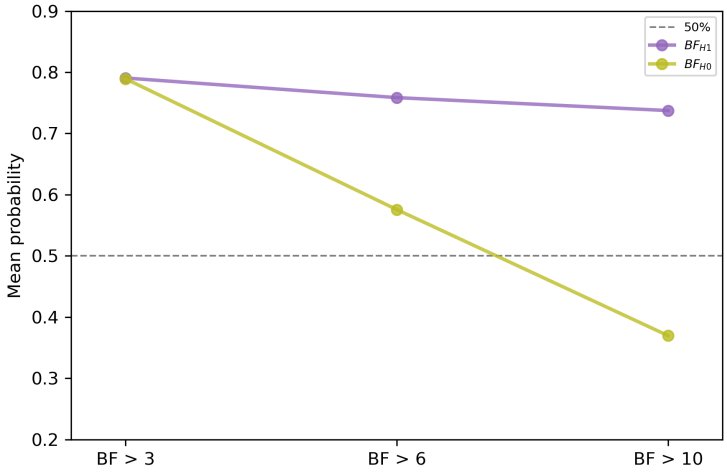


Fig 3. Overall probability of exceeding a predefined Bayes Factor threshold when H_1 is true (purple line) and when H_0 is true (yellow line).

The probabilities of exceeding the predefined *BF* threshold for different effect sizes and sample sizes are illustrated in Figure 4, for both BF_{H1} (Figure 4A) and BF_{H0} (Figure 4B). Heatmaps of the probabilities are provided in *Supplementary Material 1* (Supp. Figure 1). To test the analogy between BF_{H1} and BF_{H0} we fit a linear regression on the $logBF_{H1}/logBF_{H0}$ ratio with the expected effect size (r), the sample size (n), and the standard deviation (sd) as covariates. The model that included all covariates (*r, n, sd*) provided the best fit ($R^2 = .691$, model $logBF = \infty$). Considering the estimated mean of each coefficient, the $logBF_{H1}/logBF_{H0}$ ratio can be expressed as $logBF_{H1}/logBF_{H0} = 9.481 + 10.249(r) + 0.087(n) - 11.973(sd)$. This illustrates that the $logBF_{H1}/logBF_{H0}$ ratio is not analogous, but it increases as the expected effect size and sample size increase, and decreases with greater data variability.

Moreover, we computed two separate regression models, independently for the $logBF_{H1}$ and the $logBF_{H0}$ with *r, n,* and *sd* as covariates. For $logBF_{H1}$, the best fit was provided by a model including all covariates ($R^2 = .648$, model $logBF = \infty$), and can be expressed as $logBF_{H1} = 25.061 + 31.925(r) + 0.262(n) - 30.182(sd)$. For $logBF_{H0}$, the best model describing the data did not include the *sd* covariate ($R^2 = .907$, model $logBF = 6.453$). A comparison between the model including all covariates (*r, n, sd*) and the model including only *r* and *n*, indicated that the model with only *r* and *n* better described the data ($logBF_{r+n/r+n+sd} = 5.153$). As such, the model

describing $logBF_{H0}$ can be expressed as $logBF_{H0} = 2.104 + 1.089(r) + 0.006(n)$. Taken together, the results from the regression models illustrate that under similar circumstances, if H_1 was to be true, evidence in favor of H_1 is likely to be larger than evidence in favor of H_0 if H_0 was to be true.

Figure 4. Probability of exceeding a predefined Bayes Factor threshold for different effect sizes and sample sizes.

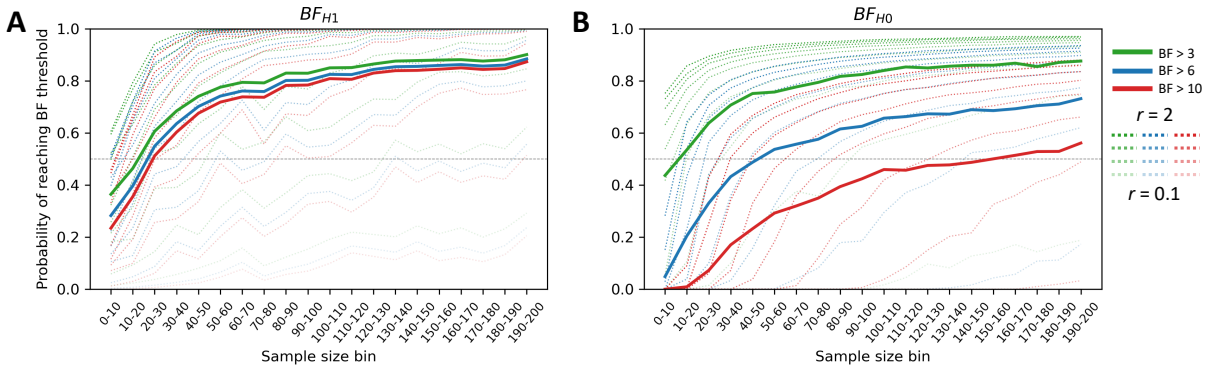


Fig 4. Probability of exceeding a predefined Bayes Factor threshold when **(A)** H_1 is true and when **(B)** H_0 is true, for different sample sizes. Thick lines represent the mean probability across all expected effect sizes (and respectively the *r* scale of the prior distribution for the *t*-test), while the dotted lines show the probabilities across different expected effect sizes. For illustration purposes, sample sizes were grouped into bins of 10 and averaged across all standard deviations.

To further explore whether a true H_1 results in higher *BF*s compared to H_0 we performed additional analyses by averaging across all *sd*s, grouping sample sizes into bins of 10, and grouping expected effect sizes (and respectively *r*) in bins of 0.2. The $logBF_{H1}/logBF_{H0}$ ratios for each effect size bin and a given sample size bin are illustrated in Figure 5A. By employing multiple one sample *t*-tests, we calculated whether the $logBF_{H1}/logBF_{H0}$ ratios were greater than 1, thus, providing evidence for higher *BF*s for a true H_1 than a true H_0 (Figure 5B). Results show that with the exception of low *r*-scales ($r < 0.4$), $logBF_{H1}/logBF_{H0}$ ratios were larger than 1 (*mean* $logBF_{10} = 14.36$, SE = 0.88, *95% CI* = [12.63, 16.09]). Similar results were found when testing whether $logBF_{10}$ were greater than $logBF_{01}$ using paired *t*-tests (*mean* $logBF_{10} = 14.09$, SE = 1.63, *95% CI* = [10.87, 17.30]; Figure 5C). The $logBF_{10}$ for each of these tests individually are provided in *Supplementary Material 2*.

Figure 5. Evidence for higher Bayes Factors when H_1 is true compared to when H_0 is true.

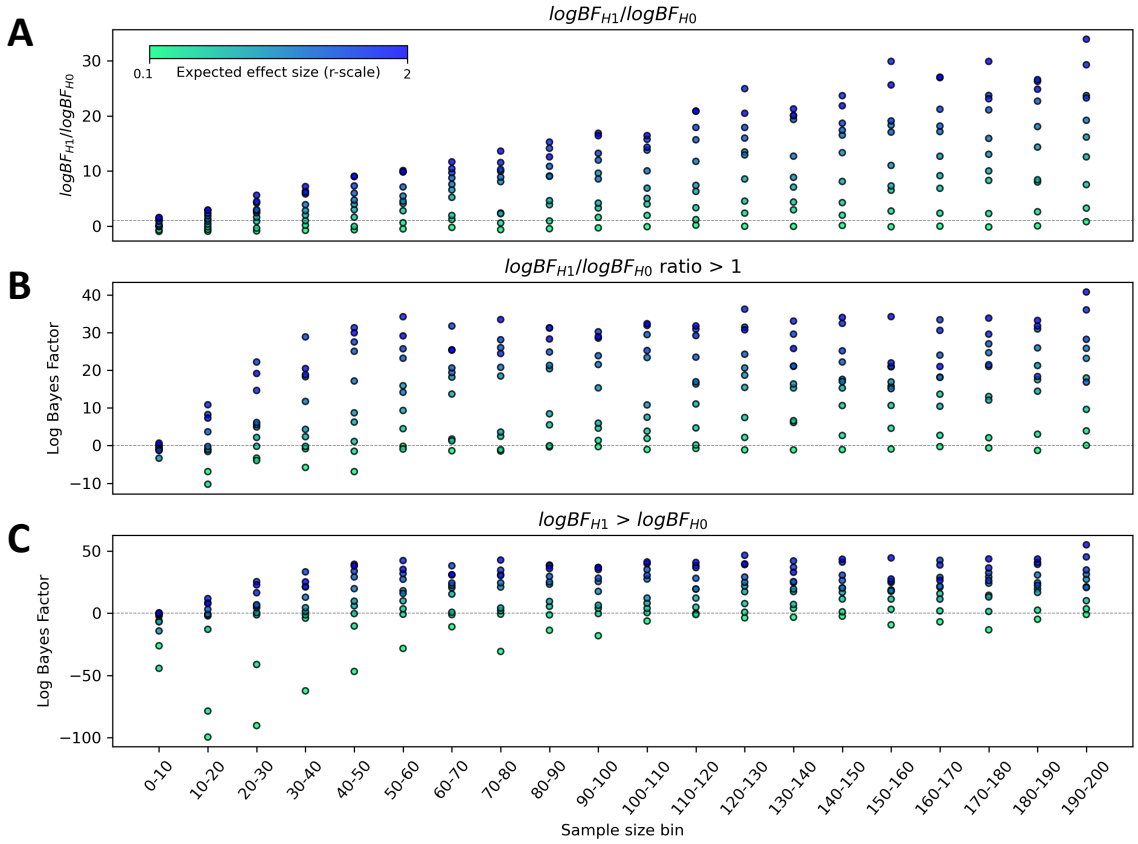


Fig 5. (A) The ratio of $logBF_{H1}$ over $logBF_{H0}$ for different expected effect size and sample size bins (grey dotted line represents a ratio of 1). **(B)** Results from one sample *t*-tests comparing whether $logBF_{10}/logBF_{01}$ is greater than 1 (grey dotted line represents a $logBF_{10} = 0$). **(C)** Results from paired t-tests comparing whether $logBF_{10}$ are larger than $logBF_{01}$ (grey dotted line represents a $logBF_{10} = 0$).

To identify a BF_{H0} threshold comparable to a BF_{H1} given this disanalogy, we explored the conditions under which the *BF* threshold was exceeded in at least 50% of the simulations (Figure 6). The results of the series of one sample *t*-tests, testing whether the *BF*s from each effect size and sample size bin exceed the 50% probability of reaching a predefined threshold are provided in *Supplementary Material 2* and are illustrated in *Supplementary Material 1* (Supp. Figure 2). In addition, heatmaps illustrate that the shift of the probabilities was larger across effect size and sample size bins for the three different *BF* thresholds for BF_{H0} compared to BF_{H1} (Figure 6A; see also *Supplementary Material 2* for the results of each *t*-test individually). This finding was supported by a repeated measures ANOVA conducted on the *logBFs* of each *t*-test, providing evidence for a model including both the hypothesis factor (H_1 vs. H_0) and the threshold factor

($BF > 3$ vs. $BF > 6$ vs. $BF > 10$), as well as their interaction ($R^2 = .74$, model $logBF = 248.29$)[3]. These analyses also demonstrated that even though the $logBF_{H1}/logBF_{H0}$ ratios for $r < .4$ are smaller than 1 (larger BF_{H0} compared to BF_{H1}) for smaller sample sizes (see Figure 5), these *BF*s fail to exceed the predefined threshold in at least 50% of the simulations (see *Supplementary Material 2*; Supp. Figure 2).

To define the comparable BF_{H0} thresholds, we pooled probabilities across the various effect sizes together (Figure 6B). In detail, BF_{H1} reaches above chance (>50%) probability of exceeding the *BF* threshold similarly for thresholds of $BF_{H1} > 3$, $BF_{H1} > 6$, and $BF_{H1} > 10$, with a sample size larger than 20. Conversely, BF_{H0} reaches above chance probability of exceeding a $BF_{H0} > 3$ with a sample size of at least 10, but a $BF_{H0} > 6$ is not reached until a sample size of 40 is surpassed. Notably, a sample size of at least 140 is required, to reach above chance probability of $BF_{H0} > 10$.

Collectively, our results provide evidence for the disanalogous relationship between BF_{H1} and BF_{H0} from a large set of simulated *BFs*, considering various assumptions including different expected effect sizes, sample sizes, and data variability. In summary, our data show that, with equivalent assumptions, under a true H_1, the resulting *BF* will be greater compared to the resulting *BF* under a true H_0. We further illustrate that even though the BF_{H1} will tend to be smaller when data variability increases, BF_{H0} remains unaffected by data variability. Finally, our simulations demonstrate that similar probabilities of exceeding a $BF > 3$ threshold can be expected when either H_1 or H_0 are true, but these probabilities differ when a $BF > 6$ or $BF > 10$ is considered. These differences are driven by the reduced probabilities of reaching the threshold under a true H_0. We next turn to a discussion of our findings.

Discussion

With a simulated sample of 200 million *BF*s, we demonstrate that, for the Bayesian *t*-test, the probability of reaching a predefined *BF* threshold when H_1 is true is not analogous to the probability of reaching a *BF* threshold under a true H_0. Even though this observation was previously noted (Phylactou et al., 2023; Phylactou & Konstantinou, 2022; Schönbrodt & Wagenmakers, 2018), this is, to the best of our knowledge, the first attempt to explicitly

[3] For the computation of this analysis $logBF = -\infty$ values were replaced with the minimum *logBF*, and $logBF = \infty$ were replaced with the maximum *logBF*.

investigate the relationship between BF_{H1} *and* BF_{H0} for different sample sizes, effect sizes, and data variances. Given the adoption of specific *BF* threshold requirements by scientific journals for publication and/or preregistration, our findings have important implications for the use of Bayesian statistics within the behavioural and psychological fields.

Figure 6. Above chance probability of exceeded the Bayes Factor Threshold.

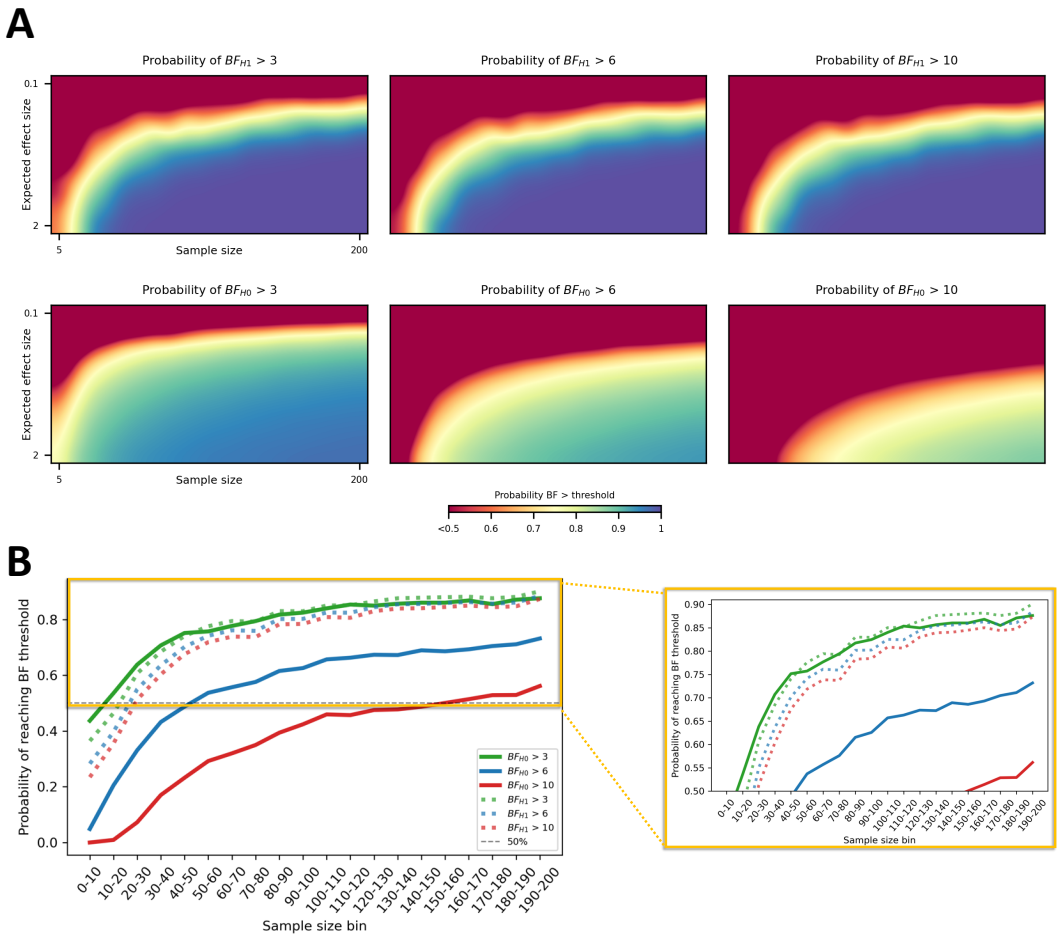


Fig 6. (A) Heatmaps illustrating the estimated probability of exceeding the predefined Bayes Factor threshold in at least 50% of the simulations, for various effects and samples. **(B)** Sample size required to exceed the Bayes Factor threshold in at least 50% of the simulations pooled across all expected effect sizes.

Even though the use of thresholds contradicts the property of the *BF* as a continuous measure of evidence, their use may benefit behavioural and psychological research. For example, aiming for a specific *BF* threshold can help researchers inform their sampling plan (e.g., sample updating; Fu et al., 2021; Schönbrodt & Wagenmakers, 2018). Further, as mentioned earlier, the use of these thresholds may also be necessary to adhere to specific scientific journal guidelines.

However, the current use of the *BF* threshold (i.e., same threshold for H_1 and H_0) may result in spending additional time and resources, if the H_0 under investigation is in reality true.

Researchers, reviewers, editors, and other stakeholders should consider this disanalogy when designing and evaluating research. As an example, consider a hypothetical scenario, where a research team is planning an experiment for an expected difference between two samples of a magnitude of 0.4, with enough resources to recruit up to 60 participants. According to our simulations, if H_1 turns out to be true, the experiment will have above chance probability of exceeding a $BF > 10$ threshold but will only exceed a $BF > 3$ if H_0 is true. Further, even with the maximum sample size simulated here ($n = 200$), if H_0 is true, the $BF > 10$ threshold will not be reached. In such a case, the research team will be able to target a scientific journal that requires a $BF > 10$ threshold if H_1 is true, but not if H_0 is true. In addition, if the research team is targeting the specific journal, they might refrain from preregistration, due to the risk of not reaching the threshold in the case of a true H_0. As such, we recommend that this disanalogy between BF_{H1} *and* BF_{H0} is taken into consideration in cases where *BF* thresholds are required.

Our findings echo previous recommendations and provide evidence from a large simulated dataset showcasing the necessity of considering flexible evidence thresholds between BF_{H1} *and* BF_{H0} (Weiss, 1997; see also Dienes, 2016; Dienes & Mclatchie, 2018). Following our results, we recommend the use of a $BF_{01} = 3$ ($BF_{10} = 1/3$) as an adequate threshold of evidence for accepting the H_0 independently of the evidence threshold used to accept H_1. Even though the probability of exceeding higher thresholds ($BF > 6$, $BF > 10$) for a true H_0 remains low (see Figure 6) scientists may wish to consider a threshold of $BF > 6$ for sample sizes greater than 40, or a threshold of $BF > 10$ for sample sizes greater than 140.

Our recommendations should be considered contiguous with our simulation study's limitations. First, it must be noted that our simulated *BF*s are limited to the values of the parameters (*r, n, sd*) used to inform the simulation (Figure 2). However, the linear relationship between BF_{H1} *and* BF_{H0} as demonstrated by our regression models, indicate that the disanalogy can likely be expected for different (i.e., higher) values of the parameters. Future work may seek to investigate if and when this disanalogy reaches a plateau. Relatedly, another limitation of the simulation relates to the choice of using the same value for the expected effect size and the *r* scale of the Cauchy prior. While this only affects the estimation of BF_{H1} (since for BF_{H0} the effect size is always 0), it may lead to a less accurate estimate of the ratio between BF_{H1} *and*

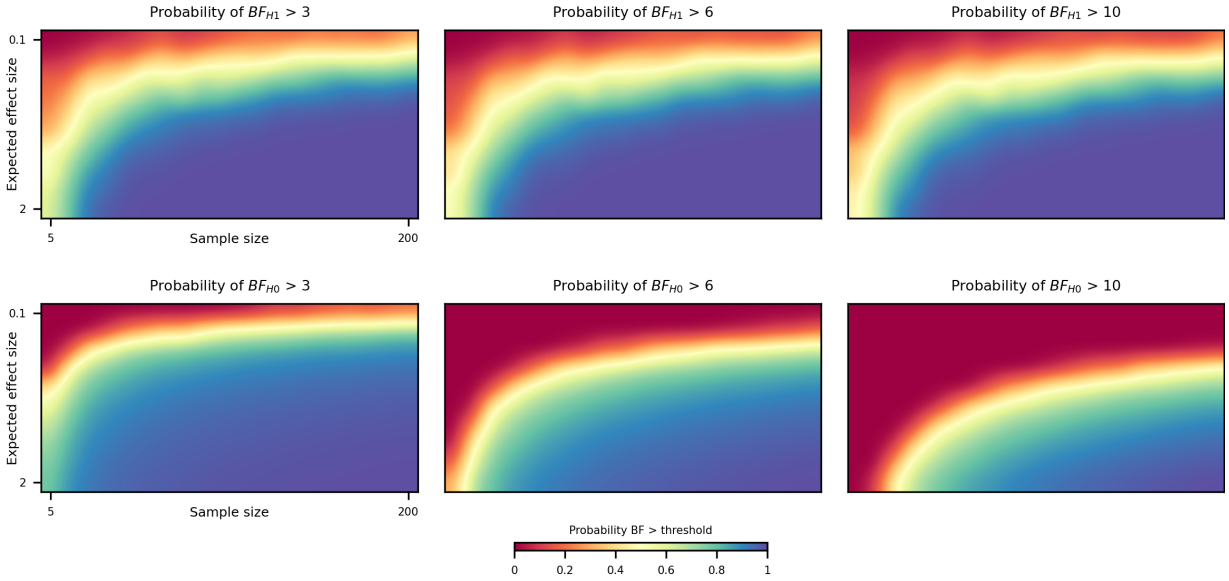
BF_{H0}. Though, given that this approach does not affect BF_{H0}, our recommendations for the use of a $BF_{H0} > 3$ as an adequate evidence threshold is not considered subject to this limitation.

A final limitation of the simulation is related to the use of the paired *t*-test to estimate the *BF*s. As such, it can be argued that our recommendations may not be translatable to an independent, a one sample, or any one-sided (directional) *t*-test. While it is likely that the ratio between BF_{H1} *and* BF_{H0} for other forms of the *t*-test will differ, the underlaying approach of calculating a *BF* remains susceptible to the same limitation of the Savage-Dickey density ratio (Figure 1). Further, the *t*-test used in our simulation can be considered equivalent of a two-sided one-sample *t*-test (i.e., testing a sample against the value 0; Phylactou et al., 2022), and an independent *t*-test assuming equal group sample size and variance (although in this case *n* represents the size per group and not the total sample size; Phylactou & Konstantinou, 2022). In terms of one-sided tests, similar results to ours can be anticipated when H_0 represents no difference (i.e., $H_0 = 0$), but the BF_{H1} to BF_{H0} ratio will most likely differ when H_0 represents a directional effect (e.g., $H_0 < 0$ or $H_0 > 0$). Subsequent work may aim to validate our findings for different forms of the *t*-tests.

In conclusion, our study uses a large sample of simulated data to provide evidence in support of the use of flexible *BF* thresholds, when such thresholds are implemented. We demonstrate that, for the *t*-test, collecting Bayesian evidence, as reflected through the *BF*, when H_0 is true is not analogous to collecting Bayesian evidence when H_1 is true. Because the probabilities of reaching a $BF_{H0} > 6$ or a $BF_{H0} > 10$ are slightly above chance, we propose that, independently of the evidence threshold used for H_1, a *BF* threshold of $BF_{H0} > 3$ is considered adequate by scientists within the field of behavioural and psychological research for H_0.

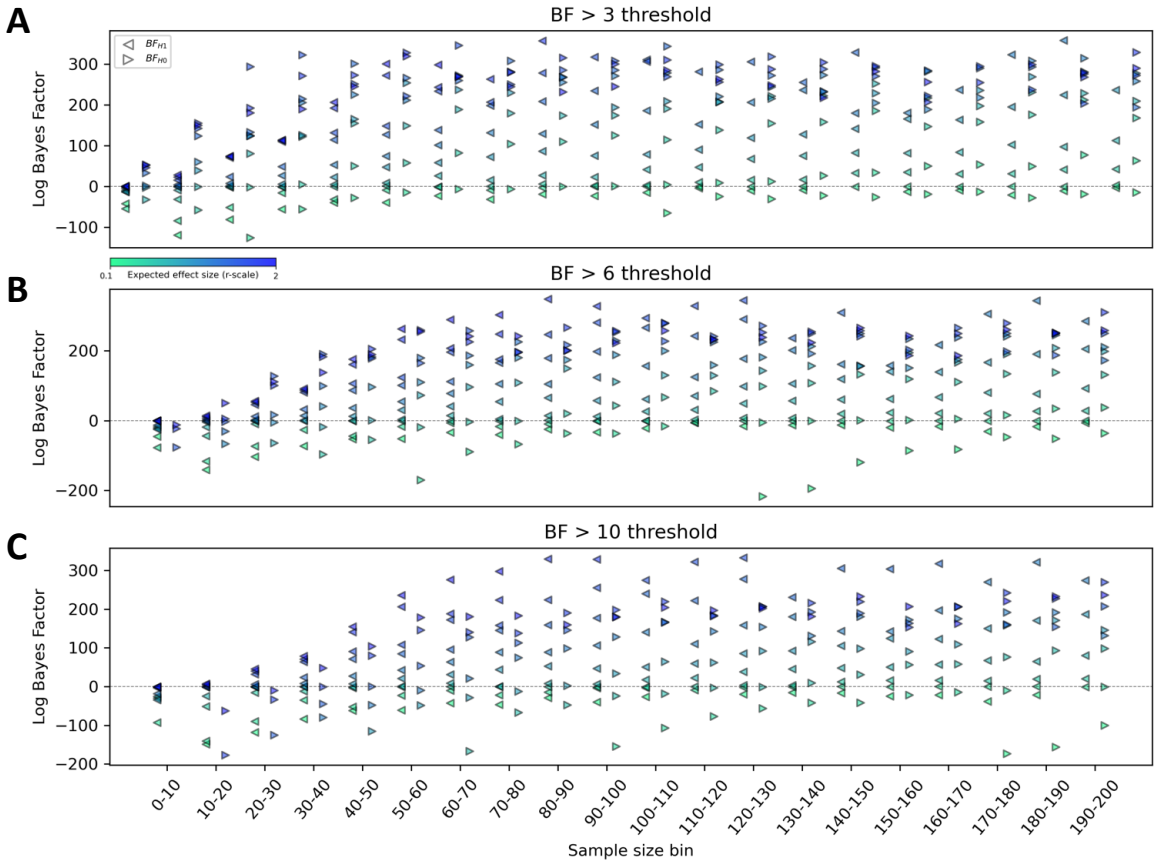
Supplementary Material 1

Supplementary Figure 1. Probabilities of reaching a predefined Bayes Factor threshold for different effect sizes and sample sizes.



Supp. Fig 1. Probability of exceeding a predefined Bayes Factor (*BF*) threshold when H_1 is true and when H_0 is true, for different sample sizes and expected effect sizes, for three different thresholds ($BF > 3$, $BF > 6$, $BF > 10$). For illustration purposes, sample sizes were grouped into bins of 10, and effect sizes into bins of 0.2.

Supplementary Figure 2. Evidence of exceeding the Bayes Factor threshold in at least 50% of the simulations.



Supp. Fig 2. Evidence of exceeding a Bayes Factor (*BF*) threshold of **(A)** 3, **(B)** 6, and **(C)** 10, in at least 50% of the simulations when H_1 is true and when H_0 is true, for different sample sizes and effect sizes (grey dotted line represents a $logBF_{10} = 0$). *Note.* Missing values are *log(Bayes Factor)* values which equal to infinity (positive or negative). The exact values for each test are provided in Supplementary Material 2.