Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-023-02295-1

THEORETICAL/REVIEW

# Investigating the relationship between the Bayes factor and the separation of credible intervals

**Zhengxiao Wei**[1] · **Farouk S. Nathoo**[1] · **Michael E. J. Masson**[2]

Accepted: 16 April 2023


## Abstract

We examined the relationship between the Bayes factor and the separation of credible intervals in between- and within-subject designs under a range of effect and sample sizes. For the within-subject case, we considered five intervals: (1) the within-subject confidence interval of Loftus and Masson (1994); (2) the within-subject Bayesian interval developed by Nathoo et al. (2018), whose derivation conditions on estimated random effects; (3) and (4) two modifications of (2) based on a proposal by Heck (2019) to allow for shrinkage and account for uncertainty in the estimation of random effects; and (5) the standard Bayesian highest-density interval. We derived and observed through simulations a clear and consistent relationship between the Bayes factor and the separation of credible intervals. Remarkably, for a given sample size, this relationship is described well by a simple quadratic exponential curve and is most precise in case (4). In contrast, interval (5) is relatively wide due to between-subjects variability and is likely to obscure effects when used in within-subject designs, rendering its relationship with the Bayes factor unclear in that case. We discuss how the separation percentage of (4), combined with knowledge of the sample size, could provide evidence in support of either a null or an alternative hypothesis. We also present a case study with example data and provide an **R** package '*rmBayes*' to enable computation of each of the within-subject credible intervals investigated here using a number of possible prior distributions.

**Keywords** Bayes factor · Credible intervals · Within-subject designs · Within-subject intervals

## Introduction

Phenomena such as *p*-hacking and the file-drawer effect are important issues in science and are associated with an over-reliance on null-hypothesis significance testing (NHST) in experimental psychology (Hu et al., 2016; Kline, 2013) and other areas. Even established researchers can sometimes misinterpret *p*-values and can, as a result, fail to replicate their studies and properly interpret their results (Etz & Vandekerckhove, 2016). Seeking alternatives to the *p*-value approach, researchers have lately advocated alternative statistical methods to assess the strength of the evidence for the presence of an effect of interest. These alternatives include interval estimation (e.g., Cumming, 2014; Eich, 2014; Heck, 2019; Loftus & Masson, 1994; Nathoo et al., 2018; Wagenmakers et al., 2022) and the Bayes factor (e.g., Kass & Raftery, 1995; Masson, 2011; Rouder et al., 2012; Rouder et al., 2017; Wagenmakers et al., 2010).

Although the relationship between confidence intervals and NHST is established and well understood in between- or within-subject designs (Franz & Loftus, 2012; Loftus & Masson, 1994; Schenker & Gentleman, 2001), the analogous relationship between credible intervals and the Bayes factor is not fully understood and is the topic of this article. Focusing on the linear mixed-effects model, we delineate this relationship and explore how it varies for different types of interval estimates and under a range of simulated effect sizes and sample sizes. Our study considers five interval estimates, and these are detailed in Table 1: (1) the within-subject confidence interval of Loftus and Masson (1994; LM-CI); (2) the analogous within-subject Bayesian interval developed by Nathoo et al. (2018; NKM-HDI), whose derivation conditions on estimated random effects; (3) and (4) two modifications of (2) based on a proposal by Heck (2019; LH-HDI and JZS-HDI) to allow

✉ Zhengxiao Wei
zhengxiao@uvic.ca

[1] Department of Mathematics and Statistics, University of Victoria, P.O. Box 1700 STN CSC, Victoria, British Columbia V8W 2Y2, Canada

[2] Department of Psychology, University of Victoria, Victoria, British Columbia, Canada

Published online: 11 May 2023

**Table 1** Overview of the interval estimates for population means

| Label | Equation | Description |
|---|---|---|
| CI | $M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{W}}}{n(n-1)a}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)}$ | Standard confidence interval |
| HDI | Markov chain Monte Carlo sampling of $\mu_i$ | Standard highest-density interval |
| LM-CI | $M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{S\times C}}}{n(n-1)(a-1)}} \cdot t^*_{1-\frac{\alpha}{2},\ (n-1)(a-1)}$ | Within-subject CI |
| NKM-HDI | $M_{i\cdot} \pm \sqrt{\frac{SS_{\mathrm{S\times C}}}{n(n-1)a}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)}$ | Conditional within-subject HDI |
| LH- or JZS-HDI | $\mathbb{E}\left[\mu_i \pm \frac{\sigma_\epsilon}{\sqrt{n}} \cdot t^*_{1-\frac{\alpha}{2},\ a(n-1)} \mid \mathrm{Data}\right]$ | Modification of NKM-HDI |

*Note*. A linear mixed-effects model is $Y_{ij} = \mu_i + b_j + \epsilon_{ij}$, $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$, for $i = 1, \cdots, a$ and $= 1, \cdots, n$; $\mu_i$ is the mean of the responses at the $i$th level; $b_j$ is a mean-zero random effect for the $j$th subject. The sample mean for the $i$th condition is $M_{i\cdot}$. The within-group sum-of-squares (SS) is $SS_{\mathrm{W}}$. The interaction SS is $SS_{\mathrm{S\times C}}$. $t^*$ refers to a critical value for the $t$-distribution

for shrinkage and account for uncertainty in the estimation of random effects; and (5) the standard Bayesian highest-density interval (HDI). The within-subject HDI labeled as JZS-HDI assumes the same prior distribution as that used to compute the default Bayes factor for analysis of variance (ANOVA) designs in Rouder et al. (2012). In the case of within-subject intervals, the intervals reflect uncertainty about the relative magnitudes of the population means rather than their absolute values.

In either a balanced between-subjects design or a within-subject design using LM-CI, the point estimates of two population means have a statistically significant difference if they are separated by at least $\sqrt{2}$ times the width of the associated frequentist interval estimate (Loftus & Masson, 1994, p. 482). Note that this multiplier is constant over sample sizes. The *width* of an interval is, by convention, the half-length from the lower bound to the upper bound of the interval. Using the pooled estimate for the standard error of the mean to compute the confidence interval when homogeneity of variance holds ensures that all conditions will have the same interval width. We calculated the *separation percentage* for the interval estimates of two population means as the absolute difference between two interval centers over the twofold interval width. Hence, 50% separation indicates that the smaller condition mean touches the lower boundary of the larger mean's interval, and 100% separation indicates that the upper bound of the smaller mean's interval touches the lower bound of the interval for the larger mean (see Fig. 1). If the intervals do not overlap at all, separation exceeds 100%. For example, if the upper bound for the smaller mean's interval is separated from the lower bound of the larger mean by a distance equal to the interval's width, then the separation percent is 150%. When it comes to unequal interval widths, we propose a more general definition of separation percent, in which the denominator of the formula is the root mean square of the two interval lengths, $l = \sqrt{\left(l_1^2 + l_2^2\right)/2}$.[1] For example, if two midpoints are one unit apart, and their interval lengths are two units and one unit, then the separation percentage is calculated as $1/\sqrt{\left(2^2 + 1^2\right)/2} \approx 63\%$ (see Fig. 1). Note that if this formula is applied to cases with equal interval widths, the denominator will reduce to the simple twofold interval width.

The Bayes factor $\mathrm{BF}_{10} = \frac{p(\mathrm{Data} \mid \mathcal{M}_1)}{p(\mathrm{Data} \mid \mathcal{M}_0)}$ is the ratio of the marginal likelihood of the observed data under the alternative model (that includes an experimental effect) to the marginal likelihood of the same data under the null model (that assumes no such effect). Bayesian approaches have advantages over NHST, which include (1) strict dependence on the observed data and not on hypothetical replicate data, (2) immunity to data collection practices such as optional stopping, and (3) quantification of statistical evidence taking both the null and alternative hypotheses into account, including the possibility of assessing the strength of evidence in favor of the null (Dienes, 2021; Nathoo & Masson, 2016; Wagenmakers, 2007). Rouder et al. (2012) developed a suite of methods for computing Bayes factors for ANOVA designs by assuming the Jeffreys prior for the overall mean and residual variance (Jeffreys, 1946), a $g$-prior structure for effects (Zellner & Siow, 1980), and independent scaled inverse-chi-square priors with one degree of freedom for the scale

[1] Other types of averages can also be used. The root mean square is preferable because it connects the pooled confidence interval width ($l = t^*_{1-\frac{\alpha}{2},\ n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$) for the difference between means in a two-sample $t$-test to the confidence interval widths ($l_i = t^*_{1-\frac{\alpha}{2},\ n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_i}}$) for the population means in an unbalanced one-way ANOVA with two conditions. $l^2 = l_1^2 + l_2^2$, and $s_p$ is the pooled estimate of the common standard deviation.

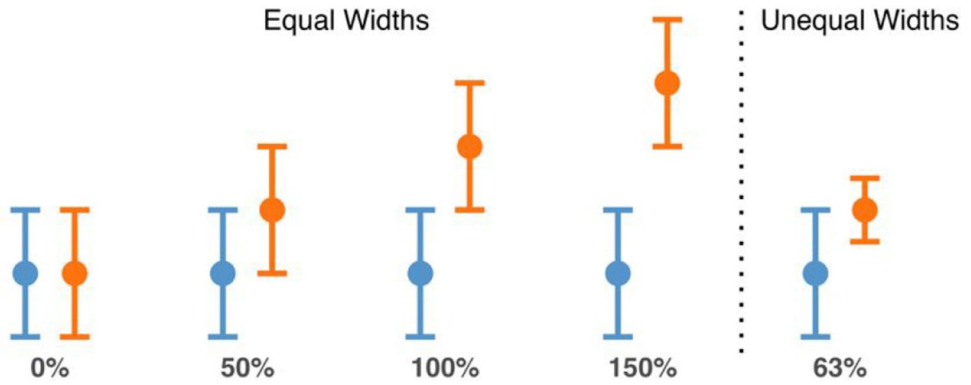


**Fig. 1** Examples of the separation percentage for the interval estimates of two population means. A general definition of the separation percentage of any two intervals is given as the distance between two interval centers divided by the root mean square of interval lengths

hyperparameters of the *g*-priors. These hierarchical specifications are widely adopted default priors for Bayesian model selection because of the consistency and computational efficiency in evaluating marginal likelihoods (Liang et al., 2008). Although the Bayes factor may not have a closed-form solution for linear models, its multi-dimensional integral computation can be fulfilled by calling a suite of functions in the '*BayesFactor*' **R** package by Morey and Rouder (2022).

Lovric (2020) investigated the relationship between decision rules based on the Bayes factor and decision rules based on credible intervals. That study compared only the operating characteristics of decision rules but did not consider any specific functional relationship between the two quantities. Within the context of testing a single normal mean with known variance, Lovric demonstrated settings where the decision rules conflict. A limitation of this study is its formulation of the comparison based on the outcome of decision rules and summary in terms of agreement or disagreement of decision rules based on the two approaches. Similarly, within the paradigm of equivalence testing (where the null hypothesis is non-equivalence), Linde et al. (2021) compared the classification performance of three approaches, namely, (1) two one-sided tests, (2) HDI-region of practical equivalence (Kruschke, 2018), and (3) interval Bayes factor, in a balanced independent-samples *t*-test, with respect to type I error rate ($\alpha$) and statistical power. Linde et al. proposed using the interval Bayes factor for finding evidence of equivalence, especially when the sample size is inadequate. However, Campbell and Gustafson (2021) argued that these three approaches would produce the same performance when calibrated to have the same predetermined maximum $\alpha$.

Our simulation study, conducted using a one-way linear model and one-way linear mixed model, investigates the relationship between the separation of credible intervals and Bayes factors for testing equality of the corresponding parameters (population means) associated with two conditions and a fixed sample size. A primary contribution of this article is thus to report the discovery of a remarkably simple quadratic exponential relationship between the two quantities, which holds for a given sample size. To our knowledge, no such relationship has been previously reported in the literature. A secondary contribution is the development of user-friendly and comprehensive software for the computation of Bayesian interval estimates for within-subject designs.

The rest of the paper proceeds as follows. We articulate the assumptions and prior distributions for the corresponding statistical models for both between- and within-subject designs. In between-subjects designs, we derive the analytic form for the function relating the separation of confidence intervals to the Bayes factor through a limit theorem. This relationship is for a fixed sample size, and the approximation becomes more accurate as the sample size increases. In the within-subject design, we discuss Table 1's methods of constructing the interval estimates for population means and introduce an **R** package to enable computation of these estimates. Next, we describe a series of Monte Carlo simulation studies and present the results across various parameter settings along with quadratic exponential curve fits. Our limit theorem for between-subjects designs suggests that the relationship observed in our simulation studies for within-subject designs is an asymptotic one. We also examine results under a range of sample sizes and provide benchmarks (1) to evaluate the Bayes factor when two JZS-HDIs barely overlap and (2) to evaluate the separation between intervals for two condition means computed using JZS-HDI when the Bayes factor presents moderate evidence for an effect. We extend the designs to multilevel and multiway ANOVA with application demonstrations on experimental data. Finally, we conclude the paper with overall recommendations and a discussion of limitations and future work.

## Between-subjects design

Consider a linear model (1) for the mean response in a one-way between-subjects design

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_\epsilon^2\right), \\ \text{for } i = 1, \cdots, a \text{ and } j = 1, \cdots, n_i, \tag{1}$$

where $Y_{ij}$ represents the response for the $j$th subject under the $i$th level of the experimental manipulation; $\mu$ is the overall mean, $\tau_i$ is the $i$th level of the experimental manipulation ($\mu_i = \mu + \tau_i$ for the means model, where $\mu_i$ is the $i$th population mean); $a$ is the number of levels; $n_i$ is the number of subjects in the $i$th group. The classical ANOVA, as well as the Bayes factor approach, tests the null hypothesis $\mathcal{M}_0 : \mu_1 = \cdots = \mu_a$ versus the alternative hypothesis $\mathcal{M}_1$ that at least one mean is different. When considering the overlap in the interval estimates of two of the population means, $\mu_p$ and $\mu_q$ at levels $p$ and $q$, respectively, the implicit null and alternative hypotheses are $\mathcal{M}_0' : \mu_p = \mu_q$ versus $\mathcal{M}_1' : \mu_p \neq \mu_q$. That is, consideration of the overlap is based on a simpler pair of models in which just two population means are involved. These two sets of null and alternative hypotheses are the same when $a = 2$.

It is common to re-parameterize effect size as $t_i = \tau_i / \sigma_\epsilon$ so that the treatment effects are standardized relative to the standard deviation of the error and become dimensionless (Jeffreys, 1961; Rouder et al., 2012, p. 359). A Jeffreys prior on $\mu$ and $\sigma_\epsilon^2$ for both models is

$$\pi\left(\mu, \sigma_\epsilon^2\right) \propto \frac{1}{\sigma_\epsilon^2}. \tag{2}$$

In the Bayesian context, we assume that

$$t_i \mid g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g), \tag{3}$$

so that the resulting Bayes estimators will exhibit a data-dependent shrinkage towards zero (the posterior estimate is shifted from the sample mean towards the prior mean; Armitage et al., 2002). Aside from the specific form adopted here, a general class of shrinkage priors can arise based on a Gaussian scale-mixture formulation through different choices of the distribution for $g$ (Carvalho et al., 2010; Casella et al., 2010). When modeling fixed effects (i.e., $t_i$ is assumed to be constant for all trials), Rouder et al. (2012) proposed default priors by projecting a set of $a$ condition effects into $a-1$ parameters, with the property that the marginal prior on all $a$ effects is identical, such that

$$t_i^\star \mid g \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, g), \tag{4}$$

$$\left(t_1^\star, \cdots, t_{a-1}^\star\right) = \left(t_1, \cdots, t_a\right) \cdot \mathbf{Q}, \tag{5}$$

and

$$\mathbf{I}_a - a^{-1}\mathbf{J}_a = \mathbf{Q} \cdot \mathbf{Q}^\top, \tag{6}$$

where $\mathbf{I}_a$ is the identity matrix of size $a \times a$, $\mathbf{J}_a$ is the all-ones matrix of size $a \times a$, $\mathbf{Q}$ is an $a \times (a-1)$ matrix of the $a-1$ eigenvectors of unit length corresponding to the nonzero eigenvalues of the left-side term in Equation 6, and $(t_1, \cdots, t_a)$ is a row vector. For example, $t_1^\star = \frac{\sqrt{2}}{2}\left(t_1 - t_2\right)$ when $a = 2$. Based on Zellner and Siow (1980), $g$-priors assume that

$$g \sim \text{Scale-inv-}\chi^2\left(1, h^2\right), \tag{7}$$

where $h$ is a tuning parameter that specifies *a priori* the expected range of effect sizes. In our study, we used a scale fixed effect $h = 0.5$ and a scale random effect $h = 1$. Rouder et al. (2012, p. 363) concluded that consideration of fixed or random effects was not critical for balanced one-way designs ($n_1 = \cdots = n_a = n$), and the resulting Bayes factors shared the same expression. The marginal prior density of the column vector $(t_1, \cdots, t_a)^\top$ is a heavy-tailed multivariate Cauchy distribution (e.g., Kotz & Nadarajah, 2004).

## Analytic Bayes factors for ANOVA

Two impediments to adopting the Bayes factor approach are usually the complexity regarding the high-dimensional integration of the marginal probabilities and the subjectivity associated with choosing the prior distributions (Craiu et al., 2022; Kass & Raftery, 1995; Morey et al., 2016). Default Bayes factors using the priors in Equations 2 to 7 have the advantage of preventing Bartlett's paradox (the Bayes factor approaches zero as the prior variance increases; Wang & Liu, 2016) and the information paradox (the Bayes factor tends to be bounded, given overwhelming information; Ly et al., 2016). In the case of the Bayesian two-sample $t$-test (equivalently, a between-subjects design where the number of conditions is $a = 2$), the default *ttestBF* and *anovaBF* functions in the '*BayesFactor*' R package by Morey and Rouder (2022) return the same Bayes factor value (Wei et al., 2022b, p. 9). Unlike the *anovaBF* function, the *ttestBF* function yields negligible Monte Carlo errors and does not require specifying the number of Markov chain Monte Carlo (MCMC) iterations. This convenience is made possible because the computation of the so-called Jeffreys-Zellner-Siow (JZS) Bayes factor for the Bayesian $t$-test is just the integration across one dimension in Equation 8 (Ly et al., 2016, p. 24; Rouder et al., 2012, p. 360; Rouder et al., 2009, p. 237; see also the balanced one-way between-subjects ANOVA in Morey et al., 2011, p. 374), so that Monte Carlo sampling is not necessary. In particular, $n = n_1 n_2 / (n_1 + n_2)$ represents the effective sample size (the default scale parameter is $h = \sqrt{2}/2$), $n_1$ and $n_2$ are the sample sizes for two groups, and $\nu = n_1 + n_2 - 2$ is the degrees of freedom in a two-sample $t$-test.

$$\text{JZS-BF}_{10} = (2\pi)^{-\frac{1}{2}} h\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}} \int_0^{\infty} (1+ng)^{-\frac{1}{2}} \left(1+\frac{t^2}{(1+ng)\nu}\right)^{-\frac{\nu+1}{2}} g^{-\frac{3}{2}} e^{-\frac{h^2}{2g}} \mathrm{d}g. \tag{8}$$

The JZS Bayes factor can be computed simply based on the $t$-statistic and the sample size using the integral representation in Equation 8. Relatedly, Jeffreys (1936, p. 417) provided a Bayes factor approximation for a point-null hypothesis test, which took the form of $\sqrt{n}$ multiplying an exponential function of the Wald test statistic ($W = t^2$; see also Lovric, 2020; Ly et al., 2018; Wagenmakers, 2022). Faulkenberry (2021) introduced the Pearson Bayes factor for the one-way between-subjects ANOVA and expressed it in an analytic form using the ANOVA $F$-statistic and its degrees of freedom, assuming the Pearson type VI distribution for the ratio of variance components $g$, i.e., $\pi(g) = \frac{g^{\beta}(1+g)^{-\gamma-\beta-2}}{\mathrm{B}(\gamma+1,\,\beta+1)}$, where $\mathrm{B}(\gamma+1, \beta+1)$ is the beta function with $\gamma, \beta > -1$. Maruyama and George (2011, p. 2749) and Wang and Sun (2014, p. 5078) imposed the shape parameters $\beta = \frac{N-a}{2} - \gamma - 2$ to further simplify the prior specification. The resulting Pearson Bayes factor is given in Equation 9, where $\Gamma(\cdot)$ is the gamma function and $df_{\mathrm{B}} = a - 1$, $df_{\mathrm{W}} = N - a$, and $df_{\mathrm{T}} = N - 1$ are the degrees of freedom for between-groups, within-group, and total sources of variation, respectively. $N = \sum_{i=1}^{a} n_i$ denotes the total number of observations. Despite their different $g$-prior assumptions, the Pearson Bayes factor matches the JZS Bayes factor, especially when $\gamma = 0$ (Faulkenberry, 2021). Based on Equation 9, Theorem 1 employs a quadratic exponential function to describe the relationship between the separation percentage of frequentist confidence intervals and the Pearson Bayes factor for a balanced one-way between-subjects design, given a fixed sample size and two conditions. This approximation gets more accurate as the sample size increases.

$$\text{P-BF}_{10} = \frac{\Gamma\left(\frac{df_{\mathrm{B}}}{2}+1+\gamma\right)\cdot\Gamma\left(\frac{df_{\mathrm{W}}}{2}\right)}{\Gamma\left(\frac{df_{\mathrm{T}}}{2}\right)\cdot\Gamma(1+\gamma)}\left(\frac{df_{\mathrm{W}}}{df_{\mathrm{W}}+df_{\mathrm{B}}\,F}\right)^{-\frac{df_{\mathrm{W}}}{2}+1+\gamma}. \tag{9}$$

**Theorem 1.** As $n \to +\infty$, the Pearson Bayes factor for the balanced one-way between-subjects ANOVA with two conditions can be approximated arbitrarily well by a quadratic exponential function of the separation between $(1-\alpha)\times 100\%$ confidence intervals for two population means $\mu_1$ and $\mu_2$,

$$\left|\text{P-BF}_{10} - A\cdot\exp\left\{B\cdot Sep^2\right\}\right| \to 0, \text{as } n \to +\infty,$$
$$\text{where } A = \frac{\Gamma(1.5+\gamma)}{\Gamma(1+\gamma)\sqrt{n}} \text{ and } B = z^2_{1-\frac{\alpha}{2}}.$$

The proof along with a graphical illustration is provided in Appendix A. Note that coefficient $A$ depends on the sample size and the hyperparameters, while coefficient $B$ depends on $\alpha$. The coefficients precisely match the remarks in Jeffreys (1961, p. 277) and Wagenmakers ($A = 2/\sqrt{n\pi}$; 2022, p. 26) when $\gamma = 0.5$. Both JZS and Pearson Bayes factors remain applicable when the design is unbalanced, and Theorem 1 can be likewise extended. Although the theorem provides the asymptotic approximation as a quadratic exponential function of the separation between confidence intervals, it is also relevant for the separation between HDIs because of the standard asymptotic normal form of the posterior distribution (e.g., Gelman et al., 2013, chap. 4; Jaynes & Kempthorne, 1976), which leads to an asymptotic relationship between the HDI and the standard confidence interval for ANOVA.

## Within-subject design

In the context of the within-subject design, the Bayes factor approach considers a linear mixed-effects model $\mathcal{M}_1$ of the response,

$$\mathcal{M}_1 : Y_{ij} = \mu + \sigma_\epsilon\left(t_i + b_j\right) + \epsilon_{ij} \quad \text{versus} \quad \mathcal{M}_0 : Y_{ij} = \mu + \sigma_\epsilon b_j + \epsilon_{ij},\ \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2_\epsilon\right), \quad \text{for } i = 1, \cdots, a \text{ and } j = 1, \cdots, n, \tag{10}$$

where the terms are as in Equation 1, and in addition, $b_j$ is the standardized random effect of each subject, $b_j \mid g_b \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, g_b\right)$, and is independent of $t_i$. Adding the subject-specific random effect can distinguish two conventional between-subjects and within-subject designs by whether each participant receives a single level of experimental manipulation or a collection of them. We again assume that $g_b \sim \text{Scale-inv-}\chi^2\left(1, h_b^2\right)$ for the $g$-priors and set $h_b = 1$ as a default. If only two conditions are considered in between-subjects or within-subject designs, it is not necessary to

introduce the quite complex ANOVA $g$-priors, but instead we can focus on the simpler case of the $t$-test ($F = t^2$). Default functions in '*BayesFactor*' version 0.9.2 and later return the same Bayes factor estimates for the independent-samples $t$-test and one-way ANOVA with two conditions. However, those functions return systematically different estimates for the paired $t$-test and one-way repeated-measures ANOVA with two conditions (see vignettes in Morey & Rouder, 2022). As it is the more general case, we consider the ANOVA $g$-priors.

## Confidence intervals

Loftus and Masson (1994) proposed a data transformation that removes the between-subjects variability prior to the construction of the confidence interval for the population means in within-subject designs. The motivation is to remove the irrelevant between-subjects variability and reveal the actual pattern of the population means in within-subject designs. As a result, the interval is based on the interaction mean-square error rather than the within-group mean-square error used in the standard confidence interval. Thus, LM-CI is not a standard confidence interval, but the practical utility of removing the nuisance between-subject variability has made it a widely used statistical method in the analysis of within-subject designs (e.g., Jusczyk et al., 1999; Urry et al., 2006; Vogel et al., 2001).

## Credible intervals

A Bayesian credible interval is interpreted with respect to posterior probability (i.e., the most likely values of the parameter, given the observed data). This interpretation is far more intuitive than that of a frequentist confidence interval (Hoekstra et al., 2014). As the most-used type of credible interval, an HDI represents the smallest interval on a (posterior) density distribution for a specified credibility level $(1-\alpha)\times 100\%$.

Nathoo et al. (2018) developed a within-subject HDI as the Bayesian analog of LM-CI for within-subject designs. The motivation for this development was to incorporate the usefulness of the within-subject interval into the Bayesian paradigm, which provides interval estimates with a conditional posterior probability interpretation. Whereas Loftus and Masson removed the irrelevant between-subjects variability using a subject-centering transformation of the data, Nathoo et al. derived a Bayesian within-subject interval by conditioning on maximum likelihood estimates of the subject-specific random effects. That is to say, Nathoo et al. considered a modified posterior distribution that is based on the conditional distribution of the parameters given the data and the subject-specific random effects and then plugged in the maximum likelihood estimates of the random effects. Nathoo et al. (2018, p. 4) showed that the original LM-CI could also be reformulated within this same framework under a specifically chosen prior distribution. This development provided a new interpretation of LM-CI as a conditional Bayesian interval. As a further development, they proposed an alternative interval based on replacing that prior with a more intuitive non-informative Jeffreys prior for the population means and residual variance. The resulting NKM-HDI was shown always to have a smaller width than LM-CI. NKM-HDI was then extended further to account for both the homoscedastic and heteroscedastic cases.

Conditioning on estimated random effects removes the irrelevant between-subjects variability when identifying the credible regions of population means that are relevant for evaluating evidence in within-subject designs. However, the maximum likelihood estimates of the random effects used in NKM-HDI do not exhibit shrinkage, and the procedure does not account for the uncertainty associated with the estimated random effects. Subsequently, Heck (2019) proposed modifying the conditional within-subject Bayesian interval to account for uncertainty and shrinkage in those effects. Heck derived a modification by applying the NKM-HDI formula at each iteration of an MCMC sampling algorithm and then taking the average interval across posterior samples (two-step approach; Ly et al., 2017). This modification introduces shrinkage estimation, as the MCMC draws will exhibit shrinkage, and accommodates the associated estimation uncertainty by computing the interval across iterations of an MCMC sampling algorithm. The posterior-averaged interval (LH-HDI) is wider than NKM-HDI, as shown in simulation studies (Heck, 2019, p. 29). This increase in width is understood to arise as a result of incorporating the uncertainty of the estimated random effects through the posterior draws.

NKM- and LH-HDIs assume improper uniform priors for the population means. In this work, we expanded the space of possible priors by assuming default $g$-priors for standardized treatment effects, as described by Rouder et al. (2017). Since the resulting HDI assumes the same priors in Equations 2 to 7 as those used to compute the default Bayes factor for within-subject designs, we will refer to it as JZS-HDI. Both LH- and JZS-HDIs take the same expression in Table 1, which is formally the Bayes estimator of the interval-valued parameter (thinking of the endpoints as a two-dimensional vector) under squared-error loss. They are centered on the posterior mean instead of the arithmetic sample mean, given that the intervals are computed from the average of MCMC draws. MCMC sampling of posterior means can also be used to obtain the standard HDI, which, unlike LM-CI and JZS-HDI, does not remove the between-subjects variability that is not of interest in within-subject designs.

### R packages

Given the relative complexity of within-subject designs,[2] we used simulations to study the fixed sample-size functional relation between interval separation and the Bayes factor. For the simulation studies below, we called two functions in the '*BayesFactor*' **R** package by Morey and Rouder (2022): the *anovaBF* function returns the Bayes factor object, and the *posterior* function returns the estimates for parameters by sampling from the posterior distribution of the numerator of the Bayes factor. According to the reference manual, the posterior is sampled with a Gibbs sampler (Morey & Rouder, 2022, p. 35). If not stated otherwise, the Bayes factors in all the figures presented here are JZS style and were computed by the '*BayesFactor*' **R** package. To provide practitioners with a high-level function for constructing within-subject HDIs without requiring any programming knowledge, we created a **Stan**-based **R** package, '*rmBayes*', available for download on the Comprehensive **R** Archive Network (CRAN). For both the homoscedastic and heteroscedastic cases in one-way within-subject designs, the *rmHDI* function provides multiple methods to construct the credible intervals for population means, with each method based on different sets of priors (see Appendices B–D and the reference manual for details; Wei et al., 2022a).

The default method (method 1) in our **R** function *rmHDI* (which corresponds to JZS-HDI) is based on the same priors that the *anovaBF* function uses to compute the Bayes factor in a within-subject design, assuming the Jeffreys prior for the overall mean and residual variance, a *g*-prior structure for effects, and independent scaled inverse-chi-square priors with one degree of freedom for the scale hyperparameters of the *g*-priors. The initial CRAN releases of '*rmBayes*', as well as Heck's (2019) scripts, miscoded the Jeffreys prior in **Stan**, resulting in a disparate prior $\pi\left(\mu,\sigma_\epsilon\right) \propto \frac{1}{\sigma_\epsilon^2}$ and slightly shorter interval estimates on average (Congdon, 2019, p. 51–52).[3] The computation of the within-subject Bayesian intervals in *anovaHDI* (see Appendix B) or *rmHDI* allows for either Gibbs sampling as implemented in the *anovaBF* function or, as an alternative, an MCMC algorithm based on the No-U-Turn sampler (NUTS) implemented in the '*rstan*' package by the Stan Development Team (2023, chap. 14). NUTS is an adaptive variant of the Hamiltonian Monte Carlo (HMC) algorithm that is designed to produce more efficient sampling than a standard HMC algorithm. The posterior sampling engine, whether it be Gibbs sampling or NUTS, can be specified by the user, with the latter being the default.

One drawback, as it were, of the option of using the '*rstan*' sampling engine in our **R** package computations is imperfect reproducibility resulting from Monte Carlo variability. Namely, the results may not be the same if practitioners call the function on different operating systems (even if the random seed is the same).[4] **Stan** results will be reproducible only if several configurations are identical, such as computer hardware and the **C++** compiler (Stan Development Team, 2023, chap. 19). Ours appears to be the first analysis of the Monte Carlo error associated with **R** packages '*BayesFactor*' and '*rmBayes*' (see Appendix C for further discussion).

## Simulation studies

### Between-subjects design

We first conducted a Monte Carlo simulation for a one-way between-subjects design with two conditions ($a = 2$). These between-subjects data were generated from the linear model in Equation 1. The number of subjects in each balanced group was set to $n = 24$ or $n = 48$. The standardized size of the difference between the two population means was set according to $d = (100 - \mu_2)/20$, and this size was adjusted to obtain cases with a power of .3 or .8 (the probability of avoiding a type II error in a standard significance test; e.g., Cohen, 1988) and a significance level of .05 while holding the common population standard deviation at $\sigma_\epsilon = 20$. Such a two-sample *t*-test power analysis can be run on the statistical software G*Power (Faul et al., 2007) to determine the required value of $\mu_2$.

The Bayes factor approach can also quantify the evidence supporting the null hypothesis as well as evidence for the alternative hypothesis. Therefore, we extended our simulations to include data simulated under the null model $\mathcal{M}_0$. The parameter settings are straightforward, letting $\mu_2 = \mu_1 = 100$. The $2 \times 3 = 6$ combinations of simulation parameters, representing the variation in sample size and power (effect size), are presented in Table 2. We generated

[2] Faulkenberry and Brennan (2022) extended Equation 9 to a closed-form expression of the Pearson Bayes factor for within-subject designs simply by substituting $N^* = n(a-1)$ for the total number of independent observations (Masson, 2011, p. 682).

[3] The Stan syntax *target += -log(sigma);* in place of *target += -2*log(sigma);* has been implemented in version 0.1.15 and later of the '*rmBayes*' **R** package to accurately reflect the Jeffreys prior in Equation 2. Regardless of which syntax is used, there is little difference in graphical results, which can be seen as an example of a sensitivity analysis for different possible priors. '*rmBayes*' 0.1.15 was used for the computations reported in this article.

[4] By calling *rmHDI(recall.long, iter = 2e4, seed = 277)$width*, macOS may return 0.5613043, Intel-based macOS may return 0.5601921, Compute Canada Cedar may return 0.5600443, and Windows may return 0.5589209.

**Table 2** Parameters for simulated data in a between-subjects design with two conditions

| Case | 1 | 2 |
|---|---|---|
| $n$ | 24 | 48 |
| $\mu_1$ | 100 | 100 |
| $\mu_2$ | (100, 91.5, 83.5) | (100, 94.1, 88.4) |
| $\sigma_\epsilon$ | 20 | 20 |
| power | (NA, .3, .8) | (NA, .3, .8) |

*Note*. 10,000 simulations were run for each effect size. $n$ is the number of subjects in each group. $\mu_1$ and $\mu_2$ are population means, and each case selects three values for $\mu_2$. $\sigma_\epsilon$ is the common standard deviation. $\mu_2$ was adjusted to obtain the desired power according to a two-sample $t$-test power analysis

10,000 between-subjects data sets under each combination of parameter settings. Then we computed the Bayes factor and two types of interval estimates (the standard frequentist 95% confidence interval and the standard Bayesian 95% HDI) for each data set.

Given a fixed sample size, data simulated under different true effect sizes merge into one curve relating the Bayes factor to interval separation. This observation is displayed in Fig. 2A for the standard confidence interval, where data points form a continuous curve with either $n = 24$ or $n = 48$, regardless of whether the true effect size is null (yellow), small (red), or large (blue). From Fig. 2A, we see that the different effect sizes make up different, but somewhat overlapping, sections of the same curve as larger Bayes factors (and greater degrees of interval separation) are usually generated under larger effect sizes. Hence, the Bayes-factor/separation relationship does not depend on the effect size.

As shown in Fig. 2E, the observed relationship between the Bayes factor (its natural logarithm) and interval separation is well described by a quadratic curve for simulations under null to large effect sizes but yields a linear function in black for the extreme effect size. Indeed, the parabola eventually opens downward once the separation exceeds 500%, given $n = 24$. Such a lack of fit of the quadratic function at the long tail arises because the component $(SS_B/SS_W)^2$ is no longer the least significant in the exponential series (see the proof of Theorem 1). Thus, a higher-order term, such as the quartic term, is necessary even for relatively large but finite sample sizes.

To compare the simulated function relating the Bayes factor and interval separation to the analytic function generated by Theorem 1, we plotted them together in Fig. 3. That figure presents two scatter plots, each obtained from 30,000 merged replicates showing the Bayes factor versus the corresponding interval separation for the standard confidence interval (left) and the standard HDI (right) in a between-subjects design with $a = 2$ and $n = 48$. The separation percentage was computed for the intervals of each pair of condition means. There is a clear quadratic exponential relationship between the interval separation and the Bayes factor for a particular value of sample size. The data points are perfectly aligned on a fitted quadratic exponential curve shown as a solid black function in Fig. 3. The fitted coefficients of the quadratic exponential are comparable to the asymptotic coefficients obtained by plugging the known values into the

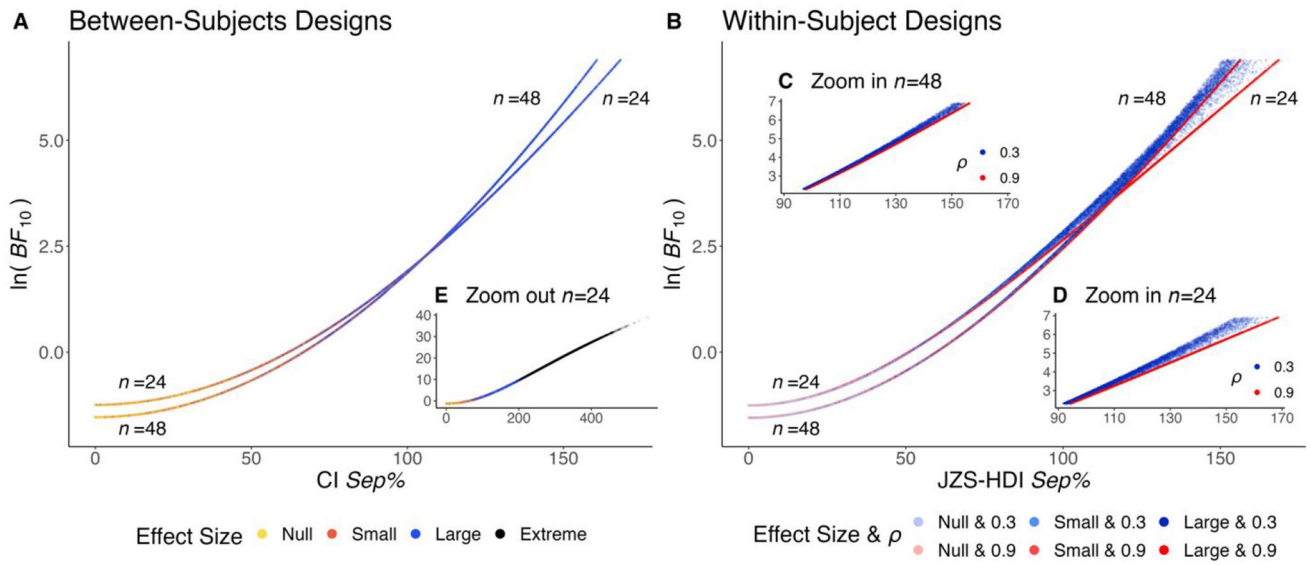


**Fig. 2** Scatter plots of the relationship between the natural logarithm of the Bayes factor and the separation (**A**) of the between-subjects confidence intervals or (**B**) of the within-subject JZS-HDIs under two fixed sample sizes and a range of effect sizes (or a combination of effect sizes and correlations) indicated by color. (**C**) and (**D**) are the partial magnifications of cases that overlay in (**B**). (**E**) displays the full scale of one case in (**A**), including an additional simulation with an extreme effect size (Cohen's $d = 2.5$)

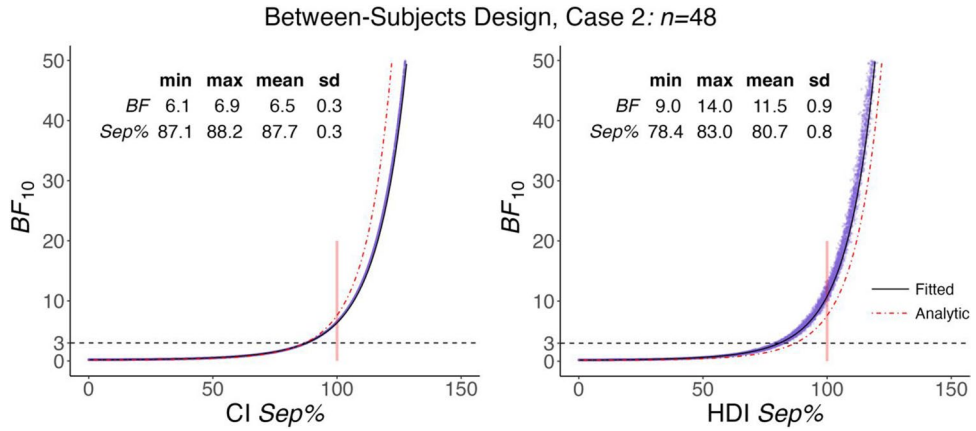


**Fig. 3** Scatter plots of the relationship between the between-subjects interval separation (standard confidence interval on the left, standard HDI on the right) and the Bayes factor (truncated at 50) for the aggregate of three effect sizes and sample size of 48 in Case 2 from Table 2. The table in each panel shows descriptive statistics for the Bayes factor when interval separation is $100 \pm 1\%$; the *Sep%* row refers to the separation value when the Bayes factor is $3 \pm 0.1$. The fitted curve is in solid black, and the analytic curve (Theorem 1) is in dash-dotted red. The horizontal dashed line indicates a Bayes factor of 3, and the vertical red line indicates a separation percentage of 100

formula of Theorem 1, with the latter being plotted as the dashed-dotted red curve in both panels of Fig. 3. The discrepancy between the fitted and analytic functions becomes apparent for confidence interval separations beyond about 80%, given $n = 48$. The approximation will be less accurate if the sample size is insufficient and the study is underpowered. As the sample size increases, conformity to the analytic curve will improve (see Appendix A). However, we still expect some difference between the fitted and analytic values even for very large $n$ because the Bayes factors are computed under different priors (JZS for the simulated data and Pearson for the theoretical function).

As discussed in Loftus and Masson (1994), two means are significantly different if separated by at least $\sqrt{2}$ times the confidence interval width when a design-specific interval is used. Therefore, from our definition of the separation percentage, at least 71% separation ($\sqrt{2}/2$) in the confidence intervals of two population means provides evidence for an effect with a standard $p$-value of .05 and a 95% confidence interval. As Fig. 3 shows, however, that with $n = 48$, the average separation of confidence intervals (88%) is greater than 71% when the Bayes factor value is about 3 (moderate evidence for an effect; Lee & Wagenmakers, 2014; Raftery, 1995; Bayes factor evidence should not be evaluated on a strict criterion, but rather with reference to the context of the research; Evett, 1987; Kruschke, 2021; Rouder et al., 2017, p. 318). This greater separation threshold occurs because the Bayes factor value of 3 usually corresponds to a stricter standard $p$-value of around .01 instead of .05, given a specific sample size (Jeffreys, 1961, p. 435; Raftery, 1995, p. 789; Wagenmakers, 2022, p. 15). Consequently, the separation threshold favoring the alternative model $\mathcal{M}_1$ is relatively conservative from the Bayesian perspective relative to NHST, which is consistent with observations from Ly et al. (2016, p. 24), Nathoo and Masson (2016), and Wetzels et al. (2011). Similarly, Fig. 3 shows that for the standard Bayesian HDI, about 81% separation is associated with a Bayes factor of about 3. The table in each panel shows descriptive statistics (1) for the Bayes factor when interval separation ranges from 99% to 101% and (2) for the separation percentage when the Bayes factor ranges from 2.9 to 3.1.

## Within-subject design

We next conducted a Monte Carlo simulation for a one-way within-subject design with two conditions. The effect size was empirically determined from the ANOVA design with treatment effects and subject-specific random effects as specified in model $\mathcal{M}_1$ from Equation 10. Simulations were run with various effect sizes until one was found that

**Table 3** Parameters for data simulation in a within-subject design with two conditions

| Case | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $n$ | 24 | 24 | 48 | 48 |
| $\mu_b$ | 100 | 100 | 100 | 100 |
| $\Delta\mu$ | (0, 3.8, 6.7) | (0, 16.5, 28.6) | (0, 2.7, 4.6) | (0, 12.0, 20.4) |
| $\sigma_b$ | 20 | 20 | 20 | 20 |
| $\sigma_\epsilon$ | 6.67 | 30.55 | 6.67 | 30.55 |
| $\rho$ | .9 | .3 | .9 | .3 |

*Note*. 10,000 simulations were run for each effect size. $n$ is the number of subjects. $\mu_b$ is the baseline mean. $\Delta\mu$ is the raw-score effect size, and each case selects three values. $\sigma_b$ is the standard deviation of the subject-specific random effect. $\sigma_\epsilon$ is the standard deviation of the error. The correlation between two conditions is $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$

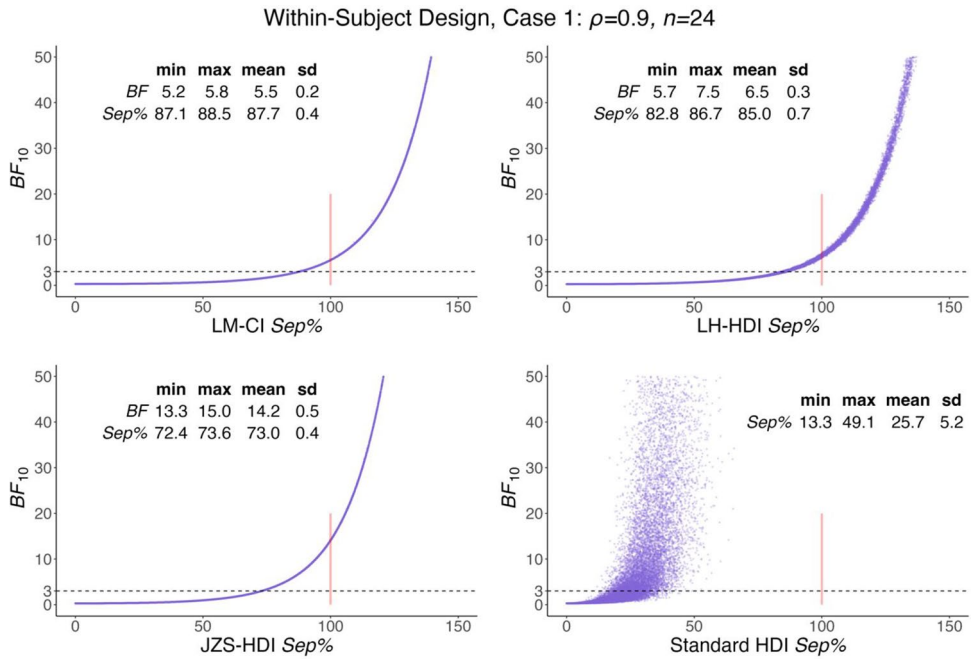


**Fig. 4** Scatter plots of the relationship between within-subject interval separation and the Bayes factor (truncated at 50) for the aggregate of three effect sizes, correlation of .9 and sample size of 24 in Case 1 from Table 3. The table in each panel shows descriptive statistics for the Bayes factor when interval separation is 100 ± 1%; the *Sep%* row refers to the separation value when the Bayes factor is 3 ± 0.1. The horizontal dashed line indicates a Bayes factor of 3, and the vertical red line indicates a separation percentage of 100

produced approximately the desired level of power (.3 or .8, probability of a Bayes factor of at least 3). We ran 10,000 simulations for each parameter set based on $2 \times 3 \times 2 = 12$ combinations of sample size, power, and correlation between conditions. The 12 settings are categorized into four cases and presented in Table 3. As in the between-subjects case, the within-subject results can be collapsed across effect sizes because reports with the same sample size and within-subject correlation generated identical curves, as shown in Fig. 2B. There may be some outliers in the computed values, with particularly large Bayes factor values occurring even when interval separation is quite low. Outliers like these are the result of Monte Carlo variability and numerical issues in the Bayes factor computation. Even though the number of MCMC iterations taken for a particular Bayes factor was as many as 100,000, such outlying Bayes factor values do occasionally occur. Fortunately, these instances are typically indicated by a large error estimate that is reported with the Bayes factor in the *anovaBF* function, suggesting that the analysis ought to be run again with a different random seed. See the discussion of this issue in Appendix C. All figures have excluded simulations that estimated proportional error on the Bayes factor to be greater than 1%.

Figures 4 and 5 display eight scatter plots showing the relationship between the Bayes factor and the interval separation, computed for data simulated from within-subject designs with two conditions under two different combinations of parameters (Cases 1 and 4 from Table 3). The separation percentage of NKM-HDI for a data set can be numerically derived from that of LM-CI by multiplying a factor of $\sqrt{\frac{a}{a-1}} \cdot \frac{t^*_{1-\frac{\alpha}{2},\,(n-1)(a-1)}}{t^*_{1-\frac{\alpha}{2},\,a(n-1)}} > 1$ (therefore, the NKM-HDI plots are omitted). The complete set of reports for the two between-subjects and four within-subject cases for each interval type can be viewed on the Open Science Framework website at https://osf.io/x2pvw/. As in Fig. 3, the horizontal dashed line in Figs. 4 and 5 indicates a Bayes factor of 3, plus or minus 0.1. The vertical red bar indicates 100% interval separation (±1%) as a reference line.

The table in each panel shows that the two benchmark values (Bayes factor associated with 100% separation and separation percent associated with a Bayes factor of 3) vary systematically across the four different within-subject intervals due to the difference in magnitude of the respective intervals computed from the same data set (LM-CI < LH-HDI < JZS-HDI widths). First, with the degree of separation fixed, methods whose estimates produce narrower interval widths are generally associated with smaller Bayes factors. For example, the average Bayes factors are 5.5, 6.5, and 14.2 across the 100% separations of LM-CI, LH-HDI, and JZS-HDI, respectively, when $n = 24$ in Fig. 4. Second, with the Bayes factor fixed (thus, the observed effect size is held constant), narrower interval widths lead to larger degrees of separation. For example, the average separations of interval

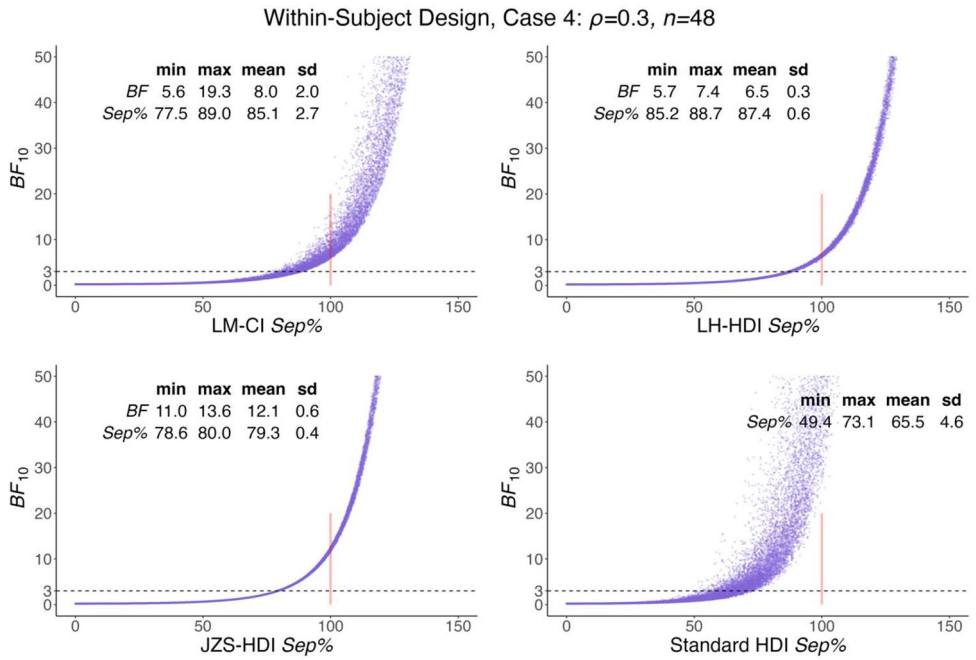


**Fig. 5** Scatter plots of the relationship between within-subject interval separation and the Bayes factor (truncated at 50) for the aggregate of three effect sizes, correlation of .3 and sample size of 48 in Case 4 from Table 3. The table in each panel shows descriptive statistics for the Bayes factor when interval separation is 100 ± 1%; the *Sep%* row refers to the separation value when the Bayes factor is 3 ± 0.1. The horizontal dashed line indicates a Bayes factor of 3, and the vertical red line indicates a separation percentage of 100

estimates are 88%, 85%, and 73% for LM-CI, LH-HDI, and JZS-HDI, respectively, when $BF_{10} \approx 3$ and $n = 24$ in Fig. 4.

For the within-subject design, we found a similar relationship between the within-subject interval separation and the Bayes factor, compared to the between-subjects case. In the within-subject design, however, the relationship between the standard HDI and the Bayes factor was evidently inconsistent for a given sample size. This result demonstrates the benefit of using within-subject intervals in within-subject designs. That is, the separation of within-subject HDIs corresponds well with the Bayes factor, whereas the separation of the standard HDIs does not. We observed that the standard HDI was so wide that it produced substantial overlap (little separation) in within-subject designs, even when there was clear evidence from the Bayes factor for a reliable effect. For example, the standard HDIs are separated at percentages from 13% to 49% when $BF_{10} \approx 3$ and $n = 24$ in Fig. 4. The standard HDI for within-subject designs is unnecessarily broad due to irrelevant between-subjects variability and tends to hide within-subject effects, rendering its relationship with the Bayes factor less clear.

In addition, the greater the correlation between the two conditions in a within-subject design, the more consistent the observed quadratic exponential relationship is, as can be seen by comparing Fig. 5 (low correlation between conditions) to Fig. 4 (high correlation), especially when the Bayes factor signals strong evidence against the null. Figure 2B depicts the Bayes-factor/HDI relationship for all four cases from Table 3. The two distinct curves are based on different sample sizes, and in each case, when the correlation is low, more dispersion is present in the relation. This trend becomes more apparent as the Bayes factor increases. Figures 2C and D present a section of each sample size's function that illustrates the dispersion related to the magnitude of correlation between conditions.

Because the Bayes-factor/HDI separation relationship was more precise for JZS-HDI than for any other intervals, we fitted a quadratic curve against the natural logarithm of the Bayes factor for each of the four simulation cases using the JZS-HDI. The results are shown in Fig. 6, where the fitted quadratic models are presented at the bottom of each panel. For all four cases, adjusted coefficients of determination, $R^2$, were 0.993 or greater. The analytic result in Theorem 1 suggests that the quadratic term is sufficient in the limit for between-subjects designs. Empirically, we found a quite good fit using the second-order quadratic equation to describe the relationship for either between- or within-subject data with moderate sample sizes. The simulated relation is concave downward when the observed effect size is large (i.e., a log Bayes factor greater than 8 in Fig. 6). The quadratic exponential relationship is a fairly simple description of the fit between the Bayes factor and HDI separation, and it reflects an asymptotic result as the sample size approaches infinity.

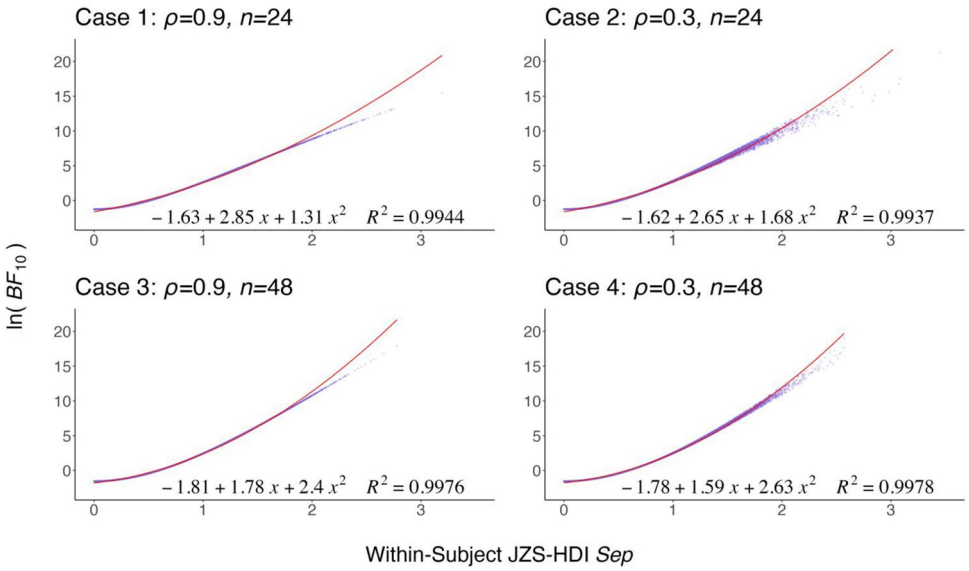


**Fig. 6** Quadratic exponential curve fits for JZS-HDI. The *x*-axis and the *y*-axis are in full scale with the *y*-axis showing ln ($BF_{10}$) values. The solid red curve indicates a multiple regression of ln ($BF_{10}$) on interval separation score (separation % divided by 100). The coefficients of determination are above 0.993 for all four cases

## Importance of sample size and evidence for the null

For within-subject data generated under the null model ($\Delta\mu = 0$), there were four cases representing variation in sample size (24 or 48) and size of correlation between conditions (.9 or .3). The benchmark values (the Bayes factor when separation percent was 100 and separation percent when the Bayes factor was 3) turned out to show similar descriptive statistics, regardless of whether the data were simulated under $\mathcal{M}_0$ or $\mathcal{M}_1$. Thus, the relationship between the Bayes factor and HDI separation does not depend on the true effect size but on the sample size and the magnitude of the correlation between conditions. Figure 7 shows the results of an additional set of simulations based on Cases 1 and 4 from Table 3. These simulations included a broader variation in sample size (from 5 to 48) and a different set of effect sizes ($\Delta\mu = 0$, 3.8, and 6.7 in Case 1, and $\Delta\mu = 0$, 12.0, and 20.4 in Case 4, representing null, small, and relatively larger effect sizes). These simulations highlight how separation percentage and Bayes factor benchmarks change with effect size and sample size. Cases 1 and 4 differ in the size of the correlation between conditions. We restricted the standardized effect sizes, $\Delta\mu/\sqrt{\sigma_b^2 + \sigma_\varepsilon^2} \equiv \Delta\mu\sqrt{1-\rho}/\sigma_\varepsilon$, on the small and relatively larger scales so that some data sets would produce a Bayes factor of 3. With even larger effect sizes, the Bayes factor would seldom be as low as 3.

First, in Figs. 7E and F, the separation percentage supporting an effect (Bayes factor of 3 ± 0.1) increases with the sample size. For sample sizes typical of within-subject designs in experimental psychology (i.e., 20–50), the separation percent associated with evidence for an effect is in the range of 70–80%. Similarly, in Figs. 7G and H, the separation percentage favoring the null ($BF_{01} = 1/BF_{10}$; reciprocal Bayes factor of 3 ± 0.1) also increases with the sample size. A blue square symbol, representing the large true effect size, is missing for $n = 48$ in Fig. 7G because few observations would be simulated in such an instance. Similarly, the separation benchmarks are absent for limited sample sizes of 5 and 10 because those functional curves never fall below the –ln 3 line (see Figs. 7A and B). With a case of $n = 10$, for example, practitioners might be tempted to draw incorrect conclusions, such as taking an interval separation of close to 0% as evidence for the null. These simulations demonstrate that obtaining a result that provides moderate evidence in favor of the null hypothesis would be virtually impossible due to the dearth of data. Interpreting HDI separation as evidence for the null depends heavily on the sample size, as Figs. 7G and H show.

Second, Figs. 7C and D show that the Bayes factor corresponding to 100 ± 1% interval separation does vary nonmonotonically with the sample size. This nonmonotonicity arises because the Bayes-factor/HDI relationship is not effectively described by the quadratic exponential when the sample size is small (5 or 10). Third, the benchmarks associated with different effect sizes are notably similar given a particular sample size, except for cases in Figs. 7C and D. When the sample size is insufficient ($n = 5$; thus, underpowered), data sets simulated from larger true effect sizes produce slightly larger Bayes factors corresponding to 100 ± 1% interval separation. The results of an additional set of

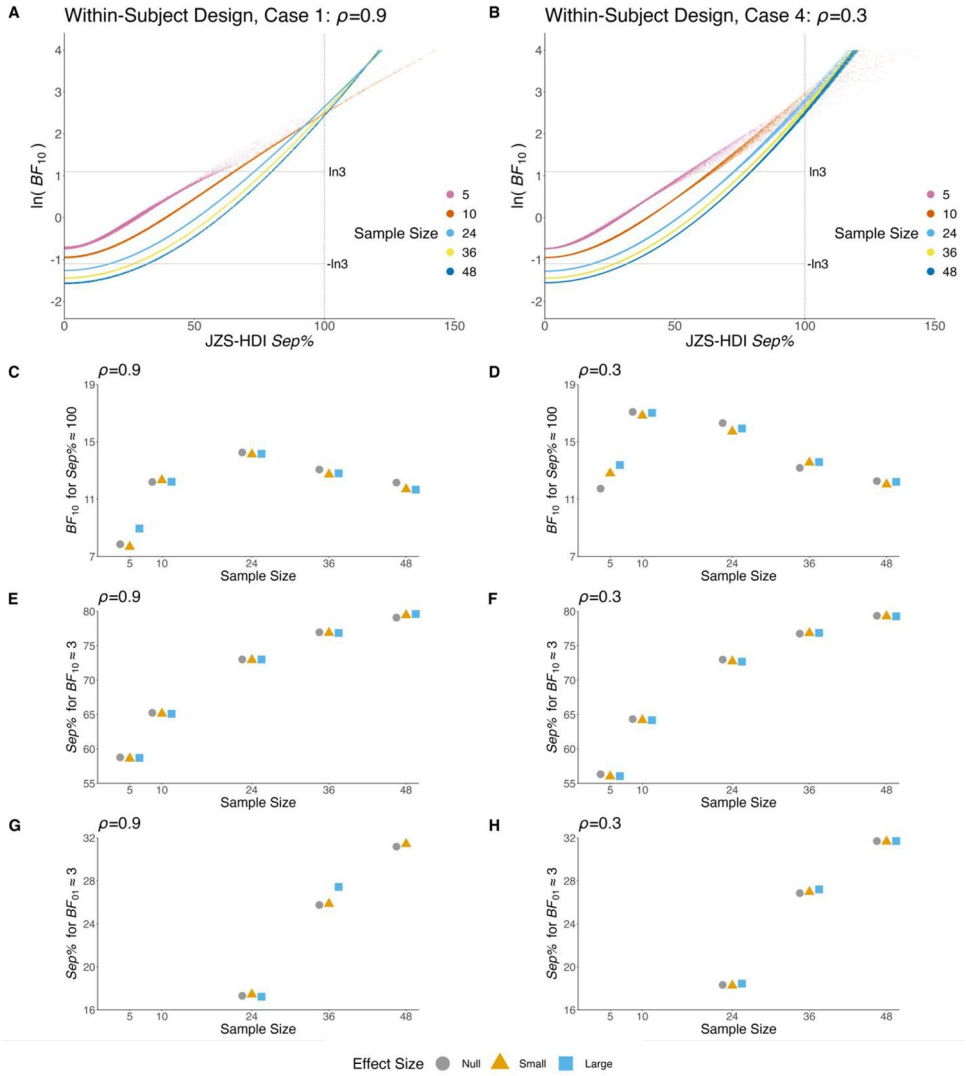


**Fig. 7** Simulations (**A** and **B**) were run according to the parameters in Cases 1 and 4 of Table 3 but with five sample sizes. Benchmarks (**C**) and (**D**) for the average Bayes factor when JZS-HDI separation is 100 ± 1%, (**E**) and (**F**) for the average separation percentage when the Bayes factor is 3 ± 0.1, and (**G**) and (**H**) for the average separation percentage when the Bayes factor is the reciprocal of 3 ± 0.1. The symbols have been horizontally jittered to avoid overplotting, and their color and shape represent the true effect sizes

simulations that assessed the benchmarks shown in Fig. 7 but for Bayes factors and standard HDIs in between-subjects designs are available at https://osf.io/x2pvw/.

Examining Figs. 3, 4, and 5, when the Bayes factor is larger than 3 and increasing, the separation is also increasing in a manner that can be predicted through a quadratic exponential function for a given sample size. Therefore, both measures can give evidence in favor of the alternative hypothesis, and the relationship holds when there is evidence for the alternative. Conversely, in the other direction, when the Bayes factor is less than one-third and approaching zero, the simulations again demonstrate that the separation percentage follows a predictable pattern as the Bayes factor becomes vanishingly small. When the Bayes factor is between 1/(3.1) and 1/(2.9), the average separation percentages of JZS-HDIs are 17% and 32% for $n = 24$ and $n = 48$, respectively. Therefore, the separation percentage, like the Bayes factor, provides evidence in favor of the null, given an adequate sample size. That is, there is a clear monotonic relationship between the two quantities for both large values (evidence for the alternative) as well as small values (evidence for the null) of the Bayes factor. It is important to note that the separation percentage should not be used as the sole basis for supporting the null hypothesis while ignoring sample size and a specific alternative hypothesis. A similar notion is also discussed in Wagenmakers (2022) within the context of the relationship between the Bayes factor and the *p*-value for a given sample size.

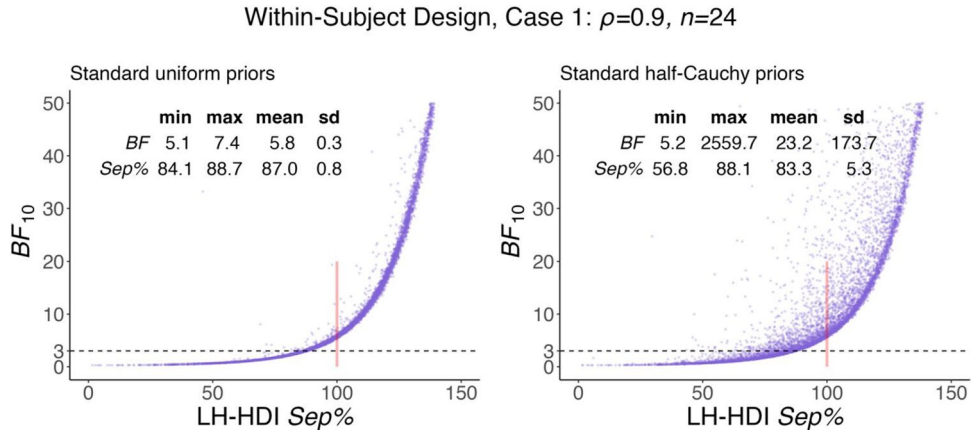


**Fig. 8** Scatter plots of the relationship between within-subject interval separation and the Bayes factor for LH-HDI for one of the sets of parameters shown in Table 3 (Case 1) and two different priors for the standard deviation of the subject-specific random effects. Left: $\sigma_\epsilon\sqrt{g_b} \sim U(0,1)$; right: $\sigma_\epsilon\sqrt{g_b} \sim HC(0,1)$

## Sensitivity analysis

We also conducted a sensitivity analysis with respect to the prior for the variance $g_b$ of the standardized subject-specific random effects by substituting for the inverse-chi-square distribution (for $g_b$) either a uniform distribution or a half-Cauchy distribution (for $\sigma_\epsilon\sqrt{g_b}$), which generated almost identical within-subject HDIs as in Heck (2019, p. 29). These priors correspond to methods 5 and 6 in our **R** function *rmHDI*. Simulations using these priors continued to show a quadratic exponential relationship between the Bayes factor and interval separation that varied somewhat with sample size (see Fig. 8 for a sample case). The lower boundaries of the curves in the scatter plots are like those in our original simulations in Fig. 4. However, the data points are somewhat more variable than the case using the default scaled inverse chi-squared priors.

## Multilevel and multiway ANOVA

We have described the relationship between a Bayes factor and an HDI only for the case of two condition means, either in a within-subject or a between-subjects design. One might ask about the nature of this relationship when there are more than two levels of the independent variable. In such cases, this relationship is quite difficult to define, given that the Bayes factor generated by analyzing data from all of the conditions could be associated with a wide range of patterns among population means and hence a large variety of possible degrees of separation between means. For example, in a within-subject design with three conditions, the three means may be equally spaced in rank order (e.g., $\mu_1 < \mu_2 < \mu_3$), or two of the means may be nearly identical and the third quite different from the first two. The Bayes factors for these two situations might be the same, and the HDI for the data might also be similar for the two cases, but the degree of separation between a given pair of means defined by the HDI metric we have used here would depend on which two means are chosen.

Our approach, therefore, has been to examine the Bayes-factor/HDI relationship specifically for the case of two conditions. The conclusions we have reached can be readily extended to designs with more than two conditions, assuming that one is interested in considering two of the population means at a time (see also testing equality constraints in Morey, 2015a; order constraints in Morey, 2015b). The computed HDI, generated as it is from all data in the design, can be taken as a general estimate of the stability of the difference between any two means in the design. Hence, in a within-subject design with three or more conditions and, for example, $n = 24$, if any two means are separated by about 73% of the length of the JZS-HDI, one can safely conclude (provided that the assumption of circularity holds reasonably well, i.e., the variances of the differences between any pair of within-subject conditions are roughly equal) that a Bayes factor computed to compare just those two conditions would turn out to be about 3; a larger separation would imply an even larger Bayes factor. Should the assumption of circularity be violated, it would be advisable to modify this approach (see below).[5] If only homoscedasticity is violated, then HDIs can be constructed using the method described by Nathoo et al. (2018, p. 5) for within-subject HDIs under

[5] As a sufficient but not necessary condition for conducting repeated-measures ANOVA, the compound symmetry assumption states that all conditions have equal population variance, and all pairs of conditions have equal covariance. Hence, compound symmetry is a restrictive form of circularity. See remarks in Cousineau (2019, p. 232).

heteroscedasticity, with Heck's (2019) modification applied. The formulation is

$$\mathbb{E}\left[\mu_i \pm \frac{\sigma_i}{\sqrt{n}} \cdot t^*_{1-\frac{\alpha}{2},\, n-1} \mid \text{Data}\right]. \tag{11}$$

An option to compute HDIs for the heteroscedastic case is available in the *rmHDI* **R** function.

If the equal covariance assumption is not met, then we recommend constructing a separate HDI for each pair of condition means to provide a more precise depiction of the relationship between those means (see Franz & Loftus, 2012, for examples of plotting multiple confidence intervals for different pairs of means in a within-subject design). One can have confidence that the relationship between HDIs constructed in this way and the corresponding Bayes factor would adhere to the benchmarks reported here.

Similar reasoning is relevant when considering more complex designs with multiple factors. Equation 12 models a two-way repeated-measures design, where $\epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$ and $Y_{ijk}$ is the response for $k$th subject at the $i$th level of factor $d$ and $j$th level of factor $p$ for $i = 1, \cdots, a$, $j = 1, \cdots, b$, and $k = 1, \cdots, n$. Such a random slope model is often confused with a two-factor factorial design with blocking in Equation 13, which assumes that the interactions between treatments and blocks are negligible, as in agricultural experiments.[6] Importantly, Equations 12 and 13 imply the same one-way model $\mathcal{M}_1$ from Equation 10, used for our reported simulations when there is only one factor and one observation per condition. In a design with multiple within-subject factors, one could treat all the conditions as levels of a single factor and compute HDIs using the methods described here. In a 2×3 within-subject design, for instance, there would be six conditions, and these could be treated as six levels of a one-way within-subject design. Concerns about the circularity assumption would be important here, and the methods discussed above for addressing violations of this assumption would need to be considered. For mixed designs containing at least one within-subject and one between-subjects factor, one might choose to plot the means with either HDIs reflecting within-subject variability or between-subjects variability, taking into account possible violations of circularity for the within-subject factor and violations of homogeneity of variance for the between-subjects factor. If all assumptions are met, it might be informative to plot both between- and within-subject HDIs for each mean. In such a case, the relationship between the HDIs and the associated Bayes factor should adhere to the pattern revealed by our simulations. We note that these relationships would be expected to hold for the main effects of the factors, but we have not explored interaction effects here. Masson and Loftus (2003) provided some suggestions about plotting and visually assessing interactions using confidence intervals.

$$Y_{ijk} = \mu + \sigma_\epsilon\left(s_k + d_i + p_j + (dp)_{ij} + (ds)_{ik} + (ps)_{jk}\right) + \epsilon_{ijk}. \tag{12}$$

$$Y_{ijk} = \mu + \sigma_\epsilon\left(s_k + d_i + p_j + (dp)_{ij}\right) + \epsilon_{ijk}. \tag{13}$$

## Example application

To illustrate the use and interpretation of within-subject HDIs, we consider a real 2×2 within-subject data set taken from a published study by Bub et al. (2021, Exp. 2), which examined speeded classification of pictured hand cues according to their laterality (left versus right hand). The hand cue showed a left- or right-handed power grasp with the palm oriented horizontally or vertically. The task was to classify the hand's laterality by making a key press using either the left or right hand, depending on which hand was shown. The hand cue was presented superimposed on a picture of a handled object (e.g., a frying pan) whose handle pointed left or right. The pictured hand's laterality was either aligned or not aligned with the object handle's location (e.g., a left hand with the object handle pointed left is an example of the aligned condition), and this variation was one of the two independent variables in the design. The other factor was compatibility between the horizontal/vertical orientation of the object's handle (e.g., a frying pan had a horizontal handle, and a beer mug had a vertical handle) and the pictured hand's palm orientation. The dependent measure was time taken to make the key press to classify the hand cue's laterality. For the present example, we used data from a condition in which the hand cue and object appeared simultaneously (0-ms stimulus onset asynchrony). The data and subsequent **R** scripts for this example are available at https://osf.io/x2pvw/. For each of the 37 subjects, a condition score was taken as the mean response time for correct responses made on trials in that condition.

Our first step in constructing 95% HDIs for the population means was to determine whether all four conditions in the design could be validly treated as four levels of a single-factor design. The advantage of treating the design this way is that a single HDI can be computed that would be suitable for making comparisons between any pair of conditions. This approach would require that the assumption of circularity holds for these four conditions. The assumption can be tested

[6] At the time of writing, the '*BayesFactor*' **R** package by Morey and Rouder (2022) implemented Equation 13 for multiway within-subject designs, but van den Bergh et al. (2022) have realized the misspecification and started to update the functions accordingly. See also Kruschke (2014, p. 606-608). Changes will not affect the one-way models used for simulations in this article.

using the *ezANOVA* function in the '*ez*' **R** package by Lawrence (2016). The *ezANOVA* function computes the ANOVA for various designs and provides a statistical test of the circularity assumption (referred to in the output of *ezANOVA* as Mauchly's test of *sphericity*). The function also provides Greenhouse-Geisser (GG) and Huynh-Feldt (HF) epsilon values, which can be used to adjust degrees of freedom for a significance test, should circularity be violated (Greenhouse & Geisser, 1959; Huynh & Feldt, 1976). In general, an epsilon value of .75 or greater can be taken as sufficient support that the circularity assumption holds (e.g., Loftus & Masson, 1994, p. 483). One could instead adopt a more stringent requirement, such as a nonsignificant Mauchly's test of the circularity assumption. The output for our example data set indicates that Mauchly's test is not significant ($p$-value of .59) and that both versions of the epsilon value are greater than .75 (GG = 0.93, HF = 1.02), so it is safe to treat the four conditions as levels of a single factor and to compute a single HDI. If the circularity assumption had not been met, it would be best to compute a separate HDI at each level of one of the two factors. For instance, in this data set, the greatest interest was in the alignment between the location of object's handle and hand laterality, so one could compute an HDI for the aligned and misaligned laterality conditions within each level of the orientation compatibility factor.

Proceeding with the computation of a 95% HDI for all four conditions, we can use the *rmHDI* function in the '*rmBayes*' **R** package. For all but the first method (method 0), MCMC sampling is used in computation of the HDI, so the function provides the option to specify the number of warmup iterations and the number of critical iterations to be used. We used more warmup iterations (2,000) and more critical iterations (10,000) than the default values because these larger values generally provide results with a low likelihood of the function producing warning messages regarding the operation of the MCMC sampling of parameters. Method 0 computes NKM-HDI, which does not use MCMC sampling and does not take into account possible shrinkage of or uncertainty about parameter values for between-subject variability. For the other methods, MCMC is used, and a Bayesian estimate of the posterior mean for each condition is generated. A 95% HDI is computed by default, although other levels of credibility can be used instead. The output shows that the resulting HDI width using method 1 (the default method, which computes JZS-HDI) is 6.78. We have plotted the posterior mean estimates and the HDI based on method 1 in Fig. 9. Keep in mind that because this is a homoscedastic within-subject HDI, the same HDI width is plotted for each mean and that it represents estimation error with respect to the relative values of these means, not their absolute values. Note that with the other methods, which differ regarding priors, the resulting HDI and posterior mean estimates vary somewhat, but they

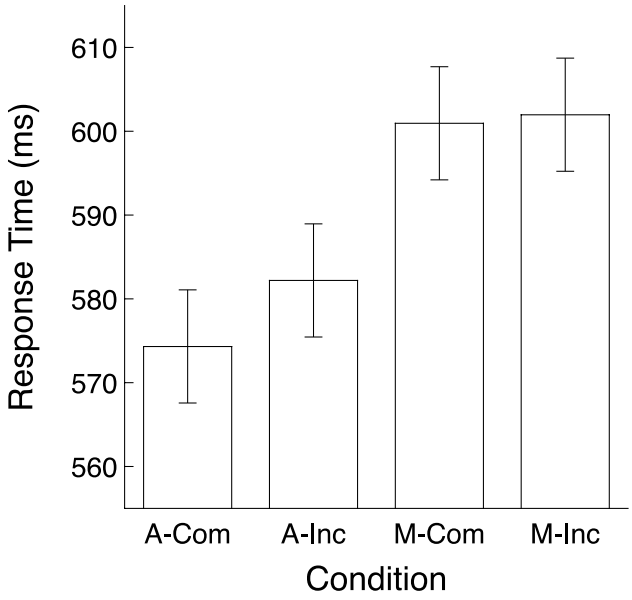


**Fig. 9** Estimated posterior means and 95% highest density interval for the four conditions in a 2×2 within-subject design ($n$=37) computed using the default method with the *rmHDI* function. The factor of alignment of object handle and hand is specified by A/M (A = aligned, M = misaligned), and the factor of orientation compatibility is specified by Com/Inc (Com = compatible, Inc = incompatible)

are not substantially different. The methods that use MCMC sampling generated for the present data have similar 95% HDI widths (around 6.78). Method 0, which does not take shrinkage or estimation error into account, produced a characteristically smaller HDI width (5.83).

The separation between the HDIs for the first two conditions, *A-Com* and *A-Inc*, is about 60%, so we can expect that there is no solid evidence for a difference between those two means, given $n$=37. The reported tests are pairwise comparisons of factor combinations and not tests of main effects that are marginalized across the other factors. The next step is to perform a Bayesian test for the difference between these two population means (the data from the two irrelevant conditions are set aside, then the *anovaBF* function is used to compute a Bayes factor). This is a test of the effect of compatibility in orientation when the laterality of the object handle and the hand cue are aligned. The Bayesian test produced a Bayes factor of 1.08 in favor of an effect, which is essentially an inconclusive result. By contrast, a comparison between the first and third conditions, *A-Com* and *M-Com*, which is a test of alignment of laterality when orientation is compatible, based on the HDI separation (over 150%) suggests a strong difference. Indeed, the corresponding Bayesian test of this effect yielded a Bayes factor of over 5,000 in favor of an effect. Finally, the HDIs for conditions *M-Com* and *M-Inc* are separated by much less than the approximately 28% indicated in Figs. 7G and H as evidence

for a null effect (separation is about 8%). This outcome provides evidence for a null effect of orientation compatibility when laterality is not aligned. Indeed, the test for a difference between those two conditions yielded $BF_{01} = 4.03$, which is moderate evidence for a null effect.

## Conclusion

We discovered a quadratic exponential relationship between the Bayes factor and the separation of credible intervals in the linear model for a between-subjects design and the linear mixed-effects model for a within-subject design, each with two levels and a known sample size. In the case of a balanced one-way between-subjects design, this relationship is stated in Theorem 1 as an asymptotic relationship. In the more complicated within-subject designs, we have presented simulation studies that demonstrate the relationship between the Bayes factor and within-subject intervals and how it varies across a number of settings. Based on the simulations, we conjecture that an asymptotic result similar to Theorem 1 will hold in the case of within-subject designs. The quadratic exponential relationship is less accurate when the sample size is insufficient or the observed effect size is enormous, placing some practical constraints on the use of this relationship.

Care must be exercised when examining the separation of standard HDIs (as is often done in practice) because these intervals can mask the presence of real effects in within-subject designs. The '*rmBayes*' `R` package (v0.1.15; Wei et al., 2022a) incorporates all the essential functions and employs default priors from *anovaBF* (v0.9.12; Morey & Rouder, 2022) and the NUTS algorithm from `Stan` (v2.21.8; Stan Development Team, 2023). We recommend using JZS-HDI (default method 1) in the *rmHDI* `R` function to implement within-subject Bayesian intervals and assess the strength of evidence for effects by examining the separation between intervals and considering the sample size. We also explored the Monte Carlo sampling error in estimating the Bayes factor. The Bayes factor was computed on a simulated data set of typical dimensions with a larger Monte Carlo sample size (100,000 iterations) than the default (10,000 iterations). Changing the random seed resulted in values ranging from 4.3 to 5.4, with one outlier of 11.7 (see Appendix C). Bayesian credible intervals are relatively more stable, and as we have shown, credible interval separation can be well calibrated to model selections based on the Bayes factor for a given sample size.

The relationship between Bayes factors and credible intervals also appears to hold when the null hypothesis is true. This relationship has been investigated using a broad set of simulations and confirmed under certain conditions through an analytic derivation. Although the relationship is approximately quadratic exponential, we note that its exact form is dependent on and must be interpreted with respect to sample size. This situation can lead to an example of the Jeffreys-Lindley paradox (Bartlett, 1957; Jeffreys, 1935; Lindley, 1957; Wagenmakers & Ly, 2023). For example, one may conceive of a scenario with a compelling Bayes factor in favor of $\mathcal{M}_1$ and a particular separation of the credible intervals. If one were to increase the sample sizes while keeping the degree of separation fixed, the Bayes factor should be less compelling. As the sample size continues to grow, the same separation will ultimately signal strong evidence against the presence of an effect. Practically speaking, however, HDI separation is the distance between two means divided by the twofold interval width, and in order to keep separation fixed as the sample size grows (narrower HDI width), it would be necessary to bring the two population means closer together. This scenario contrasts with an actual experimental situation where the true values of the parameters will not change if the sample size increases. A similar rule that approximates the objective Bayes factors from *p*-values and sample size is discussed in Wagenmakers (2022).

Future work will consider more general settings for the development of the within-subject Bayesian credible interval. The inclusion of Bayesian semiparametric mixed models and exponential family mixed models, including Poisson and binomial regression, will be investigated.

## Appendix A

**Proof of Theorem 1**

After some substitutions, the Pearson Bayes factor in Equation 9 becomes

$$\text{P-BF}_{10} = \frac{\Gamma\left(\frac{a+1}{2}+\gamma\right)\cdot\Gamma\left(\frac{a(n-1)}{2}\right)}{\Gamma\left(\frac{an-1}{2}\right)\cdot\Gamma(1+\gamma)}\left(1+\frac{SS_{\text{B}}}{SS_{\text{W}}}\right)^{\frac{a(n-1)}{2}-1-\gamma}. \tag{A1}$$

By applying Stirling's formula $\Gamma(y+z)\sim y^{z}\Gamma(y)$ as $y\to+\infty$, the gamma ratio in Equation A1 becomes $\frac{\Gamma\left(\frac{a(n-1)}{2}\right)}{\Gamma\left(\frac{an-1}{2}\right)}\sim\left(\frac{an}{2}\right)^{\frac{1-a}{2}}$ as $n\to+\infty$.

We calculated the separation for the standard between-subjects confidence interval in Equation A2 as the absolute value of the difference between two sample means over the twofold interval width. Here, we consider only $a=2$.

$$Sep = |M_{1.}-M_{2.}|\Big/\left(2t^{*}_{1-\frac{\alpha}{2},\,a(n-1)}\sqrt{\frac{SS_{\text{W}}}{n(n-1)a}}\right). \tag{A2}$$

As $n\to+\infty$, $t^{*}_{1-\frac{\alpha}{2},\,a(n-1)}\sim z_{1-\frac{\alpha}{2}}$, the sample means converge to the population means, and $(SS_{\text{B}}/SS_{\text{W}})^2$ is assumed

to be the least significant in the exponential series. $SS_{\mathrm{B}} = n\sum_{i=1}^{a}\left(M_{i\cdot} - M\right)^2$ reduces to $\frac{1}{2}n\left(M_{1\cdot} - M_{2\cdot}\right)^2$ when $a=2$. Plugging Equation A2 in the limit below, we obtain

$$\begin{aligned}
&\lim_{n\to+\infty}\left[\left(1+\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}\right)^{\frac{a(n-1)}{2}-1-\gamma} - \exp\left\{z^2_{1-\frac{\alpha}{2}}\cdot Sep^2\right\}\right]\\
&= \lim_{n\to+\infty}\left[\left(1+\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}\right)^{\frac{a(n-1)}{2}-1-\gamma} - \exp\left\{\frac{1}{2}a(n-1)\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}\right\}\right]\\
&= \lim_{n\to+\infty}\left[\left(1+\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}\right)^{\frac{a(n-1)}{2}-1-\gamma} - \left(1+\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}+O\left(\left(\frac{SS_{\mathrm{B}}}{SS_{\mathrm{W}}}\right)^2\right)\right)^{\frac{a(n-1)}{2}}\right] = 0.
\end{aligned}$$

Hence, the asymptotic approximation of the log Pearson Bayes factor is a quadratic function of the separation of the standard confidence interval for population means in balanced one-way between-subjects designs, as the number of subjects goes to infinity. To check the convergence of the limit, we plot the (solid black) fitted line for the relationship between the log JZS-$BF_{10}$ and squared separation score, along with the (dashed red) analytic line from plugging the known values into the formula of Theorem 1, in Fig. 10. As the sample size increases, the two lines become closer. We still expect some variations between these lines even for very large $n$ because Theorem 1 applies for the Pearson Bayes factor, whereas the quadratic exponential is interpolated for the JZS Bayes factor.

## Appendix B

### R packages to perform Bayesian inference

The '*rmBayes*' package performs Bayesian interval estimation for both the homoscedastic and heteroscedastic cases in either between- or within-subject designs that include a single independent variable. The **Stan**-based **R** source package installation will take a few minutes because models need to be compiled into dynamic shared objects. We recommend using **R** version 4.0.1 or later and installing the pre-compiled binary package so users do not have to worry about **C++** compiler issues. The relevant commands are:

```
> install.packages("rmBayes", type =
"binary")
> library(rmBayes)
```

The *rmHDI* function in '*rmBayes*' provides multiple methods to construct the credible intervals for population means, with each method based on different sets of priors. The default method implements the NUTS algorithm and constructs the within-subject HDI corresponding to the JZS-HDI case in Table 1. More methods documentation can be viewed on GitHub, https://zhengxiaouvic.github.io/rmBayes/. The following example includes a partial data set, a call to the *rmHDI* function, and the resulting output. The partial data set is also shown in wide format.

```
> ## Data are in the long format. 10
subjects. 3 conditions.
> head(recall.long, 2)
Subject Level Response
1 s1 Level1 10
2 s2 Level1 6
> rmHDI(recall.long, whichSubject =
"Subject", whichLevel = "Level", whi-
chResponse = "Response", seed = 277)
#macOS (Apple chip)
$HDI
lower upper
Level1 10.47101 11.59361
Level2 12.39176 13.51436
Level3 13.55086 14.67346
$`posterior means`
Level1 Level2 Level3
11.03231 12.95306 14.11216
$width
[1] 0.5613014
> ## Same data are in the wide format.
> head(recall.wide, 2)
Level1 Level2 Level3
s1 10 13 13
s2 6 8 8
> rmHDI(data.wide= recall.wide, seed
= 277)
```

An alternative method for computing HDIs is possible using the '*BayesFactor*' package, which computes Bayes factors for several experimental designs. The *anovaBF* function can first be used to generate the Bayes factor for a within-subject design.

```
> library(BayesFactor); set.seed(277)
> anovaBF(Response ~ Level + Subject,
data = recall.long, whichRandom = "Sub-
ject", iterations = 100000, progress =
FALSE)
Bayes factor analysis
--------------
[1] Level + Subject : 36469.12 ±0.32%
Against denominator:
Response ~ Subject
---
Bayes factor type: BFlinearModel, JZS
```

Then, Gibbs sampling can obtain parameter estimates from the posterior distribution of the Bayes factor object numerator. Those estimates are plugged into the interval

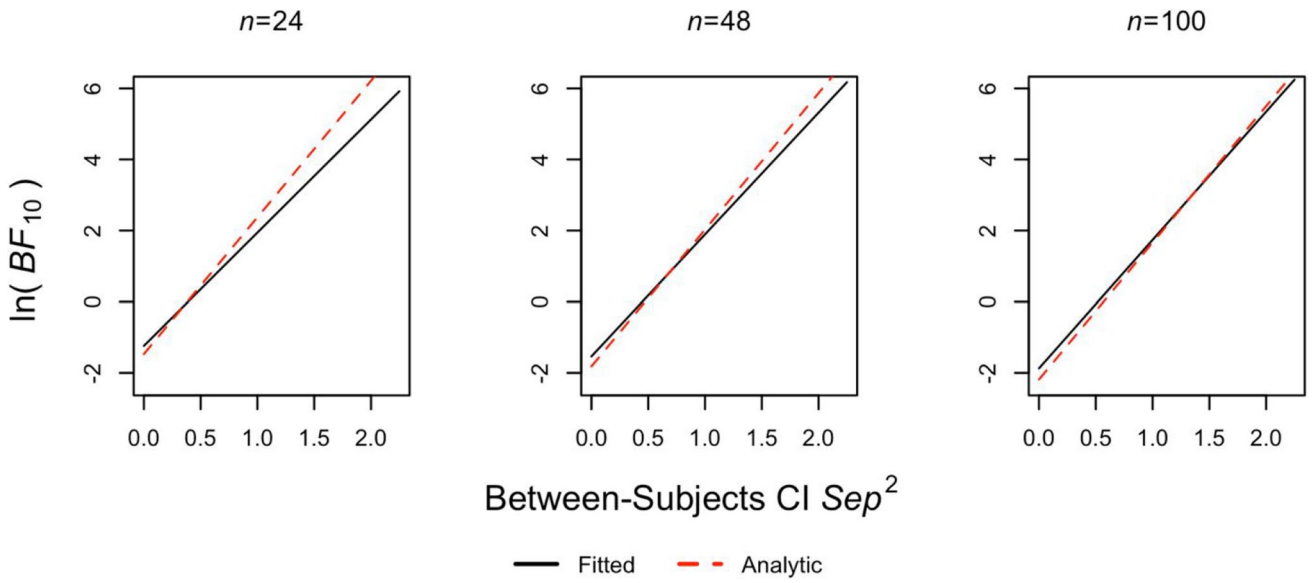


**Fig. 10** Plots of the relationship between the log Bayes factor and the squared separation score of the standard confidence interval for population means in between-subjects designs, with the fitted line in solid black and the analytic line in dashed red

equations for LH- or JZS-HDI in Table 1 to construct the within-subject HDI. This method can be implemented using the **R** code below, which defines a new *anovaHDI* function. Both *anovaHDI* and the default *rmHDI* functions assume all the same priors but use different sampling algorithms for establishing posterior distributions.

```
  > anovaHDI <- function(data, whichSubject, whichLevel, whichResponse, cred, iter) {
  #' input arguments are defined as in the rmHDI function
  n <- length(unique(data[,whichSubject]))
  a <- length(unique(data[,whichLevel]))
  BF <- BayesFactor::anovaBF(as.formula(paste(whichResponse, "~", whichLevel, "+", whichSubject)), data = data, whichRandom = whichSubject, iterations = iter, progress = FALSE)
  chains <- BayesFactor::posterior(BF, iterations = iter, progress = FALSE)
  mu.chains <- chains[,2:(a+1)] + chains[,1]
  widths <- qt((1 + cred) / 2, df = a * (n - 1)) * sqrt(chains["sig2"] / n)
  uprs <- mu.chains + widths
  lwrs <- mu.chains - widths
  matrix(c(colMeans(lwrs),colMeans(uprs)), nrow = a, dimnames = list(paste("Level",1:a), c("lower","upper")))
  }
  > set.seed(277)
  > anovaHDI(recall.long, "Subject", "Level", "Response", .95, 100000)
  lower upper
  Level 1 10.50752 11.64187
  Level 2 12.41498 13.54934
  Level 3 13.56208 14.69644
```

## Appendix C

### Monte Carlo error and a data permutation issue

Users should expect different results if they vary the number of iterations or the random seed used in MCMC. Such variability is referred to as *Monte Carlo error*. We examined Monte Carlo error in computing Bayes factors by applying the *anovaBF* function 500 times (each containing 100,000 MCMC iterations; the default value is 10,000) with different random seeds on the same set of simulated within-subject data. Among these 500 runs, one Bayes factor outlier was as extreme as 11.7, although the vast majority of values ranged from 4.3 to 5.4. R scripts for this and the following examples are available at https://osf.io/x2pvw/.

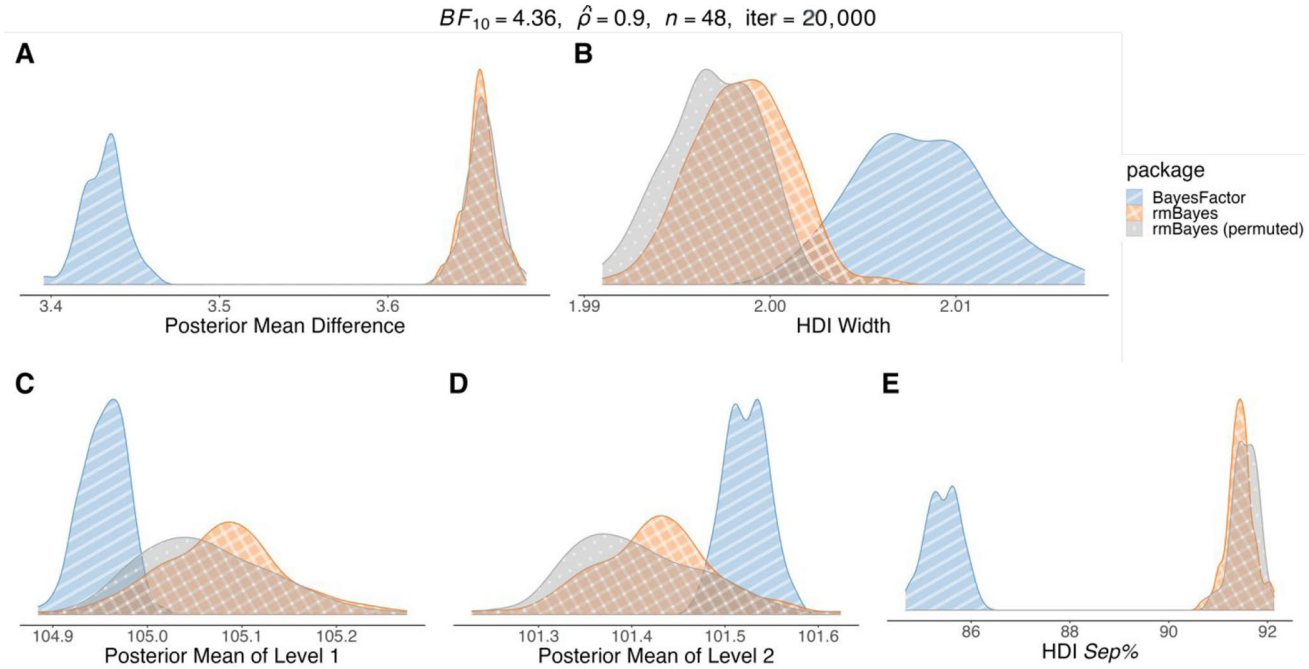


**Fig. 11** Density plots of the simulations from replicating R functions. **A:** posterior mean difference; **B:** HDI width; **C:** posterior mean of one condition; **D:** posterior mean of the other condition; **E:** HDI separation percentage. The same random seed was fixed when investigating the permutation issue, denoted as '*rmBayes (permuted)*', whereas different random seeds were used for investigating the Monte Carlo error in '*BayesFactor*' and '*rmBayes*'

Similarly, Monte Carlo error is associated with the Bayesian interval estimation. We replicated *rmHDI* and *anovaBF* functions 100 times (each containing 20,000 MCMC draws – 2 chains with 10,000 iterations each in *rmHDI*) with different random seeds on the same simulated within-subject data. Furthermore, we visualized the resulting variability of the posterior mean estimates, HDI widths, and HDI separation via density plots. The whole process was repeated several times for data sets having different Bayes factors, correlations between conditions, and sample sizes. One example is exhibited in Fig. 11. In different realizations of the draws from the posterior distribution, it is also worthwhile to note a *data permutation* issue that affects the simulation. That is, the same experimental data are used but permuted by row (e.g., switch Subject 10 up to the second place) or by column (e.g., Level-high and Level-low rather than Level-low and Level-high order). Permutation of the entries in a data file will result in slightly different estimates even if the random seed stays the same. We randomly permuted the data by row but fixed the same random seed when calling *rmHDI* to assess the magnitude of the permutation issue relative to Monte Carlo error. In Fig. 11, two density plots (permuted data but using a constant random seed, or not permuted but varying the random seed) generated from results provided by the *rmHDI* function in the '*rmBayes*' package highly overlap, indicating that permutation of the data produces variability in outcomes of a similar magnitude to setting different random seeds. Moreover, the functions in the '*BayesFactor*' package returned less variability in estimates for posterior means but more variability in estimates for standard error of the mean (thus, interval width) and posterior mean difference, whereas the *rmHDI* performance is quite the opposite. The separation percentage is less variable when calling *rmHDI*, as shown in panel E of Fig. 11. Although the models and priors assumed by the two packages are identical, there may be differences in the actual code implementation, especially for Equations 2 and 4 and MCMC samplers (Gibbs sampling in the *anovaBF* and NUTS in the *rmHDI*), leading to differences in the variable results.

In the *rmHDI* function, the default setting for the argument `permuted` is `TRUE`, meaning the converted wide-format data are first ordered by their column names in alphabetic order. Then, the data are placed in ascending order by the first and second columns.

## Appendix D

### Warning messages regarding sampling and effective Monte Carlo sample size

The Stan website https://mc-stan.org/misc/warnings lists all the potential warnings in running an MCMC. Three common warnings are related to the exceeded maximum tree depth (a concern for long execution time), low bulk effective samples size (ESS, indicating posterior means and medians may be unreliable), and low tail ESS (indicating

posterior variances and tail quantiles may be unreliable). The relevance of these warnings depends on the specific data being analyzed. Visit the website https://osf.io/x2pvw/ for an example.

We suspect that a high correlation between conditions in a within-subject design might result in slower, inefficient sampling due to a computed likelihood with elongated elliptical contours. The latter two warnings indicate that the sampler is moving slowly. After accounting for the correlation across successive draws of the Markov chain sampler, the ESS is low. For example, if the lag-1 autocorrelation of the MCMC sampling output is high (e.g., above .97), then 2,000 iterations can be worth, say fewer than 100 independent draws. The warning disappears with 10,000 iterations because the effective sample size may then be sufficiently high to cross the threshold in `Stan` (it might be an ESS of approximately 500).

**Acknowledgements** We thank Eric-Jan Wagenmakers for bringing to our attention the potential link between the separation of credible intervals and the Jeffreys-Lindley paradox. We are also grateful for an anonymous referee's helpful comments on the likelihood principle, the importance of sample size, and evidence for the null hypothesis.

**Funding** This work was supported by discovery grants to Farouk S. Nathoo (RGPIN-04044-2020) and Michael E. J. Masson (RGPIN-2015-04773) from the Natural Sciences and Engineering Research Council. Farouk S. Nathoo holds a Tier II Canada Research Chair in Biostatistics for Spatial and High-Dimensional Data.

**Data availability** The data used in the analyses are available via the Open Science Framework at https://osf.io/x2pvw/.

**Code availability** All R code is available via the Open Science Framework at https://osf.io/x2pvw/.

## Declarations

**Conflict of interest** The authors declare no conflicts of interest.

## References

Armitage, P., Berry, G., & Matthews, J. N. S. (2002). *Statistical methods in medical research* (4th ed.). Bodmin, UK: Blackwell Science. https://doi.org/10.1002/9780470773666

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44*, 533–534. https://doi.org/10.1093/biomet/44.3-4.533

Bub, D. N., Masson, M. E., & van Noordenne, M. (2021). Motor representations evoked by objects under varying action intentions. *Journal of Experimental Psychology: Human Perception and Performance, 47*, 53–80.

Campbell, H., & Gustafson, P. (2021). re: Linde et al. (2021) - The Bayes factor, HDI-ROPE and frequentist equivalence testing are actually all equivalent. *ArXiv*. 1–22. https://doi.org/10.48550/arXiv.2104.07834

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*, 465–480.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5*, 369–411.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9780203771587

Congdon, P. D. (2019). *Bayesian hierarchical models with applications using R* (2nd ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780429113352

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology, 15*, 226–241.

Craiu, R. V., Gustafson, P., & Rosenthal, J. S. (2022). Reflections on Bayesian inference and Markov chain Monte Carlo. *The Canadian Journal of Statistics, 50*, 1213–1227.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*, 7–29.

Dienes, Z. (2021). Obtaining evidence for no effect. *Collabra. Psychology, 7*, 1–15.

Eich, E. (2014). Business not as usual. *Psychological Science, 25*, 3–6.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE, 11*, 1–12.

Evett, I. W. (1987). Bayesian inference and forensic science: Problems and perspectives. *Journal of the Royal Statistical Society, 36*, 99–105.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Faulkenberry, T. J. (2021). The Pearson Bayes factor: An analytic formula for computing evidential value from minimal summary statistics. *Biometrical Letters, 58*, 1–26.

Faulkenberry, T. J., & Brennan, K. B. (2022). Computing analytic Bayes factors from summary statistics in repeated-measures designs. *ArXiv.*, 1–25. https://doi.org/10.48550/arXiv.2209.08159

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review, 19*, 395–404.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York: Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112.

Heck, D. W. (2019). Accounting for estimation uncertainty and shrinkage in Bayesian within-subject intervals: A comment on Nathoo, Kilshaw, and Masson (2018). *Journal of Mathematical Psychology, 88*, 27–31.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*, 1157–1164.

Hu, C., Wang, F., Guo, J., Song, M., Sui, J., & Peng, K. (2016). The replication crisis in psychological research. *Advances in Psychological Science, 24*, 1504–1518.

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomised block and split-plot designs. *Journal of Educational Statistics, 1*, 69–82.

Jaynes, E. T., & Kempthorne, O. (1976). Confidence intervals vs Bayesian intervals. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, *6b*, 175–257. https://doi.org/10.1007/978-94-010-1436-6_6

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31*, 203–222.

Jeffreys, H. (1936). Further significance tests. *Mathematical Proceedings of the Cambridge Philosophical Society, 32*, 416–445.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 186*, 453–461.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press. https://global.oup.com/academic/product/theory-of-probability-9780198503682

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, D.C.: American Psychological Association. https://doi.org/10.1037/14136-000

Kotz, S., & Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511550683

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London, UK: Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.09999-2

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*, 270–280.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour, 5*, 1282–1291.

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments*. R package version 4.4-0. https://cran.r-project.org/package=ez

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

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*, 410–423.

Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E.-J., & van Ravenzwaaij, D. (2021). Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *Psychological Methods*. 1–16. https://doi.org/10.1037/met0000402

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192. https://doi.org/10.2307/2333251

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476–490.

Lovric, M. M. (2020). Conflicts in Bayesian statistics between inference based on credible intervals and Bayes factors. *Journal of Modern Applied Statistical Methods, 18*, 1–27.

Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (2017). A flexible and efficient hierarchical Bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. *Computational Models of Brain and Behavior*, 467–479. https://doi.org/10.1002/9781119159193.ch34

Ly, A., Raj, A., Etz, A., Marsman, M., Gronau, Q. F., & Wagenmakers, E.-J. (2018). Bayesian reanalyses from summary statistics: a guide for academic consumers. *Advances in Methods and Practices in Psychological Science, 1*, 367–374.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32.

Maruyama, Y., & George, E. I. (2011). Fully Bayes factors with a generalized *g*-prior. *The Annals of Statistics, 39*, 2740–2765.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679–690.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57*, 203–220.

Morey, R. D. (2015a). *Multiple comparisons with BayesFactor, Part 1*. R-Bloggers. https://www.r-bloggers.com/2015/01/multiple-comparisons-with-bayesfactor-part-1/

Morey, R. D. (2015b). *Multiple comparisons with BayesFactor, Part 2 - Order restrictions*. BayesFactor. https://bayesfactor.blogspot.com/2015/01/multiple-comparisons-with-bayesfactor-2.html

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18.

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.4. 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*, 368–378.

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. J. (2018). A better (Bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology, 86*, 1–9.

Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null-hypothesis significance testing for repeated-measures designs. *Journal of Mathematical Psychology, 72*, 144–157.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

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

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22*, 304–321.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55*, 182–186.

Stan Development Team (2023). *RStan: The R interface to Stan*. R package version 2.21.8. https://mc-stan.org/

Urry, H. L., van Reekum, C. M., Johnstone, T., Kalin, N. H., Thurow, M. E., Schaefer, H. S., Jackson, C. A., Frye, C. J., Greischar, L. L., Alexander, A. L., & Davidson, R. J. (2006). Amygdala and ventromedial prefrontal cortex are inversely coupled during regulation of negative affect and predict the diurnal pattern of cortisol secretion among older adults. *Journal of Neuroscience, 26*, 4415–4425.

van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (2022). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *PsyArXiv*. 1-28. 10.31234/osf.io/fb8zn

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 92–114.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779–804.

Wagenmakers, E.-J. (2022). Approximate objective Bayes factors from *p*-values and sample size: The $3p\sqrt{n}$ rule. *PsyArXiv*. 1-50. https://doi.org/10.31234/osf.io/egydq

Wagenmakers, E.-J., Gronau, Q. F., Dablander, F., & Etz, A. (2022). The support interval. *Erkenntnis, 87*, 589–601.

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

Wagenmakers, E.-J., & Ly, A. (2023). History and nature of the Jeffreys-Lindley paradox. *Archive for History of Exact Sciences, 77*, 25–72.

Wang, M., & Liu, G. (2016). A simple two-sample Bayesian *t*-test for hypothesis testing. *The American Statistician, 70*, 195–201.

Wang, M., & Sun, X. (2014). Bayes factor consistency for one-way random effects model. *Communications in Statistics - Theory and Methods, 43*, 5072–5090.

Wei, Z., Nathoo, F. S., & Masson, M. E. J. (2022a). *rmBayes: Performing Bayesian inference for repeated-measures designs*. R package version 0.1.15. https://cran.r-project.org/package=rmBayes

Wei, Z., Yang, A., Rocha, L., Miranda, M. F., & Nathoo, F. S. (2022b). A review of Bayesian hypothesis testing and its practical implementations. *Entropy, 24*, 1–15.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*, 291–298.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística Y de Investigación Operativa, 31*, 585–603.