

¹ Indices of Effect Existence and Significance in the Bayesian Framework

² Dominique Makowski¹, Mattan S. Ben-Shachar², Daniel Lüdecke³, & S.H. Annabel Chen^{1,4}

³ ¹ Nanyang Technological University, Singapore

⁴ ² Ben-Gurion University of the Negev, Israel

⁵ ³ University Medical Center Hamburg-Eppendorf, Germany

⁶ ⁴ Centre for Research and Development in Learning (CRADLE), Singapore

⁷ Author Note

⁸ Daniel Lüdecke and S.H. Annabel Chen share senior authorship.

⁹ Correspondence concerning this article should be addressed to Dominique Makowski,
¹⁰ HSS 04-18, 48 Nanyang Avenue, Singapore. E-mail: dmakowski@ntu.edu.sg

11

Abstract

12 Turmoil has engulfed psychological science. Causes and consequences of the
13 reproducibility crisis are in dispute. With the hope of addressing some of its aspects,
14 Bayesian methods are gaining increasing attention in psychological science. Some of their
15 advantages, as opposed to the frequentist framework, are the ability to describe parameters
16 in probabilistic terms and explicitly incorporate prior knowledge about them into the
17 model. These issues are crucial in particular regarding the current debate about statistical
18 significance. Bayesian methods are not necessarily the only remedy against incorrect
19 interpretations or wrong conclusions, but there is an increasing agreement that they are
20 one of the keys to avoid such fallacies. Nevertheless, its flexible nature is its power and
21 weakness, for there is no agreement about what indices of “significance” should be
22 computed or reported. This lack of a consensual index or guidelines, such as the frequentist
23 *p*-value, further contributes to the unnecessary opacity that many non-familiar readers
24 perceive in Bayesian statistics. Thus, this study describes and compares several Bayesian
25 indices, provide intuitive visual representation of their “behavior” in relationship with
26 common sources of variance such as sample size, magnitude of effects and also frequentist
27 significance. The results contribute to the development of an intuitive understanding of the
28 values that researchers report, allowing to draw sensible recommendations for Bayesian
29 statistics description, critical for the standardization of scientific reporting.

30

Keywords: Bayesian, significance, NHST, *p*-value, Bayes factors

31

Word count: 6293

32 Indices of Effect Existence and Significance in the Bayesian Framework

33 **Introduction**

34 The Bayesian framework is quickly gaining popularity among psychologists and
35 neuroscientists (Andrews & Baguley, 2013). Reasons to prefer this approach are reliability,
36 better accuracy in noisy data, better estimation for small samples, less proneness to type I
37 errors, the possibility of introducing prior knowledge into the analysis and the intuitiveness
38 and straightforward interpretation of results (Dienes & Mcclatchie, 2018; Etz &
39 Vandekerckhove, 2016; Kruschke, 2010; Kruschke, Aguinis, & Joo, 2012; Wagenmakers et
40 al., 2018; Wagenmakers, Morey, & Lee, 2016). On the other hand, the frequentist approach
41 has been associated with the focus on *p*-values and null hypothesis significance testing
42 (NHST). The misinterpretation and misuse of *p*-values, so called “p-hacking” (Simmons,
43 Nelson, & Simonsohn, 2011), has been shown to critically contribute to the reproducibility
44 crisis in psychological science (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014;
45 Szucs & Ioannidis, 2016). Not only are *p*-values used to draw inappropriate inferences from
46 noisy data, but even when used properly, effects are drastically overestimated, sometimes
47 even in the wrong direction, when estimation is tied to statistical significance in highly
48 variable data (Gelman, 2018). In response, there is a general agreement that the
49 generalization and utilization of the Bayesian framework is one way of overcoming these
50 issues (Benjamin et al., 2018; Etz & Vandekerckhove, 2016; Halsey, 2019; Marasini, Quatto,
51 & Ripamonti, 2016; Maxwell, Lau, & Howard, 2015; Wagenmakers et al., 2017).

52 The tenacity and resilience of the *p*-value as an index of significance is remarkable,
53 despite the long-lasting criticism and discussion about its misuse and misinterpretation
54 (Anderson, Burnham, & Thompson, 2000; Cohen, 2016; Fidler, Thomason, Cumming,
55 Finch, & Leeman, 2004; Finch et al., 2004; Gardner & Altman, 1986). This endurance
56 might be informative on how such indices, and the accompanying heuristics applied to
57 interpret them (e.g., assigning thresholds like .05, .01 and .001 to certain levels of

58 significance), are useful and necessary for researchers to gain an intuitive (although
59 possibly simplified) understanding of the interactions and structure of their data.
60 Moreover, the utility of such an index is most salient in contexts where decisions must be
61 made and rationalized (e.g., in medical settings). Unfortunately, these heuristics can
62 become severely rigidified, and meeting significance has become a goal unto itself rather
63 than a tool for understanding the data (Cohen, 2016; Kirk, 1996). This is particularly
64 problematic given that *p*-values can only be used to reject the null hypothesis and not to
65 accept it as true, because a statistically non-significant result does not mean that there is
66 no difference between groups or no effect of a treatment (Amrhein, Greenland, & McShane,
67 2019; Wagenmakers, 2007).

68 While significance testing (and its inherent categorical interpretation heuristics)
69 might have its place as a complementary perspective to effect estimation, it does not
70 preclude the fact that drastic improvements are needed. For instance, one possible advance
71 could focus on improving the mathematical understanding (e.g., through a new simpler
72 index) of the values being used (as opposed to the obscure mathematical definition of the
73 *p*-value, contributing to its common misinterpretation). Another improvement could be
74 found in providing an intuitive understanding (e.g., by visual means) of the behavior of the
75 indices in relationship with main sources of variance, such as sample size, noise or effect
76 presence. Such better overall understanding of the indices would hopefully act as a barrier
77 against their mindless reporting by allowing the users to nuance the interpretations and
78 conclusions that they draw.

79 The Bayesian framework offers several alternative indices for the *p*-value. To better
80 understand these indices, it is important to point out one of the core differences between
81 Bayesian and frequentist methods. From a frequentist perspective, the effects are fixed (but
82 unknown) and data are random. On the other hand, instead of having single estimates of
83 some “true effect” (for instance, the “true” correlation between *x* and *y*), Bayesian methods
84 compute the probability of different effects values *given* the observed data (and some prior

85 expectation), resulting in a distribution of possible values for the parameters, called the
86 posterior distribution. The description of the posterior distribution (e.g., through its
87 centrality, dispersion, etc.) allows to draw conclusions from Bayesian analyses.

88 Bayesian “significance” testing indices could be roughly grouped into three
89 overlapping categories: Bayes factors, posterior indices and Region of Practical Equivalence
90 (ROPE)-based indices. Bayes factors are a family of indices of relative evidence of one
91 model over another (e.g., the null *vs.* the alternative hypothesis; Jeffreys, 1998; Ly,
92 Verhagen, & Wagenmakers, 2016). They provide many advantages over the *p*-value by
93 having a straightforward interpretation as well as allowing to quantify evidence in favor of
94 the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). However, its use for parameters
95 description in complex models is still a matter of debate (Heck, 2019; Wagenmakers,
96 Lodewyckx, Kuriyal, & Grasman, 2010), being highly dependent on the specification of
97 priors (Etz, Haaf, Rouder, & Vandekerckhove, 2018; Kruschke & Liddell, 2018). On the
98 contrary, “posterior indices” reflect objective characteristics of the posterior distribution,
99 for instance the proportion of strictly positive values. While the simplicity of their
100 computation and interpretation is an asset, it might also limit the information that they
101 provide. Finally, ROPE-based indices are related to the redefinition of the null hypothesis
102 from the classic point-null hypothesis to a range of values considered negligible or too small
103 to be of any practical relevance (the Region of Practical Equivalence - ROPE; Kruschke,
104 2014; Lakens, 2017; Lakens, Scheel, & Isager, 2018), usually spread equally around 0 (e.g.,
105 [-0.1; 0.1]). It is interesting to note that this perspective unites significance testing with the
106 focus on effect size (involving a discrete separation between at least two categories:
107 negligible and non-negligible), which finds an echo in recent statistical recommendations
108 (Ellis & Steyn, 2003; Simonsohn, Nelson, & Simmons, 2014; Sullivan & Feinn, 2012).

109 Despite the richness provided by the Bayesian framework and the availability of
110 multiple indices, no consensus has yet emerged on which ones to be used. Literature
111 continues to bloom in a raging debate, often polarized between proponents of the Bayes

factor as the supreme index and its detractors (Robert, 2014, 2016; Spanos, 2013; Wagenmakers, Lee, Rouder, & Morey, 2019), with strong theoretical arguments being developed on both sides. Yet no practical, empirical and direct comparison between these indices has been done. This might be a deterrent for scientists interested in adopting the Bayesian framework. Moreover, this grey area can increase the difficulty of readers or reviewers unfamiliar with the Bayesian framework to follow the assumptions and conclusions, which could in turn generate unnecessary doubt upon an entire study. While we think that such indices of significance and their interpretation guidelines (in the form of rules of thumb) are useful in practice, we also strongly believe that they should be accompanied with the understanding of their “behavior” in relationship with major sources of variance, such as sample size, noise or effect presence. This knowledge is important for people to implicitly and intuitively appraise the meaning and implication of the mathematical values they report. Such an understanding could prevent the crystallization of the possible heuristics and categories derived from such indices, as has unfortunately occurred for the *p*-values.

Thus, based on the simulation of linear and logistic regressions (arguably some of the most widely used models in the psychological sciences), the present work aims at comparing several indices of effect “significance”, provide visual representations of the “behavior” of such indices in relationship with sample size, noise and effect presence, as well as their relationship to frequentist *p*-values (an index which, beyond its many flaws, is well known and could be used as a reference for Bayesian neophytes), and finally draw recommendations for Bayesian statistics reporting.

134

Methods

135 Data Simulation

136 We simulated datasets suited for linear and logistic regression and started by
137 simulating an independent, normally distributed x variable (with mean 0 and SD 1) of a
138 given sample size. Then, the corresponding y variable was added, having a perfect
139 correlation (in the case of data for linear regressions) or as a binary variable perfectly
140 separated by x . The case of no effect was simulated by creating a y variable that was
141 independent of (i.e. not correlated to) x . Finally, a Gaussian noise was added to the x
142 variable (the error).

143 The simulation aimed at modulating the following characteristics: *outcome type*
144 (linear or logistic regression), *sample size* (from 20 to 100 by steps of 10), *null hypothesis*
145 (original regression coefficient from which data is drawn prior to noise addition, 1 -
146 presence of “true” effect, or 0 - absence of “true” effect) and *noise* (Gaussian noise applied
147 to the predictor with SD uniformly spread between 0.666 and 6.66, with 1000 different
148 values), which is directly related to the absolute value of the coefficient (i.e., the effect
149 size). We generated a dataset for each combination of these characteristics, resulting in a
150 total of 36,000 (2 model types * 2 presence/absence of effect * 9 sample sizes * 1,000 noise
151 variations) datasets. The code used for data generation is available on GitHub
152 (https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian/data). Note that it takes usually several days/weeks for the generation to
153 complete.

155 Indices

156 For each of these datasets, Bayesian and frequentist regressions were fitted to predict
157 y from x as a single unique predictor. We then computed the following seven indices from
158 all simulated models (see **Figure 1**), related to the effect of x .

159 **Frequentist *p*-value.** This was the only index computed by the frequentist version
160 of the regression. The *p*-value represents the probability that for a given statistical model,
161 when the null hypothesis is true, the effect would be greater than or equal to the observed
162 coefficient (Wasserstein, Lazar, & others, 2016).

163 **Probability of Direction (*pd*).** The *Probability of Direction* (*pd*) varies between
164 50% and 100% and can be interpreted as the probability that a parameter (described by its
165 posterior distribution) is strictly positive or negative (whichever is the most probable). It is
166 mathematically defined as the proportion of the posterior distribution that is of the
167 median's sign (Makowski, Ben-Shachar, & Lüdecke, 2019).

168 **MAP-based *p*-value.** The *MAP-based p-value* is related to the odds that a
169 parameter has against the null hypothesis (Mills, 2017; Mills & Parent, 2014). It is
170 mathematically defined as the density value at 0 divided by the density at the Maximum A
171 Posteriori (MAP), *i.e.*, the equivalent of the mode for continuous distributions.

172 **ROPE (95%).** The *ROPE (95%)* refers to the percentage of the 95% HDI that lies
173 within the ROPE. As suggested by Kruschke (2014), the Region of Practical Equivalence
174 (ROPE) was defined as range from -0.1 to 0.1 for linear regressions and its equivalent, -0.18
175 to 0.18, for logistic models (based on the $\pi/\sqrt{3}$ formula to convert log odds ratios to
176 standardized differences; Cohen, 1988).

177 **ROPE (full).** The *ROPE (full)* is similar to *ROPE (95%)*, with the exception that
178 it refers to the percentage of the *whole* posterior distribution that lies within the ROPE.

179 **Bayes factor (*vs.* 0).** The Bayes Factor (*BF*) used here is based on prior and
180 posterior distributions of a single parameter. In this context, the Bayes factor indicates the
181 degree by which the mass of the posterior distribution has shifted further away from or
182 closer to the null value (0), relative to the prior distribution, thus indicating if the null
183 hypothesis has become less or more likely given the observed data. The *BF* was computed
184 as a Savage-Dickey density ratio, which is also an approximation of a Bayes factor

¹⁸⁵ comparing the marginal likelihoods of the model against a model in which the tested
¹⁸⁶ parameter has been restricted to the point-null (Wagenmakers et al., 2010).

¹⁸⁷ **Bayes factor (*vs.* ROPE).** The *Bayes factor* (*vs.* ROPE) is similar to the *Bayes*
¹⁸⁸ *factor* (*vs.* 0), but instead of a point-null, the null hypothesis is a range of negligible values
¹⁸⁹ (defined here same as for the ROPE indices). The *BF* was computed by comparing the
¹⁹⁰ prior and posterior odds of the parameter falling within vs. outside the ROPE (see
¹⁹¹ *Non-overlapping Hypotheses* in Morey & Rouder, 2011). This measure is closely related to
¹⁹² the *ROPE (full)*, as it can be formally defined as the ratio between the *ROPE (full)* odds
¹⁹³ for the posterior distribution and the *ROPE (full)* odds for the prior distribution:

$$BF_{rope} = \frac{odds(ROPE_{\text{full posterior}})}{odds(ROPE_{\text{full prior}})}$$

¹⁹⁴ **Data Analysis**

¹⁹⁵ In order to achieve the two-fold aim of this study; 1) comparing Bayesian indices and
¹⁹⁶ 2) provide visual guides for an intuitive understanding of the numeric values in relation to
¹⁹⁷ a known frame of reference (the frequentist *p*-value), we will start by 1) presenting the
¹⁹⁸ relationship between these indices and main sources of variance, such as sample size, noise
¹⁹⁹ and null hypothesis (true if absence of effect, false if presence of effect). We will then 2)
²⁰⁰ compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds
²⁰¹ (.05, .01, .001). Finally, we will show the mutual relationship between 3 recommended
²⁰² Bayesian candidates. Taken together, these results will help us outline guides to ease the
²⁰³ reporting and interpretation of the indices.

²⁰⁴ In order to provide an intuitive understanding of values, data processing will focus on
²⁰⁵ creating clear visual figures to help the user grasp the patterns and variability that exists
²⁰⁶ when computing the investigated indices. Nevertheless, we decided to also mathematically
²⁰⁷ test our claims in cases where the graphical representation begged for a deeper

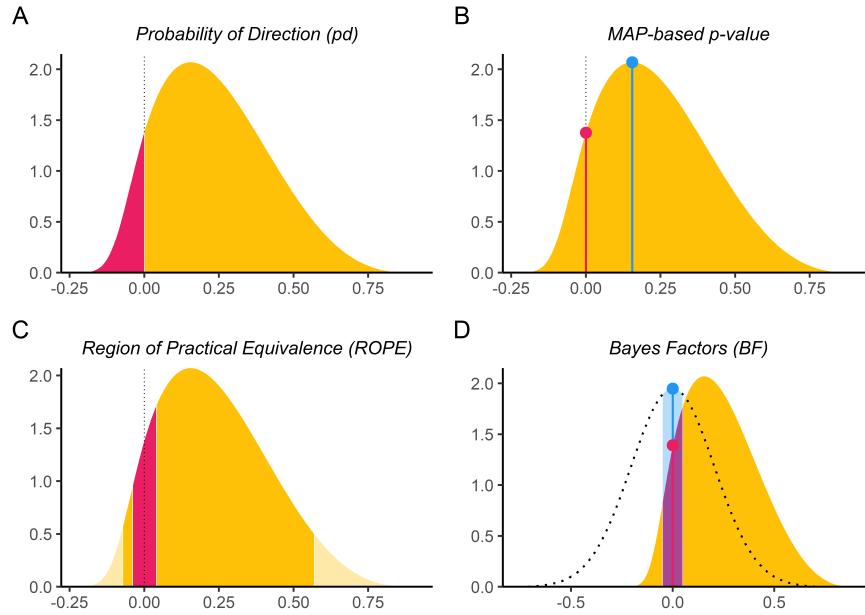


Figure 1. Bayesian indices of effect existence and significance. (A) The Probability of Direction (*pd*) is defined as the proportion of the posterior distribution that is of the median's sign (the size of the yellow area relative to the whole distribution). (B) The MAP-based *p*-value is defined as the density value at 0, - the height of the red lollipop, divided by the density at the Maximum A Posteriori (MAP), - the height of the blue lollipop. (C) The percentage in ROPE corresponds to the red area relative to the distribution (with or without tails for ROPE (*full*) and ROPE (*95%*), respectively). (D) The Bayes factor (vs. 0) corresponds to the point-null density of the prior (the blue lollipop on the dotted distribution) divided by that of the posterior (the red lollipop on the yellow distribution), and the Bayes factor (vs. ROPE) is calculated as the odds of the prior falling within vs. outside the ROPE (the blue area on the dotted distribution) divided by that of the posterior (the red area on the yellow distribution).

208 investigation. Thus, we fitted two regression models to assess the impact of sample size and
 209 noise, respectively. To ensure that any differences between the indices are not due to
 210 differences in their scale or distribution, we converted all indices to the same scale by
 211 normalizing the indices between 0 and 1 (note that *BFs* were transformed to posterior

212 probabilities, assuming uniform prior odds) and reversing the p -values, the MAP-based
213 p -values and the ROPE indices so that a higher value corresponds to stronger “significance”.

214 The statistical analyses were conducted using R (R Core Team, 2019). Computations
215 of Bayesian models were done using the *rstanarm* package (Goodrich, Gabry, Ali, &
216 Brilleman, 2019), a wrapper for Stan probabilistic language (Carpenter et al., 2017). We
217 used Markov Chain Monte Carlo sampling (in particular, Hamiltonian Monte Carlo;
218 Gelman et al., 2014) with 4 chains of 2000 iterations, half of which used for warm-up.
219 Mildly informative priors (a normal distribution with mean 0 and SD 1) were used for the
220 parameter in all models. The Bayesian indices were calculated using the *bayestestR*
221 package (Makowski et al., 2019).

222

Results

223 **Impact of Sample Size**

224 **Figure 2** shows the sensitivity to sample size of the indices. The p -value, the pd and
225 the MAP-based p -value are sensitive to sample size only in case of the presence of a true
226 effect (when the null hypothesis is false). When the null hypothesis is true, all three indices
227 are unaffected by sample size. In other words, these indices reflect the amount of observed
228 evidence (the sample size) for the presence of an effect (i.e., against the null hypothesis
229 being true), but not for the absence of an effect. The *ROPE* indices, however, appear as
230 strongly modulated by the sample size when there is no effect, suggesting their sensitivity
231 to the amount of evidence for the absence of effect. Finally, the figure suggests that *BFs*
232 are sensitive to sample size for both presence and absence of true effect.

233 Consistently with **Figure 2**, the model investigating the sensitivity of sample size on
234 the different indices suggests that *BF* indices are sensitive to sample size both when an
235 effect is present (null hypothesis is false) and absent (null hypothesis is true). *ROPE*
236 indices are particularly sensitive to sample size when the null hypothesis is true, while

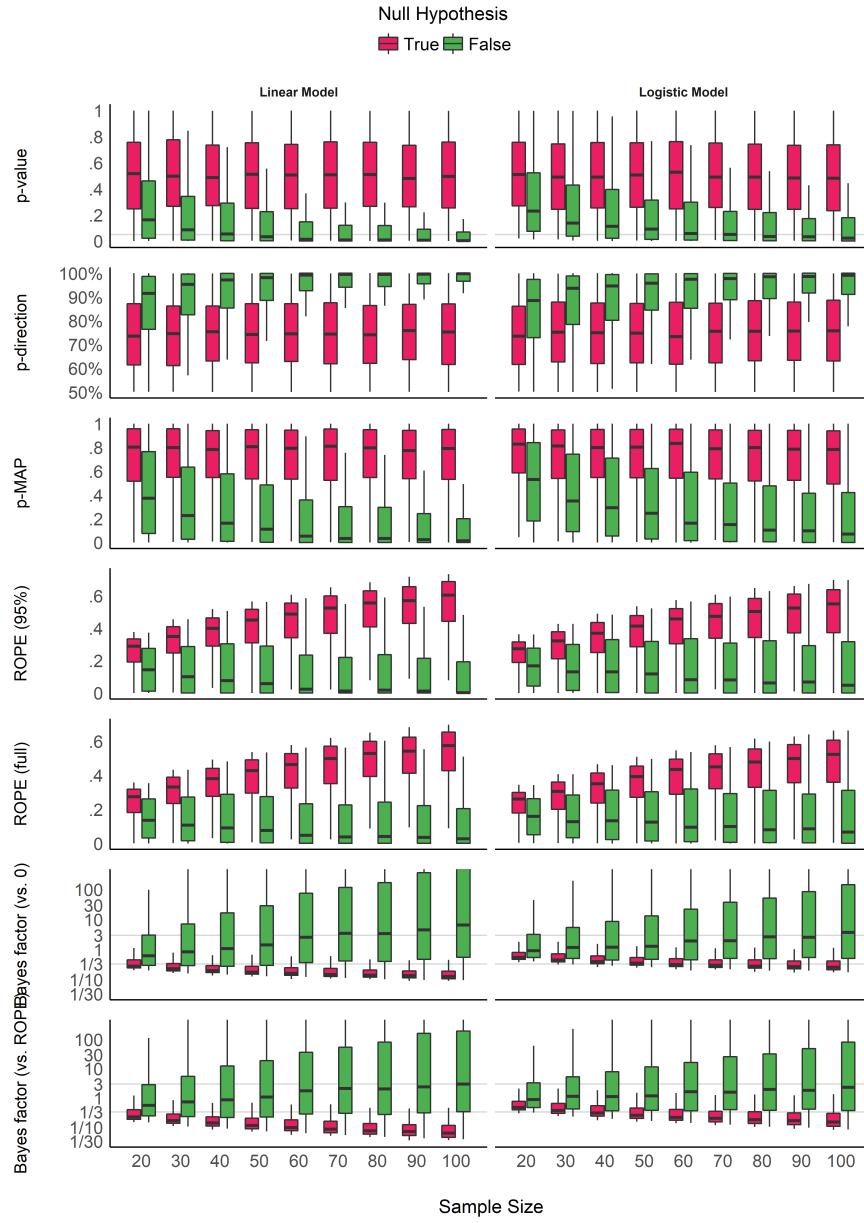


Figure 2. Impact of Sample Size on the different indices, for linear and logistic models, and when the null hypothesis is true or false. Grey vertical lines for $*p*$ -values and Bayes factors represent commonly used thresholds.

²³⁷ p -value, pd and MAP-based p -value are only sensitive to sample size when the null

²³⁸ hypothesis is false, in which case they are more sensitive than $ROPE$ indices. These
²³⁹ findings can be related to the concept of consistency: as the number of data points

Table 1

Sensitivity to sample size. This table shows the standardized coefficient between the sample size and the value of each index, adjusted for error, and stratified by model type and presence of true effect. The stronger the coefficient is, the stronger the relationship with sample size.

Index	Linear Models /	Linear Models /	Logistic Models /	Logistic Models /
	Presence of Effect	Absence of Effect	Presence of Effect	Absence of Effect
p-value	0.17	0.01	0.16	0.02
p-direction	0.17	0.01	0.15	0.02
p-MAP	0.24	0.00	0.24	0.03
ROPE (95%)	0.03	0.36	0.01	0.31
ROPE (full)	0.03	0.36	0.02	0.31
Bayes factor (vs. 0)	0.20	0.12	0.12	0.14
Bayes factor (vs. ROPE)	0.15	0.14	0.08	0.18

²⁴⁰ increases, the statistic converges toward some “true” value. Here, we observe that *p*-value,
²⁴¹ *pd* and the MAP-based *p*-value are consistent only when the null hypothesis is false. In
²⁴² other words, as sample size increases, they tend to reflect more strongly that the effect is
²⁴³ present. On the other hand, *ROPE* indices appear as consistent when the effect is absent.
²⁴⁴ Finally, *BFs* are consistent both when the effect is absent and when it is present. Note also
²⁴⁵ that *BF* (*vs. ROPE*), compared to *BF* (*vs. 0*), is more sensitive to sample size when the null
²⁴⁶ hypothesis is true, and *ROPE* (*full*) is overall slightly more consistent than *ROPE* (95%).

²⁴⁷ Impact of Noise

²⁴⁸ **Figure 3** shows the indices’ sensitivity to noise. Unlike the patterns of sensitivity to
²⁴⁹ sample size, the indices display more similar patterns in their sensitivity to noise (or
²⁵⁰ magnitude of effect). All indices are unidirectional impacted by noise: as noise increases,
²⁵¹ the observed coefficients decrease in magnitude, and the indices become less “pronounced”

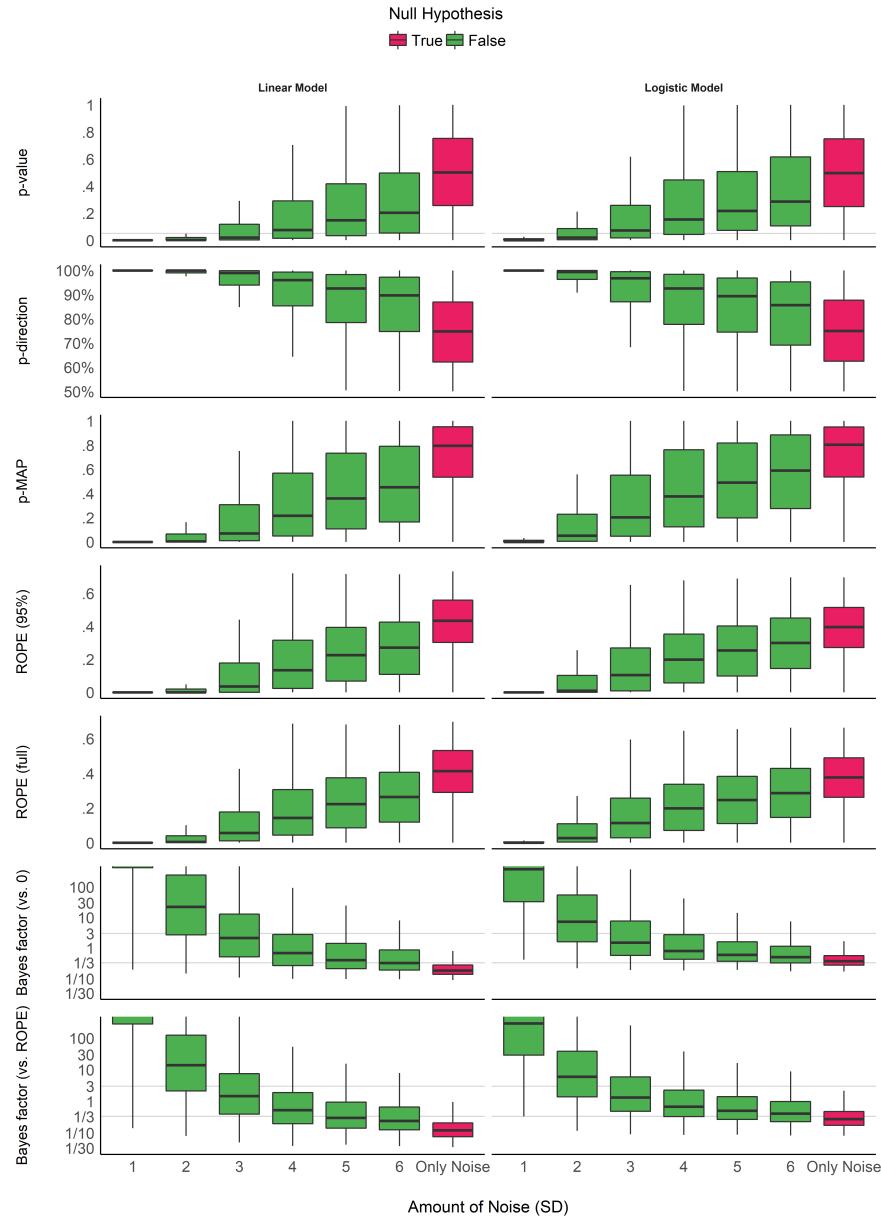


Figure 3. Impact of Noise. The noise corresponds to the standard deviation of the Gaussian noise that was added to the generated data. It is related to the magnitude the parameter (the more noise there is, the smaller the coefficient). Grey vertical lines for $*p*$ -values and Bayes factors represent commonly used thresholds. The scale is capped for the Bayes factors as these extend to infinity.

Table 2

Sensitivity to noise. This table shows the standardized coefficient between noise and the value of each index when the true effect is present, adjusted for sample size and stratified by model type. The stronger the coefficient is, the stronger the relationship with noise.

Index	Linear Models /	Logistic Models /
	Presence of Effect	Presence of Effect
p-value	0.35	0.40
p-direction	0.36	0.40
p-MAP	0.55	0.60
ROPE (95%)	0.45	0.45
ROPE (full)	0.46	0.45
Bayes factor (vs. 0)	0.79	0.65
Bayes factor (vs. ROPE)	0.81	0.67

252 (respectively to their direction). However, it is interesting to note that the variability of
 253 the indices seems differently impacted by noise. For the *p*-values, the *pd* and the ROPE
 254 indices, the variability increases as the noise increases. In other words, small variation in
 255 small observed coefficients can yield very different values. On the contrary, the variability
 256 of BFs decreases as the true effect tends toward 0. For the MAP-based *p*-value, the
 257 variability appears to be the highest for moderate amount of noise. This behavior seems
 258 consistent across model types.

259 Consistently with **Figure 3**, the model investigating the sensitivity of noise when an
 260 effect is present (as there is only noise in the absence of effect), adjusted for sample size,
 261 suggests that BFs (especially *vs.* ROPE), followed by the MAP-based *p*-value and
 262 percentages in ROPE, are the most sensitive to noise. As noise is a proxy of effect size
 263 (linearly related to the absolute value of the coefficient of the parameter), this result
 264 highlights the fact that these indices are sensitive to the magnitude of the effect. For
 265 example, as noise increases, evidence for an effect becomes weak, and data seems to support

the absence of an effect (or at the very least the presence of a negligible effect), which is reflected in *BFs* being consistently smaller than 1. On the other hand, as the *p*-value and the *pd* quantify evidence only for the presence of an effect, as noise increases, they are become more dependent on larger sample size to be able to detect the presence of an effect.

Relationship with the frequentist *p*-value

Figure 4 suggests that the *pd* has a 1:1 correspondence with the frequentist *p*-value (through the formula $p_{two-sided} = 2 * (1 - p_d)$). *BF* indices still appear as having a severely non-linear relationship with the frequentist index, mostly due to the fact that smaller *p*-values correspond to stronger evidence in favor of the presence of an effect, but the reverse is not true. *ROPE*-based percentages appear to be only weakly related to *p*-values. Critically, their relationship seems to be strongly dependent on sample size.

Figure 5 shows equivalence between *p*-value thresholds (.1, .05, .01, .001) and the Bayesian indices. As expected, the *p*-direction has the sharpest thresholds (95%, 97.5%, 99.5% and 99.95%, respectively). For logistic models, these threshold points appear as more conservative (i.e., Bayesian indices have to be more “pronounced” to reach the same level of significance). This sensitivity to model type is the strongest for *BFs* (which is possibly related to the difference in the prior specification for these two types of models).

Relationship between ROPE (full), pd and BF (vs. ROPE)

Figure 6 suggests that the relationship between the *ROPE (full)* and the *pd* might be strongly affected by the sample size, and subject to differences across model types. This seems to echo the relationship between *ROPE (full)* and *p*-value, the latter having a 1:1 correspondence with *pd*. On the other hand, the *ROPE (full)* and the *BF (vs. ROPE)* seem very closely related within the same model type, reflecting their formal relationship (see definition of *BF (vs. ROPE)* above). Overall, these results help to demonstrate *ROPE*

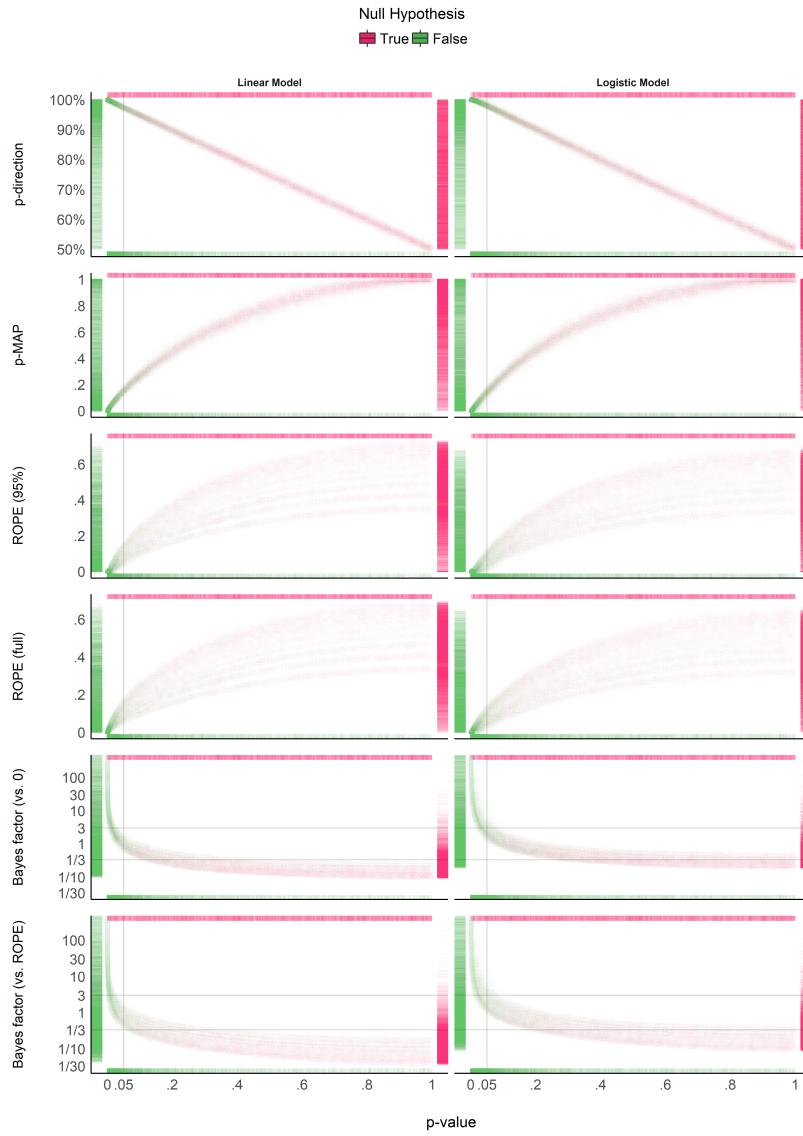


Figure 4. Relationship with the frequentist $*p*$ -value. In each plot, the $*p*$ -value densities are visualized by the marginal top (absence of true effect) and bottom (presence of true effect) markers, whereas on the left (presence of true effect) and right (absence of true effect), the markers represent the density of the index of interest. Different point shapes, representing different sample sizes, specifically illustrate its impact on the percentages in ROPE, for which each "curve line" is associated with one sample size (the bigger the sample size, the higher the percentage in ROPE).

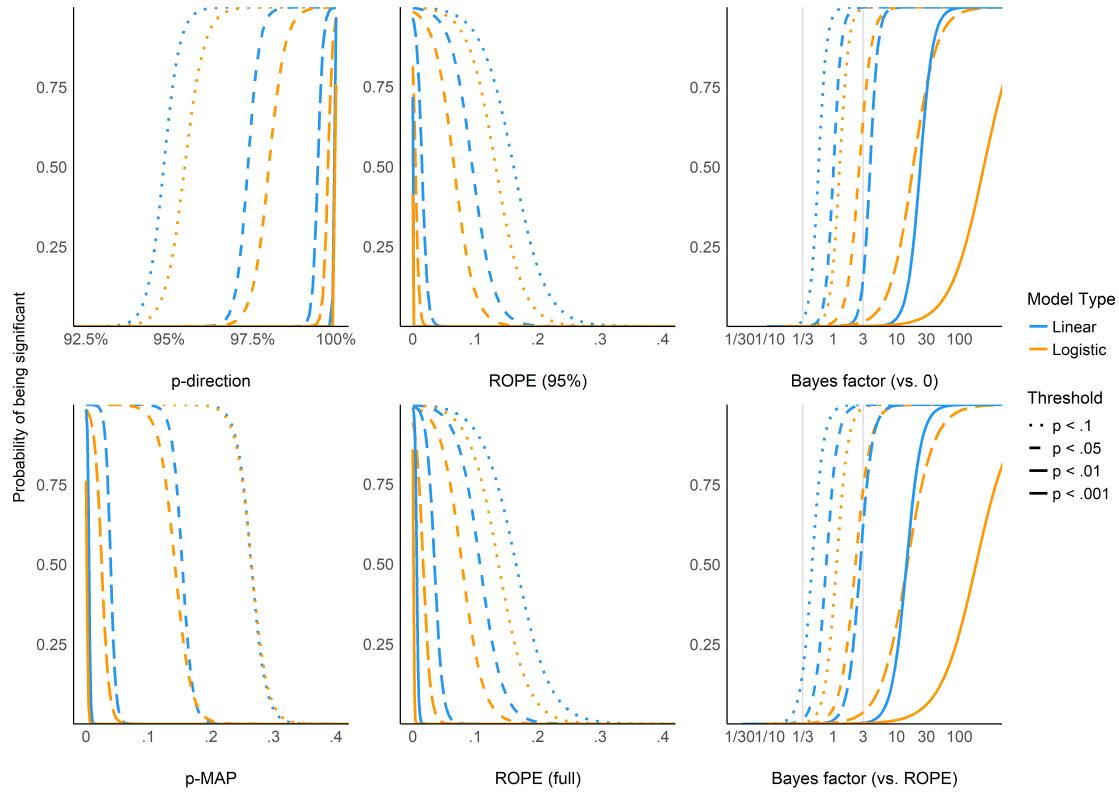


Figure 5. The probability of reaching different $*p*$ -value based significance thresholds (.1, .05, .01, .001 for solid, long-dashed, short-dashed and dotted lines, respectively) for different values of the corresponding Bayesian indices.

(full) and BF (vs. $ROPE$)’s consistency both in case of presence and absence of a true effect, whereas the pd , being equivalent to the p -value, is only consistent when the true effect is absent.

293

Discussion

Based on the simulation of linear and logistic models, the present work aimed at comparing several Bayesian indices of effect “significance” (see **Table 3**), providing visual representations of the “behavior” of such indices in relationship with important sources of variance such as sample size, noise and effect presence, as well as comparing them with the well-known and widely used frequentist p -value and its arbitrary interpretation thresholds.

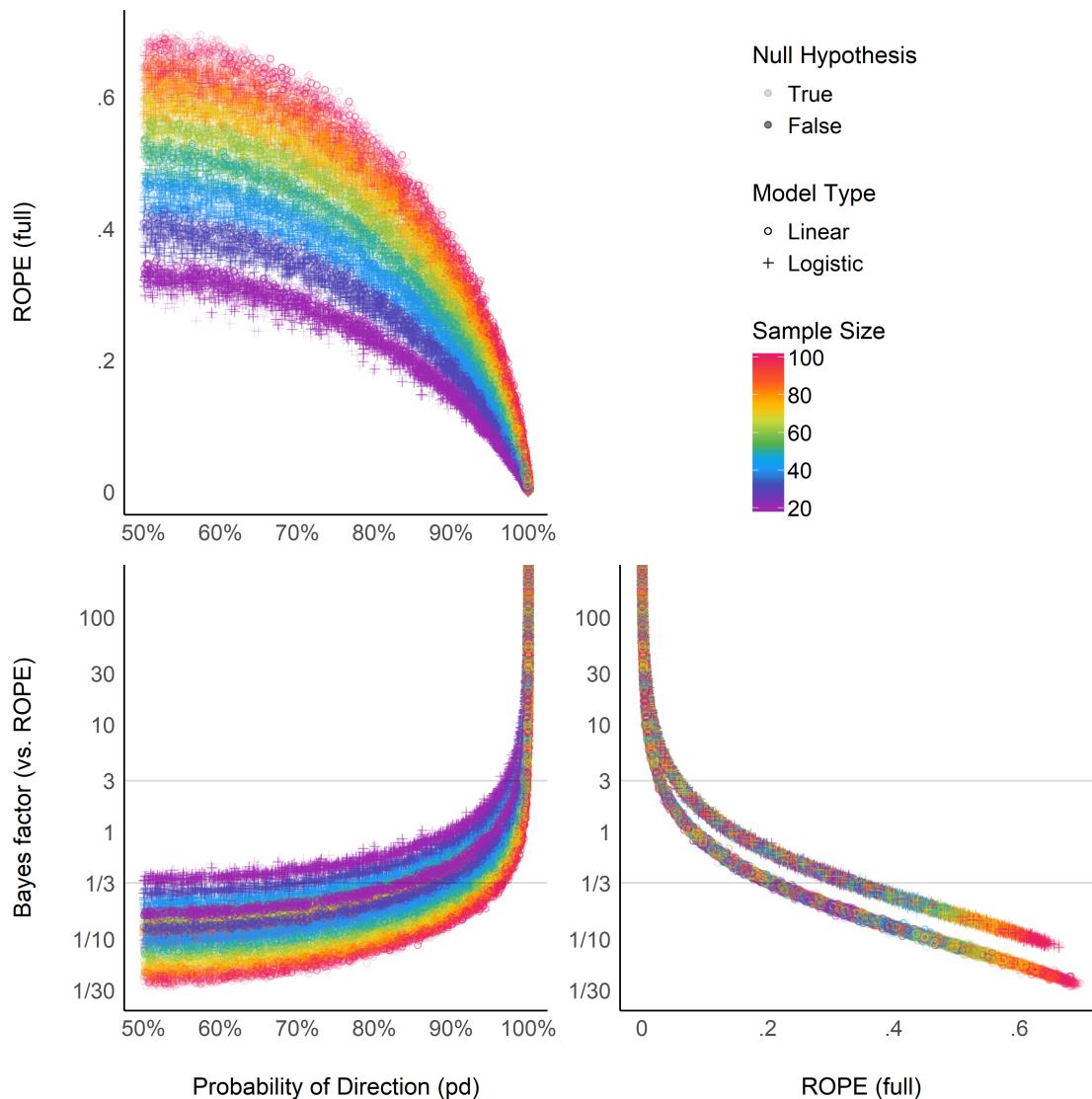


Figure 6. Relationship between three Bayesian indices: The Probability of Direction (*pd*), the percentage of the full posterior distribution in the ROPE, and the Bayes factor (*vs.* ROPE).

The results tend to suggest that the investigated indices could be separated into two categories. The first group, including the pd and the MAP-based p -value, presents similar properties to those of the frequentist p -value: they are sensitive to the amount of evidence for the alternative hypothesis only (i.e., when an effect is truly present). In other words, these indices are not able to reflect the amount of evidence in favor of the null hypothesis

304 (Rouder & Morey, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). A high value
305 suggests that the effect exists, but a low value indicates *uncertainty* about its existence
306 (but not certainty that it is non-existent). The second group, including ROPE and Bayes
307 factors, seem sensitive to both presence and absence of effect, accumulating evidence as the
308 sample size increases. However, the ROPE seems particularly suited to provide evidence in
309 favor of the null hypothesis. Consistently with this, combining Bayes factors with the
310 ROPE (*BF vs.* ROPE), as compared to Bayes factors against the point-null (*BF vs.* 0),
311 leads to a higher sensitivity to null-effects (Morey & Rouder, 2011; Rouder & Morey, 2012).

312 We also showed that besides sharing similar properties, the *pd* has a 1:1
313 correspondence with the frequentist *p*-value, being its Bayesian equivalent. On the contrary
314 Bayes factors appear as having a severely non-linear relationship with the frequentist
315 index, which is to be expected from their mathematical definition and their sensitivity
316 when the null hypothesis is true. This in turn can lead to surprising conclusions. For
317 instance, Bayes factors lower than 1, which are considered as providing evidence *against*
318 the presence of an effect, can still correspond to a “significant” frequentist *p*-value (see
319 **Figures 3 and 4**). ROPE indices are more closely related to the *p*-value, as their
320 relationship appears dependent on another factor, the sample size. This suggests that the
321 ROPE encapsulates additional information about the strength of evidence.

322 What is the point of comparing Bayesian indices with the frequentist *p*-value,
323 especially after having pointed out to its many flaws? While this comparison may seem
324 counter-intuitive (as Bayesian thinking is intrinsically different from the frequentist
325 framework), we believe that this juxtaposition is interesting for didactic reasons. The
326 frequentist *p*-value “speaks” to many and can thus be seen as a reference and a way to
327 facilitate the shift toward the Bayesian framework. Thus, pragmatically documenting such
328 bridges can only foster the understanding of the methodological issues that our field is
329 facing, and in turn act against dogmatic adherence to a framework. This does not
330 preclude, however, that a change in the general paradigm of significance seeking and

³³¹ “p-hacking” is necessary, and that Bayesian indices are fundamentally different from the
³³² frequentist *p*-value, rather than mere approximations or equivalents.

³³³ Critically, while the purpose of these indices was solely referred to as *significance*
³³⁴ until now, we would like to emphasize the nuanced perspective of the existence-significance
³³⁵ testing as a dual-framework for parameters description and interpretation. The idea
³³⁶ supported here is that there is a conceptual and practical distinction, and possible
³³⁷ dissociation to be made, between an effect’s existence *and* significance. In this context,
³³⁸ *existence* is simply defined as the consistency of an effect in one particular direction (i.e.,
³³⁹ positive or negative), without any assumptions or conclusions as to its size, importance,
³⁴⁰ relevance or meaning. It is an objective feature of an estimate (tied to its uncertainty). On
³⁴¹ the other hand, *significance* would be here re-framed following its original literally
³⁴² definition such as “being worthy of attention” or “importance”. An effect can be considered
³⁴³ significant if its magnitude is higher than some given threshold. This aspect can be
³⁴⁴ explored, to a certain extent, in an objective way with the concept of *practical equivalence*
³⁴⁵ (Kruschke, 2014; Lakens, 2017; Lakens et al., 2018), which suggests the use of a range of
³⁴⁶ values assimilated to the absence of an effect (the ROPE). If the effect falls within this
³⁴⁷ range, it is considered as non-significant *for practical reasons*: the magnitude of the effect is
³⁴⁸ likely to be too small to be of high importance in real-world scenarios or applications.
³⁴⁹ Nevertheless, *significance* also withholds a more subjective aspect, corresponding to its
³⁵⁰ contextual meaningfulness and relevance. This, however, is usually dependent on the
³⁵¹ literature, priors, novelty, context or field, and thus cannot be objectively or neutrally
³⁵² assessed with a statistical index alone.

³⁵³ While indices of existence and significance can be numerically related (as shown in
³⁵⁴ our results), the former is conceptually independent from the latter. For example, an effect
³⁵⁵ for which the whole posterior distribution is concentrated within the [0.0001, 0.0002] range
³⁵⁶ would be considered as positive with a high certainty (and thus, *existing* in a that
³⁵⁷ direction), but also not significant (i.e., too small to be of any practical relevance).

Table 3

Summary of Bayesian Indices of Effect Existence and Significance.

Index	Interpretation	Definition	Strengths	Limitations
Probability of Direction (pd)	Probability that an effect is of the same sign as the median's.	Proportion of the posterior distribution of the same sign than the median's.	Straightforward computation and interpretation. Objective property of the posterior distribution. 1:1 correspondence with the frequentist p-value.	Limited information favoring the null hypothesis.
MAP-based p-value	Relative odds of the presence of an effect against 0.	Density value at 0 divided by the density value at the mode of the posterior distribution.	Straightforward computation. Objective property of the posterior distribution	Limited information favoring the null hypothesis. Relates on density approximation. Indirect relationship between mathematical definition and interpretation.
ROPE (95%)	Probability that the credible effect values are not negligible.	Proportion of the 95% CI inside of a range of values defined as the ROPE.	Provides information related to the practical relevance of the effects.	A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. Not sensitive to highly significant effects.
ROPE (full)	Probability that the effect possible values are not negligible.	Proportion of the posterior distribution inside of a range of values defined as the ROPE.	Provides information related to the practical relevance of the effects.	A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors.
Bayes factor (vs. 0)	The degree by which the probability mass has shifted away from or towards the null value, after observing the data.	Ratio of the density of the null value between the posterior and the prior distributions.	An unbounded continuous measure of relative evidence. Allows statistically supporting the null hypothesis.	Sensitive to selection of prior distribution shape, location and scale.
Bayes factor (vs. ROPE)	The degree by which the probability mass has into or outside of the null interval (ROPE), after observing the data.	Ratio of the odds of the posterior vs the prior distribution falling inside of the range of values defined as the ROPE.	An unbounded continuous measure of relative evidence. Allows statistically supporting the null hypothesis. Compared to the BF (vs. 0), evidence is accumulated faster for the null when the null is true.	Sensitive to selection of prior distribution shape, location and scale. Additionally, a ROPE range needs to be arbitrarily defined, which is sensitive to the scale (the unit) of the predictors.

358 Acknowledging the distinction and complementary of these two aspects can in turn enrich
359 the information and usefulness of the results reported in psychological science (for practical
360 reasons, the implementation of this dual-framework of existence-significance testing is
361 made straightforward through the *bayestestR* open-source package for R; Makowski et al.,
362 2019). In this context, the *pd* and the MAP-based *p*-value appear as indices of effect
363 existence, mostly sensitive to the certainty related to the direction of the effect.
364 ROPE-based indices and Bayes factors are indices of effect significance, related to the
365 magnitude and the amount of evidence in favor of it (see also a similar discussion of
366 statistical significance vs. effect size in the frequentist framework; e.g., Cohen, 2016)

367 The inherent subjectivity related to the assessment of significance is one of the
368 practical limitation the ROPE-based indices (although being, conceptually, an asset,
369 allowing for contextual nuance in the interpretation), as they require an explicit definition
370 of the non-significant range (the ROPE). Although default values were reported in the
371 literature (for instance, half of a “negligible” effect size reference value; Kruschke, 2014), it
372 is critical for the reproducibility and transparency that the researcher’s choice is explicitly
373 stated (and, if possible, justified). Beyond being arbitrary, this range also has hard bounds
374 (for instance, contrary to a value of 0.0499, a value of 0.0501 would be considered as
375 non-negligible if the range ends at 0.05). This reinforces a categorical and clustered
376 perspective of what is by essence a continuous space of possibilities. Importantly, as this
377 range is fixed to the scale of the response (it is expressed in the unit of the response),
378 ROPE indices are sensitive to changes in the scale of the predictors. For instance,
379 negligible results may change into non-negligible results when predictors are scaled up
380 (e.g. express reaction times in seconds instead of milliseconds), which one inattentive or
381 malicious researcher could misleadingly present as “significant” (note that indices of
382 existence, such as the *pd*, would not be affected). Finally, the ROPE definition is also
383 dependent on the model type, and selecting a consistent or homogeneous range for all the
384 families of models is not straightforward. This can make comparisons between model types

385 difficult, and an additional burden when interpreting ROPE-based indices. In summary,
386 while a well-defined ROPE can be a powerful tool to give a different and new perspective,
387 it also requires extra caution from the authors and the readers.

388 As for the difference between ROPE (95%) and ROPE (full), we suggest reporting
389 the latter (i.e., the percentage of the whole posterior distribution that falls within the
390 ROPE instead of a given proportion of CI). This bypass the usage of another arbitrary
391 range (95%) and appears to be more sensitive to delineate highly significant effects).
392 Critically, rather than using the percentage in ROPE as a dichotomous, all-or-nothing
393 decision criterion, such as suggested by the original equivalence test (Kruschke, 2014), we
394 recommend using the percentage as a continuous index of significance (with explicitly
395 specified cut-off points if categorization is needed, for instance 5% for significance and 95%
396 for non-significance).

397 Our results underline Bayes factor as an interesting index, able to provide evidence in
398 favor or against the presence of an effect. Moreover, its easy interpretation in terms of odds
399 in favor, or against, one or the other hypothesis makes it a compelling index for
400 communication. Nevertheless, one of the main critiques of Bayes factors, is its sensitivity to
401 priors (shown in our results here through its sensitivity to model types, as priors' odds for
402 logistic and linear models are different). Moreover, while the BF against a ROPE appears
403 as even better than the BF against a point-null, it also carries all the limitations related to
404 the ROPE specification mentioned above. Thus, we recommend using Bayes factors
405 (preferentially *vs.* a ROPE) if the user has explicitly specified (and have a rationale for)
406 informative priors (often called “subjective” priors; Wagenmakers, 2007). In the end, there
407 is a relative proximity between Bayes factors (*vs.* ROPE) and the percentage in ROPE
408 (full), consistently with their mathematical relationship.

409 Being quite different from the Bayes factors and the ROPE indices, the Probability of
410 Direction (*pd*) is an index of effect existence representing the certainty with which an effect

411 goes in a particular direction (i.e., is positive or negative). Beyond its simplicity of
412 interpretation, understanding and computation, this index also presents other interesting
413 properties. It is independent from the model, i.e., it is solely based on the posterior
414 distributions and does not require any additional information from the data or the model.
415 Contrary to ROPE-based indices, it is robust to the scale of both the response variable and
416 the predictors. Nevertheless, this index also presents some limitations. Most importantly,
417 the *pd* is not relevant to assess size or importance of the effect and is not able to give
418 information *in favor* of the null hypothesis. In other words, a high *pd* suggests the presence
419 of an effect but a small *pd* does not give us any information about how much the null
420 hypothesis is plausible, suggesting that this index can only be used to eventually reject the
421 null hypothesis (which is consistent with the interpretation of the frequentist *p*-value). On
422 the contrary, the BFs (and to some extent the percentage in ROPE) increase or decrease as
423 the evidence becomes stronger (more data points), in both directions.

424 Much of the strengths of the *pd* also apply to the MAP-based *p*-value. Although
425 possibly showing some superiority in terms of sensitivity as compared to it, it also presents
426 an important limitation. Indeed, the MAP is mathematically dependent on the density at
427 0 and at the mode. However, the density estimation of a continuous distribution is a
428 statistical problem on its own and many different methods exist. It is possible that
429 changing the density estimation might impact the MAP-based *p*-value with unknown
430 results. The *pd*, however, has a linear relationship with the frequentist *p*-value, which is in
431 our opinion an asset.

432 After all the criticism regarding the frequentist *p*-value, it might appear as
433 contradictory to suggest the usage of its Bayesian empirical equivalent. The subtler
434 perspective that we support is that the *p*-value is not an intrinsically bad, or wrong, index.
435 Instead, it is its misuse, misunderstanding and misinterpretation that fuels the decay of the
436 situation into the crisis. Interestingly, the proximity between the *pd* and the *p*-value
437 suggests that the latter is more an index of effect *existence* than *significance* (as in “worth

438 of interest”; Cohen, 2016). Addressing this confusion, the Bayesian equivalent has an
439 intuitive meaning and interpretation, contributing to making more obvious the fact that all
440 thresholds and heuristics are arbitrary. In summary, its mathematical and interpretative
441 transparency of the pd , and its conceptualization as an index of effect existence, offers a
442 valuable insight into the characterization of Bayesian results, and its practical proximity
443 with the frequentist p -value makes it a perfect metric to ease the transition of psychological
444 research into the adoption of the Bayesian framework.

445 Our study has some limitations. First, our simulations were based on simple linear
446 and logistic regression models. Although these models are widely spread, the behavior of
447 the presented indices for other model families or types, like count models or mixed effects
448 models, still needs to be explored. Furthermore, we only tested continuous predictors. The
449 indices might behave differently when varying the type of predictor (binary, ordinal) as
450 well. Finally, we limited our simulations to small sample sizes, for reasons that data is
451 particularly noisy in small samples, and experiments in psychology often include only a
452 limited number of subjects. However, it is possible that the indices converge (or diverge),
453 for larger samples. Importantly, before being able to draw a definitive conclusion about the
454 qualities of these indices, further studies need to investigate the robustness of these indices
455 to sampling characteristics (*e.g.*, sampling algorithm, number of iterations, chains,
456 warm-up) and the impact of prior specification (Kass & Raftery, 1995; Kruschke, 2011;
457 Vanpaemel, 2010), all of which are important parameters of Bayesian statistics.

458 Reporting Guidelines

459 How can the current observations be used to improve statistical good practices in
460 psychological science? Based on the present comparison, we can start outlining the
461 following guidelines. As *existence* and *significance* are complementary perspectives, we
462 suggest using at minimum one index of each category. As an objective index of effect
463 existence, the pd should be reported, for its simplicity of interpretation, its robustness and

its numeric proximity to the well-known frequentist p -value; As an index of significance either the BF (*vs.* $ROPE$) or the $ROPE$ (*full*) should be reported, for their ability to discriminate between presence and absence of effect (De Santis, 2007), and the information they provide related to evidence of the size of the effect. Selection between the BF (*vs.* $ROPE$) or the $ROPE$ (*full*) should depend on the informativeness of the priors used - when uninformative priors are used, and there is little prior knowledge regarding the expected size of the effect, the $ROPE$ (*full*) should be reported as it reflects only the posterior distribution, and is not sensitive to the width of a wide-range of prior scales (Rouder, Haaf, & Vandekerckhove, 2018). On the other hand, in cases where informed priors are used, reflecting prior knowledge regarding the expected size of the effect, BF (*vs.* $ROPE$) should be used.

Defining appropriate heuristics to help the interpretation is beyond the scope of this paper, as it would require testing them on more natural datasets. Nevertheless, if we take the frequentist framework and the existing literature as a reference point, it seems that 95%, 97% and 99% might be relevant reference points (i.e., easy-to-remember values) for the pd and 3, 10 and 30 (weak evidence) appropriate for the BF . A concise, standardized, reference template sentence to describe the parameter of a model including an index of point-estimate, uncertainty, existence, significance and effect size (Cohen, 1988) could be, in the case of pd and BF :

“There is moderate evidence ($BFRope = 3.44$) [BF (*vs.* $ROPE$)] in favor of the presence of effect of X, which has a probability of 98.14% [pd] of being negative (Median = -5.04, 89% CI [-8.31, 0.12]), and can be considered as small (Std. Median = -0.29) [*standardized coefficient*]”

And if the user decides to use the percentage in $ROPE$ instead of the BF :

“The effect of X has a probability of 98.14% [pd] of being negative (Median = -5.04, 89% CI [-8.31, 0.12]), and can be considered as small (Std. Median = -0.29) [*standardized*

490 coefficient] and significant (0.82% in ROPE) [ROPE (full)]”

491 Data Availability

492 In the spirit of open and honest science, the full R code used for data generation,
493 data processing, figures creation and manuscript compiling is available on GitHub at https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian.
494

495 Ethics Statement

496 No human participants, but the authors of the present manuscript, were used to
497 produce the current study. The latter verbally reported being endowed with a feeling of
498 free-will at the moment of writing.

499 Author Contributions

500 DM conceived and coordinated the study. DM, MSB and DL participated in the
501 study design, statistical analysis, data interpretation and manuscript drafting. DL
502 supervised the manuscript drafting. AC performed a critical review of the manuscript,
503 assisted with manuscript drafting and provided funding for publication. All authors read
504 and approved the final manuscript.

505 Conflict of Interest Statement

506 The authors declare that the research was conducted in the absence of any
507 commercial or financial relationships that could be construed as a potential conflict of
508 interest.

509

Acknowledgments

510 This study was made possible by the development of the **bayestestR** package, itself
511 part of the *easystats* ecosystem (Lüdecke, Waggoner, & Makowski, 2019), an open-source
512 and collaborative project created to facilitate the usage of R. Thus, there is substantial
513 evidence in favor of the fact that we thank the masters of easystats and all the other
514 padawan following the way of the Bayes.

515

References

- 516 Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical
517 significance. *Nature*, 567(7748), 305–307. doi:10.1038/d41586-019-00857-9
- 518 Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing:
519 Problems, prevalence, and an alternative. *The Journal of Wildlife Management*,
520 912–923.
- 521 Andrews, M., & Baguley, T. (2013). Prior approval: The growth of bayesian methods in
522 psychology. *British Journal of Mathematical and Statistical Psychology*, 66(1), 1–7.
- 523 Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk,
524 R., ... others. (2018). Redefine statistical significance. *Nature Human Behaviour*,
525 2(1), 6.
- 526 Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ...
527 Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of
528 Statistical Software*, 76(1). doi:10.18637/jss.v076.i01
- 529 Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead
530 of 'playing the game' it is time to change the rules: Registered reports at aims
531 neuroscience and beyond. *AIMS Neuroscience*, 1(1), 4–17.
- 532 Cohen, J. (1988). Statistical power analysis for the social sciences.
- 533 Cohen, J. (2016). The earth is round ($p < .05$). In *What if there were no significance tests?*
534 (pp. 69–82). Routledge.
- 535 De Santis, F. (2007). Alternative bayes factors: Sample size determination and
536 discriminatory power assessment. *Test*, 16(3), 504–522.
- 537 Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in
538 Psychology*, 5, 781.

- 539 Dienes, Z., & McIatchie, N. (2018). Four reasons to prefer bayesian analyses over
540 significance testing. *Psychonomic Bulletin & Review*, 25(1), 207–218.
- 541 Ellis, S., & Steyn, H. (2003). Practical significance (effect sizes) versus or in combination
542 with statistical significance (p-values): Research note. *Management Dynamics:*
543 *Journal of the Southern African Institute for Management Scientists*, 12(4), 51–53.
- 544 Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and
545 testing any hypothesis you can specify. *Advances in Methods and Practices in*
546 *Psychological Science*, 2515245918773087.
- 547 Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility
548 project: Psychology. *PloS One*, 11(2), e0149794.
- 549 Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead
550 researchers to confidence intervals, but can't make them think: Statistical reform
551 lessons from medicine. *Psychological Science*, 15(2), 119–126.
- 552 Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., ... Goodman,
553 O. (2004). Reform of statistical inference in psychology: The case of Memory &
554 cognition. *Behavior Research Methods, Instruments, & Computers*, 36(2), 312–324.
- 555 Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values:
556 Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, 292(6522),
557 746–750.
- 558 Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying
559 Incremental Changes, and What to Do About It. *Personality and Social Psychology*
560 *Bulletin*, 44(1), 16–23. doi:10.1177/0146167217729162
- 561 Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014).
562 *Bayesian data analysis*. (Third edition.). Boca Raton: CRC Press.
- 563 Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2019). Rstanarm: Bayesian applied

564 regression modeling via Stan. Retrieved from <http://mc-stan.org/>

565 Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we
566 employ to fill the power vacuum? *Biology Letters*, 15(5), 20190174.

567 Heck, D. W. (2019). A caveat on the savage–dickey density ratio: The case of computing
568 bayes factors for regression parameters. *British Journal of Mathematical and
569 Statistical Psychology*, 72(2), 316–333.

570 Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and
571 reporting bayes factors. *The Journal of Problem Solving*, 7(1), 2.

572 Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

573 Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical
574 Association*, 90(430), 773–795.

575 Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational
576 and Psychological Measurement*, 56(5), 746–759.

577 Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*.
578 Academic Press.

579 Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in
580 Cognitive Sciences*, 14(7), 293–300.

581 Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and
582 model comparison. *Perspectives on Psychological Science*, 6(3), 299–312.

583 Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for
584 data analysis in the organizational sciences. *Organizational Research Methods*,
585 15(4), 722–752.

586 Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing,
587 estimation, meta-analysis, and power analysis from a bayesian perspective.
588 *Psychonomic Bulletin & Review*, 25(1), 178–206.

- 589 Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and
590 meta-analyses. *Social Psychological and Personality Science*, 8(4), 355–362.
- 591 Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological
592 research: A tutorial. *Advances in Methods and Practices in Psychological Science*,
593 2515245918770963.
- 594 Lüdecke, D., Waggoner, P., & Makowski, D. (2019). Insight: A unified interface to access
595 information from model objects in r. *Journal of Open Source Software*, 4(38), 1412.
596 doi:10.21105/joss.01412
- 597 Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor
598 hypothesis tests: Explanation, extension, and application in psychology. *Journal of
599 Mathematical Psychology*, 72, 19–32.
- 600 Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and
601 their Uncertainty, Existence and Significance within the Bayesian Framework.
602 *Journal of Open Source Software*, 4(40), 1541. doi:10.21105/joss.01541
- 603 Marasini, D., Quatto, P., & Ripamonti, E. (2016). The use of p-values in applied research:
604 Interpretation and new trends. *Statistica*, 76(4), 315–325.
- 605 Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a
606 replication crisis? What does “failure to replicate” really mean? *American
607 Psychologist*, 70(6), 487.
- 608 Mills, J. A. (2017). Objective bayesian precise hypothesis testing. *University of Cincinnati
609 [Original Version: 2007]*.
- 610 Mills, J. A., & Parent, O. (2014). Bayesian mcmc estimation. In *Handbook of regional
611 science* (pp. 1571–1595). Springer.
- 612 Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null
613 hypotheses. *Psychological Methods*, 16(4), 406.

- 614 R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna,
615 Austria: R Foundation for Statistical Computing. Retrieved from
616 <https://www.R-project.org/>
- 617 Robert, C. P. (2014). On the jeffreys-lindley paradox. *Philosophy of Science*, 81(2),
618 216–232.
- 619 Robert, C. P. (2016). The expected demise of the bayes factor. *Journal of Mathematical*
620 *Psychology*, 72, 33–37.
- 621 Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for
622 psychology, part iv: Parameter estimation and bayes factors. *Psychonomic Bulletin*
623 & Review, 25(1), 102–113.
- 624 Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in
625 regression. *Multivariate Behavioral Research*, 47(6), 877–903.
- 626 Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t
627 tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin &*
628 *Review*, 16(2), 225–237.
- 629 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology:
630 Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything
631 as Significant. *Psychological Science*, 22(11), 1359–1366.
632 doi:10.1177/0956797611417632
- 633 Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting
634 for publication bias using only significant results. *Perspectives on Psychological*
635 *Science*, 9(6), 666–681.
- 636 Spanos, A. (2013). Who should be afraid of the jeffreys-lindley paradox? *Philosophy of*
637 *Science*, 80(1), 73–93.
- 638 Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough.

- 639 *Journal of Graduate Medical Education*, 4(3), 279–282.
- 640 Szucs, D., & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and
641 power in the recent cognitive neuroscience and psychology literature. *BioRxiv*,
642 071530.
- 643 Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes
644 factor. *Journal of Mathematical Psychology*, 54(6), 491–498.
- 645 Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values.
646 *Psychonomic Bulletin & Review*, 14(5), 779–804.
- 647 Wagenmakers, E.-J., Lee, M., Rouder, J., & Morey, R. (2019, August). *Another statistical
648 paradox*. Retrieved from
649 <http://www.ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf>
- 650 Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian
651 hypothesis testing for psychologists: A tutorial on the savage–dickey method.
652 *Cognitive Psychology*, 60(3), 158–189.
- 653 Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . others.
654 (2018). Bayesian inference for psychology. Part i: Theoretical advantages and
655 practical ramifications. *Psychonomic Bulletin & Review*, 25(1), 35–57.
- 656 Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the
657 pragmatic researcher. *Current Directions in Psychological Science*, 25(3), 169–176.
- 658 Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., &
659 Morey, R. D. (2017). The need for bayesian hypothesis testing in psychological
660 science. *Psychological Science Under Scrutiny: Recent Challenges and Proposed
661 Solutions*, 123–138.
- 662 Wasserstein, R. L., Lazar, N. A., & others. (2016). The asa’s statement on p-values:
663 Context, process, and purpose. *The American Statistician*, 70(2), 129–133.