

1 Indices of Effect Existence and Significance in the Bayesian Framework

2 Dominique Makowski ^{1,*}, Mattan S. Ben-Shachar ², S.H. Annabel Chen ^{1, 3, 4, *, †}, & Daniel
3 Lüdecke ^{5, †}

4 ¹ School of Social Sciences, Nanyang Technological University, Singapore

5 ² Department of Psychology, Ben-Gurion University of the Negev, Israel

6 ³ Centre for Research and Development in Learning, Nanyang Technological University,
7 Singapore

8 ⁴ Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

9 ⁵ Department of Medical Sociology, University Medical Center Hamburg-Eppendorf,
10 Germany

11 Author Note

12 * Correspondence concerning this article should be addressed to Dominique Makowski

13 (HSS 04-18, 48 Nanyang Avenue, Singapore; dmakowski@ntu.edu.sg) and S.H. Annabel

14 Chen (HSS 04-19, 48 Nanyang Avenue, Singapore; annabelchen@ntu.edu.sg).

15 † S.H. Annabel Chen and Daniel Lüdecke share senior authorship.

16

Abstract

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

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

36 Word count: 6194

37 Indices of Effect Existence and Significance in the Bayesian Framework

38 **Introduction**

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

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

significance), are useful and necessary for researchers to gain an intuitive (although possibly simplified) understanding of the interactions and structure of their data.

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

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

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

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

93 Bayesian “significance” testing indices could be roughly grouped into three
94 overlapping categories: Bayes factors, posterior indices and Region of Practical Equivalence
95 (ROPE)-based indices. Bayes factors are a family of indices of relative evidence of one
96 model over another (e.g., the null *vs.* the alternative hypothesis; Jeffreys, 1998; Ly,
97 Verhagen, & Wagenmakers, 2016). They provide many advantages over the *p*-value by
98 having a straightforward interpretation as well as allowing to quantify evidence in favor of
99 the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). However, its use for parameters
100 description in complex models is still a matter of debate (Heck, 2019; Wagenmakers,
101 Lodewyckx, Kuriyal, & Grasman, 2010), being highly dependent on the specification of
102 priors (Etz, Haaf, Rouder, & Vandekerckhove, 2018; Kruschke & Liddell, 2018). On the
103 contrary, “posterior indices” reflect objective characteristics of the posterior distribution,
104 for instance the proportion of strictly positive values. While the simplicity of their
105 computation and interpretation is an asset, it might also limit the information that they
106 provide. Finally, ROPE-based indices are related to the redefinition of the null hypothesis
107 from the classic point-null hypothesis to a range of values considered negligible or too small
108 to be of any practical relevance (the Region of Practical Equivalence - ROPE; Kruschke,
109 2014; Lakens, 2017; Lakens, Scheel, & Isager, 2018), usually spread equally around 0 (e.g.,
110 [-0.1; 0.1]). It is interesting to note that this perspective unites significance testing with the
111 focus on effect size (involving a discrete separation between at least two categories:
112 negligible and non-negligible), which finds an echo in recent statistical recommendations
113 (Ellis & Steyn, 2003; Simonsohn, Nelson, & Simmons, 2014; Sullivan & Feinn, 2012).

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

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

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

139

Methods

140 Data Simulation

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

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

160 Indices

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

¹⁶⁴ **Frequentist *p*-value.** This was the only index computed by the frequentist version
¹⁶⁵ of the regression. The *p*-value represents the probability that for a given statistical model,
¹⁶⁶ when the null hypothesis is true, the effect would be greater than or equal to the observed
¹⁶⁷ coefficient (Wasserstein, Lazar, & others, 2016).

¹⁶⁸ **Probability of Direction (*pd*).** The *Probability of Direction* (*pd*) varies between
¹⁶⁹ 50% and 100% and can be interpreted as the probability that a parameter (described by its
¹⁷⁰ posterior distribution) is strictly positive or negative (whichever is the most probable). It is
¹⁷¹ mathematically defined as the proportion of the posterior distribution that is of the
¹⁷² median's sign (Makowski, Ben-Shachar, & Lüdecke, 2019).

¹⁷³ **MAP-based *p*-value.** The *MAP-based p-value* is related to the odds that a
¹⁷⁴ parameter has against the null hypothesis (Mills, 2017; Mills & Parent, 2014). It is
¹⁷⁵ mathematically defined as the density value at 0 divided by the density at the Maximum A
¹⁷⁶ Posteriori (MAP), i.e., the equivalent of the mode for continuous distributions.

¹⁷⁷ **ROPE (95%).** The *ROPE (95%)* refers to the percentage of the 95% Highest
¹⁷⁸ Density Interval (HDI) that lies within the ROPE. As suggested by Kruschke (2014), the
¹⁷⁹ Region of Practical Equivalence (ROPE) was defined as range from -0.1 to 0.1 for linear
¹⁸⁰ regressions and its equivalent, -0.18 to 0.18, for logistic models (based on the $\pi/\sqrt{3}$ formula
¹⁸¹ to convert log odds ratios to standardized differences; Cohen, 1988).

¹⁸² **ROPE (full).** The *ROPE (full)* is similar to *ROPE (95%)*, with the exception that
¹⁸³ it refers to the percentage of the *whole* posterior distribution that lies within the ROPE.

¹⁸⁴ **Bayes factor (*vs.* 0).** The Bayes Factor (*BF*) used here is based on prior and
¹⁸⁵ posterior distributions of a single parameter. In this context, the Bayes factor indicates the
¹⁸⁶ degree by which the mass of the posterior distribution has shifted further away from or
¹⁸⁷ closer to the null value (0), relative to the prior distribution, thus indicating if the null
¹⁸⁸ hypothesis has become less or more likely given the observed data. The *BF* was computed
¹⁸⁹ as a Savage-Dickey density ratio, which is also an approximation of a Bayes factor

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

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

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

199 **Data Analysis**

200 In order to achieve the two-fold aim of this study; 1) comparing Bayesian indices and
 201 2) provide visual guides for an intuitive understanding of the numeric values in relation to
 202 a known frame of reference (the frequentist *p*-value), we will start by presenting the
 203 relationship between these indices and main sources of variance, such as sample size, noise
 204 and null hypothesis (true if absence of effect, false if presence of effect). We will then
 205 compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds
 206 (.05, .01, .001). Finally, we will show the mutual relationship between three recommended
 207 Bayesian candidates. Taken together, these results will help us outline guides to ease the
 208 reporting and interpretation of the indices.

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

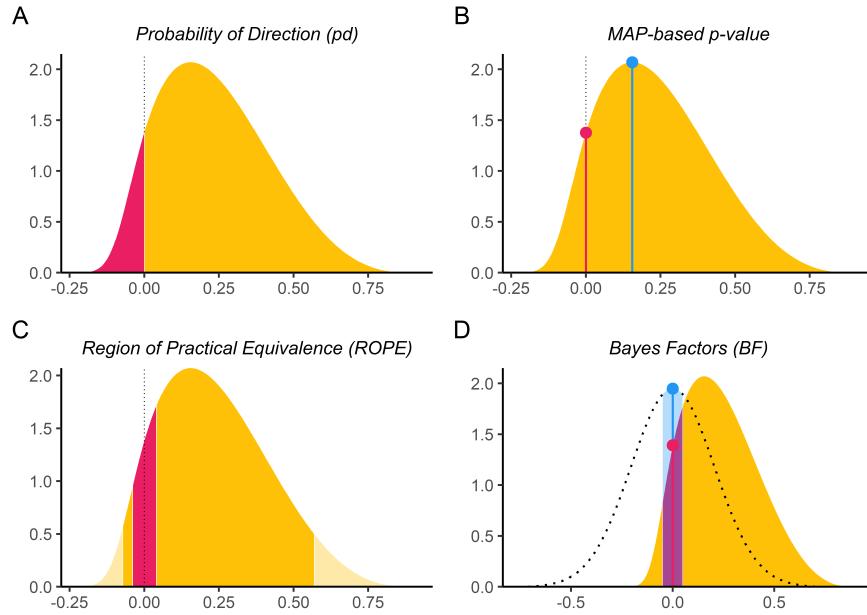


Figure 1. Bayesian indices of effect existence and significance. (A) The Probability of Direction (*pd*) is defined as the proportion of the posterior distribution that is of the median's sign (the size of the yellow area relative to the whole distribution). (B) The MAP-based *p*-value is defined as 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).

213 investigation. Thus, we fitted two regression models to assess the impact of sample size and
 214 noise, respectively. For these models (but not for the figures), to ensure that any
 215 differences between the indices are not due to differences in their scale or distribution, we
 216 converted all indices to the same scale by normalizing the indices between 0 and 1 (note

217 that BF s were transformed to posterior probabilities, assuming uniform prior odds) and
218 reversing the p -values, the MAP-based p -values and the ROPE indices so that a higher
219 value corresponds to stronger “significance”.

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

228

Results

229 **Impact of Sample Size**

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

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

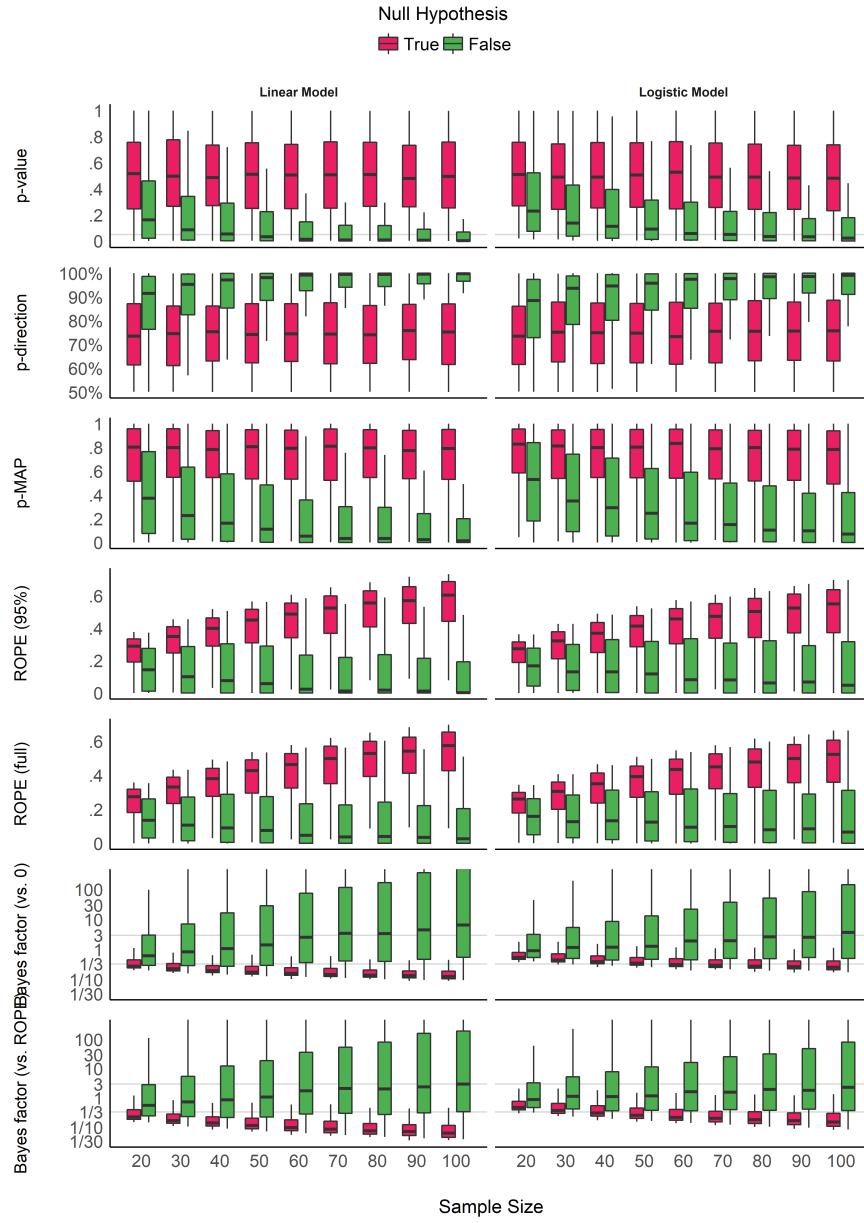


Figure 2. Impact of Sample Size on the different indices, for linear and logistic models, and when the null hypothesis is true or false. Grey vertical lines for $*p*$ -values and Bayes factors represent commonly used thresholds.

- ²⁴² indices are particularly sensitive to sample size when the null hypothesis is true, while
²⁴³ p -value, pd and MAP-based p -value are only sensitive to sample size when the null
²⁴⁴ hypothesis is false, in which case they are more sensitive than $ROPE$ indices. These

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

245 findings can be related to the concept of consistency: as the number of data points
 246 increases, the statistic converges toward some “true” value. Here, we observe that *p*-value,
 247 *pd* and the MAP-based *p*-value are consistent only when the null hypothesis is false. In
 248 other words, as sample size increases, they tend to reflect more strongly that the effect is
 249 present. On the other hand, *ROPE* indices appear as consistent when the effect is absent.
 250 Finally, *BFs* are consistent both when the effect is absent and when it is present, and *BF*
 251 (*vs. ROPE*), compared to *BF* (*vs. 0*), is more sensitive to sample size when the null
 252 hypothesis is true, and *ROPE (full)* is overall slightly more consistent than *ROPE (95%)*.

253 Impact of Noise

254 **Figure 3** shows the indices’ sensitivity to noise. Unlike the patterns of sensitivity to
 255 sample size, the indices display more similar patterns in their sensitivity to noise (or
 256 magnitude of effect). All indices are unidirectional impacted by noise: as noise increases,

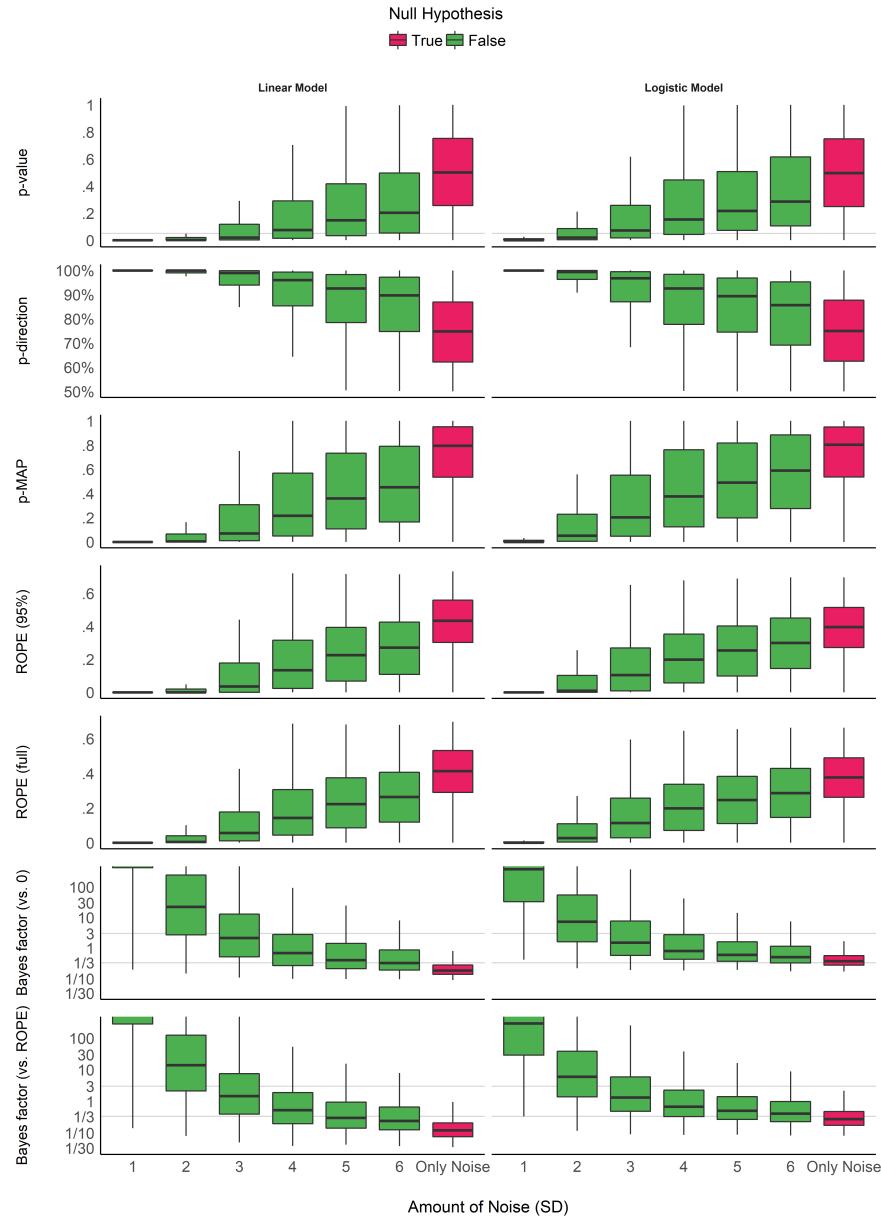


Figure 3. Impact of Noise. The noise corresponds to the standard deviation of the Gaussian noise that was added to the generated data. It is related to the magnitude the parameter (the more noise there is, the smaller the coefficient). Grey vertical lines for $*p*$ -values and Bayes factors represent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

257 the observed coefficients decrease in magnitude, and the indices become less “pronounced”
 258 (respectively to their direction). However, it is interesting to note that the variability of
 259 the indices seems differently impacted by noise. For the p -values, the pd and the ROPE
 260 indices, the variability increases as the noise increases. In other words, small variation in
 261 small observed coefficients can yield very different values. On the contrary, the variability
 262 of BFs decreases as the true effect tends toward 0. For the MAP-based p -value, the
 263 variability appears to be the highest for moderate amount of noise. This behavior seems
 264 consistent across model types.

265 Consistently with **Figure 3**, the model investigating the sensitivity of noise when an
 266 effect is present (as there is only noise in the absence of effect), adjusted for sample size,
 267 suggests that BFs (especially *vs.* ROPE), followed by the MAP-based p -value and
 268 percentages in ROPE, are the most sensitive to noise. As noise is a proxy of effect size
 269 (linearly related to the absolute value of the coefficient of the parameter), this result
 270 highlights the fact that these indices are sensitive to the magnitude of the effect. For

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 *pd* 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

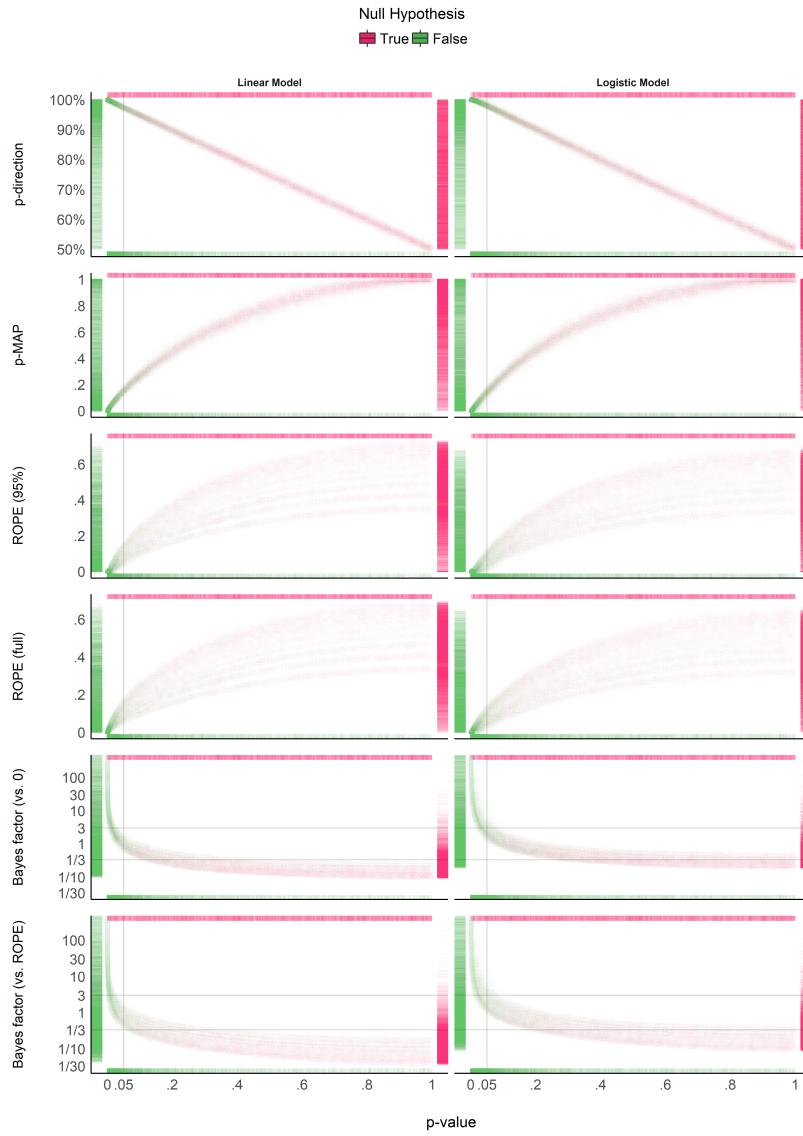


Figure 4. Relationship with the frequentist $*p*$ -value. In each plot, the $*p*$ -value densities are visualized by the marginal top (absence of true effect) and bottom (presence of true effect) markers, whereas on the left (presence of true effect) and right (absence of true effect), the markers represent 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).

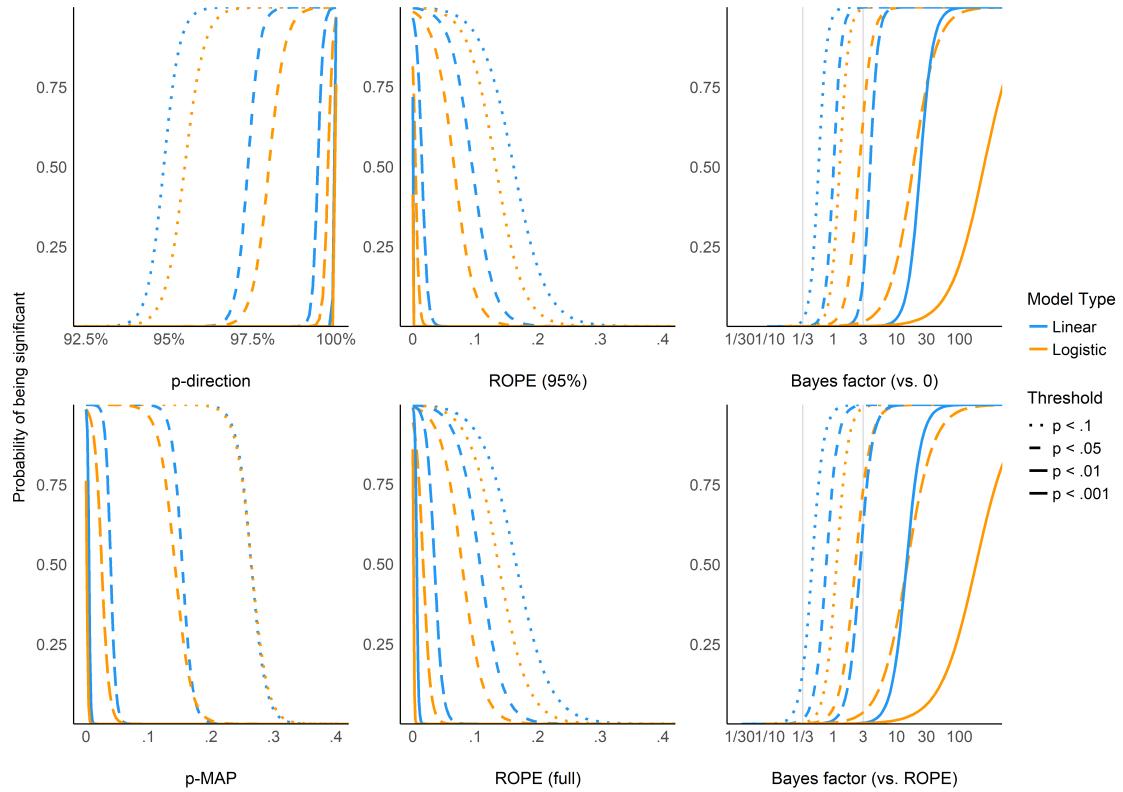


Figure 5. The probability of reaching different $*p$ -value based significance thresholds (.1, .05, .01, .001 for solid, long-dashed, short-dashed and dotted lines, respectively) for different values of the corresponding Bayesian indices.

295 definition of BF (*vs. ROPE*) above). Overall, these results help to demonstrate $ROPE$
 296 (*full*) and BF (*vs. ROPE*)’s consistency both in case of presence and absence of a true
 297 effect, whereas the pd , being equivalent to the p -value, is only consistent when the true
 298 effect is absent.

299 Discussion

300 Based on the simulation of linear and logistic models, the present work aimed at
 301 comparing several Bayesian indices of effect “significance” (see **Table 3**), providing visual
 302 representations of the “behavior” of such indices in relationship with important sources of
 303 variance such as sample size, noise and effect presence, as well as comparing them with the

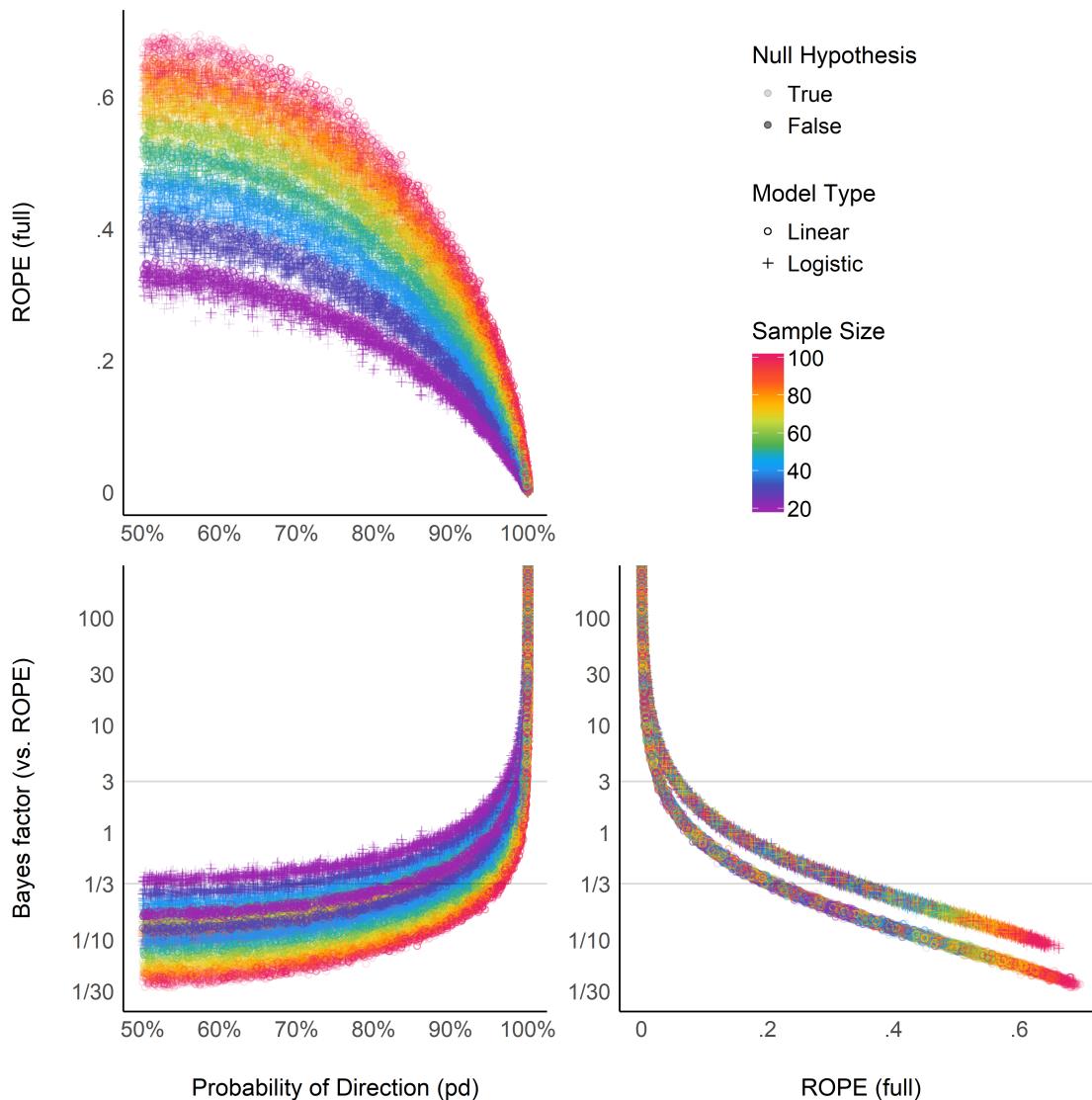


Figure 6. Relationship between three Bayesian indices: The Probability of Direction (*pd*), the percentage of the full posterior distribution in the ROPE, and the Bayes factor (*vs.* ROPE).

304 well-known and widely used frequentist p -value and its arbitrary interpretation thresholds.

305 The results tend to suggest that the investigated indices could be separated into two
 306 categories. The first group, including the pd and the MAP-based p -value, presents similar
 307 properties to those of the frequentist p -value: they are sensitive to the amount of evidence
 308 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 (Rouder & Morey, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). A high value suggests that the effect exists, but a low value indicates *uncertainty* about its existence (but not certainty that it is non-existent). The second group, including ROPE and Bayes factors, seem sensitive to both presence and absence of effect, accumulating evidence as the sample size increases. However, the ROPE seems particularly suited to provide evidence in favor of the null hypothesis. Consistently with this, combining Bayes factors with the ROPE (*BF vs.* ROPE), as compared to Bayes factors against the point-null (*BF vs.* 0), leads to a higher sensitivity to null-effects (Morey & Rouder, 2011; Rouder & Morey, 2012).

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

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

336 preclude, however, that a change in the general paradigm of significance seeking and
337 “p-hacking” is necessary, and that Bayesian indices are fundamentally different from the
338 frequentist *p*-value, rather than mere approximations or equivalents.

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

359 While indices of existence and significance can be numerically related (as shown in
360 our results), the former is conceptually independent from the latter. For example, an effect
361 for which the whole posterior distribution is concentrated within the [0.0001, 0.0002] range
362 would be considered as positive with a high certainty (and thus, *existing* in a that

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.

363 direction), but also not significant (i.e., too small to be of any practical relevance).
364 Acknowledging the distinction and complementary of these two aspects can in turn enrich
365 the information and usefulness of the results reported in psychological science (for practical
366 reasons, the implementation of this dual-framework of existence-significance testing is
367 made straightforward through the *bayestestR* open-source package for R; Makowski et al.,
368 2019). In this context, the *pd* and the MAP-based *p*-value appear as indices of effect
369 existence, mostly sensitive to the certainty related to the direction of the effect.
370 ROPE-based indices and Bayes factors are indices of effect significance, related to the
371 magnitude and the amount of evidence in favor of it (see also a similar discussion of
372 statistical significance vs. effect size in the frequentist framework; e.g., Cohen, 2016)

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

³⁹⁰ families of models is not straightforward. This can make comparisons between model types
³⁹¹ difficult, and an additional burden when interpreting ROPE-based indices. In summary,
³⁹² while a well-defined ROPE can be a powerful tool to give a different and new perspective,
³⁹³ it also requires extra caution from the authors and the readers.

³⁹⁴ As for the difference between ROPE (95%) and ROPE (full), we suggest reporting
³⁹⁵ the latter (i.e., the percentage of the whole posterior distribution that falls within the
³⁹⁶ ROPE instead of a given proportion of CI). This bypass the usage of another arbitrary
³⁹⁷ range (95%) and appears to be more sensitive to delineate highly significant effects).
³⁹⁸ Critically, rather than using the percentage in ROPE as a dichotomous, all-or-nothing
³⁹⁹ decision criterion, such as suggested by the original equivalence test (Kruschke, 2014), we
⁴⁰⁰ recommend using the percentage as a continuous index of significance (with explicitly
⁴⁰¹ specified cut-off points if categorization is needed, for instance 5% for significance and 95%
⁴⁰² for non-significance).

⁴⁰³ Our results underline Bayes factor as an interesting index, able to provide evidence in
⁴⁰⁴ favor or against the presence of an effect. Moreover, its easy interpretation in terms of odds
⁴⁰⁵ in favor, or against, one or the other hypothesis makes it a compelling index for
⁴⁰⁶ communication. Nevertheless, one of the main critiques of Bayes factors, is its sensitivity to
⁴⁰⁷ priors (shown in our results here through its sensitivity to model types, as priors' odds for
⁴⁰⁸ logistic and linear models are different). Moreover, while the BF against a ROPE appears
⁴⁰⁹ as even better than the BF against a point-null, it also carries all the limitations related to
⁴¹⁰ the ROPE specification mentioned above. Thus, we recommend using Bayes factors
⁴¹¹ (preferentially *vs.* a ROPE) if the user has explicitly specified (and have a rationale for)
⁴¹² informative priors (often called “subjective” priors; Wagenmakers, 2007). In the end, there
⁴¹³ is a relative proximity between Bayes factors (*vs.* ROPE) and the percentage in ROPE
⁴¹⁴ (full), consistently with their mathematical relationship.

⁴¹⁵ Being quite different from the Bayes factors and the ROPE indices, the Probability of

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

430 Much of the strengths of the pd also apply to the MAP-based p -value. Although
431 possibly showing some superiority in terms of sensitivity as compared to it, it also presents
432 an important limitation. Indeed, the MAP is mathematically dependent on the density at
433 0 and at the mode. However, the density estimation of a continuous distribution is a
434 statistical problem on its own and many different methods exist. It is possible that
435 changing the density estimation might impact the MAP-based p -value with unknown
436 results. The pd , however, has a linear relationship with the frequentist p -value, which is in
437 our opinion an asset.

438 After all the criticism regarding the frequentist p -value, it might appear as
439 contradictory to suggest the usage of its Bayesian empirical equivalent. The subtler
440 perspective that we support is that the p -value is not an intrinsically bad, or wrong, index.
441 Instead, it is its misuse, misunderstanding and misinterpretation that fuels the decay of the
442 situation into the crisis. Interestingly, the proximity between the pd and the p -value

443 suggests that the latter is more an index of effect *existence* than *significance* (as in “worth
444 of interest”; Cohen, 2016). Addressing this confusion, the Bayesian equivalent has an
445 intuitive meaning and interpretation, contributing to making more obvious the fact that all
446 thresholds and heuristics are arbitrary. In summary, its mathematical and interpretative
447 transparency of the *pd*, and its conceptualization as an index of effect existence, offers a
448 valuable insight into the characterization of Bayesian results, and its practical proximity
449 with the frequentist *p*-value makes it a perfect metric to ease the transition of psychological
450 research into the adoption of the Bayesian framework.

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

464 Reporting Guidelines

465 How can the current observations be used to improve statistical good practices in
466 psychological science? Based on the present comparison, we can start outlining the
467 following guidelines. As *existence* and *significance* are complementary perspectives, we
468 suggest using at minimum one index of each category. As an objective index of effect

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

481 Defining appropriate heuristics to help the interpretation is beyond the scope of this
482 paper, as it would require testing them on more natural datasets. Nevertheless, if we take
483 the frequentist framework and the existing literature as a reference point, it seems that
484 95%, 97% and 99% might be relevant reference points (i.e., easy-to-remember values) for
485 the *pd*. A concise, standardized, reference template sentence to describe the parameter of a
486 model including an index of point-estimate, uncertainty, existence, significance and effect
487 size (Cohen, 1988) could be, in the case of *pd* and *BF*:

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

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

493 “The effect of X has a probability of 98.14% [*pd*] of being negative ($Median = -5.04$,
494 $89\%CI[-8.31, 0.12]$), and can be considered as small ($Std.Median = -0.29$) [*standardized*

495 coefficient] and significant (0.82% in ROPE) [ROPE (full)]".

496 Data Availability

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

500 Ethics Statement

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

504 Author Contributions

505 DM conceived and coordinated the study. DM, MSB and DL participated in the
506 study design, statistical analysis, data interpretation and manuscript drafting. DL
507 supervised the manuscript drafting. AC performed a critical review of the manuscript,
508 assisted with manuscript drafting and provided funding for publication. All authors read
509 and approved the final manuscript.

510 Conflict of Interest Statement

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

514

Acknowledgments

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

520

References

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

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

- 570 regression modeling via Stan. Retrieved from <http://mc-stan.org/>
- 571 Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we
572 employ to fill the power vacuum? *Biology Letters*, 15(5), 20190174.
- 573 Heck, D. W. (2019). A caveat on the savage–dickey density ratio: The case of computing
574 bayes factors for regression parameters. *British Journal of Mathematical and
575 Statistical Psychology*, 72(2), 316–333.
- 576 Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and
577 reporting bayes factors. *The Journal of Problem Solving*, 7(1), 2.
- 578 Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.
- 579 Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical
580 Association*, 90(430), 773–795.
- 581 Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational
582 and Psychological Measurement*, 56(5), 746–759.
- 583 Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*.
584 Academic Press.
- 585 Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in
586 Cognitive Sciences*, 14(7), 293–300.
- 587 Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and
588 model comparison. *Perspectives on Psychological Science*, 6(3), 299–312.
- 589 Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for
590 data analysis in the organizational sciences. *Organizational Research Methods*,
591 15(4), 722–752.
- 592 Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing,
593 estimation, meta-analysis, and power analysis from a bayesian perspective.
594 *Psychonomic Bulletin & Review*, 25(1), 178–206.

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

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

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