Indices of Effect Existence and Significance in the Bayesian Framework

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

[1] School of Social Sciences, Nanyang Technological University, Singapore

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

[3] Centre for Research and Development in Learning, Nanyang Technological University, Singapore

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

[5] Department of Medical Sociology, University Medical Center Hamburg-Eppendorf, Germany

Author Note

[*] Correspondence concerning this article should be addressed to Dominique Makowski (HSS 04-18, 48 Nanyang Avenue, Singapore; dmakowski@ntu.edu.sg) and S.H. Annabel Chen (HSS 04-19, 48 Nanyang Avenue, Singapore; annabelchen@ntu.edu.sg).

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

Abstract

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

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

Word count: 6326

Indices of Effect Existence and Significance in the Bayesian Framework

Introduction

The Bayesian framework is quickly gaining popularity among psychologists and neuroscientists (Andrews & Baguley, 2013), for reasons such as flexibility, better accuracy in noisy data and small samples, less proneness to type I errors, the possibility of introducing prior knowledge into the analysis and the intuitiveness and straightforward interpretation of results (Dienes & Mclatchie, 2018; Etz & Vandekerckhove, 2016; Kruschke, 2010; Kruschke, Aguinis, & Joo, 2012; Wagenmakers et al., 2018; Wagenmakers, Morey, & Lee, 2016). On the other hand, the frequentist approach has been associated with the focus on *p*-values and null hypothesis significance testing (NHST). The misinterpretation and misuse of *p*-values, so called "p-hacking" (Simmons, Nelson, & Simonsohn, 2011), has been shown to critically contribute to the reproducibility crisis in psychological science (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Szucs & Ioannidis, 2016). The reliance on *p*-values has been criticized for its association with inappropriate inference, and effects can be drastically overestimated, sometimes even in the wrong direction, when estimation is tied to statistical significance in highly variable data (Gelman, 2018). Power calculations allow researchers to control the probability of falsely rejecting the null hypothesis, but do not completely solve this problem. For instance, the "false-alarm probability" of even very small *p*-values can be much higher than expected (Nuzzo, 2014). In response, there is an increasing belief that the generalization and utilization of the Bayesian framework is one way of overcoming these issues (Benjamin et al., 2018; Etz & Vandekerckhove, 2016; Halsey, 2019; Marasini, Quatto, & Ripamonti, 2016; Maxwell, Lau, & Howard, 2015; Wagenmakers et al., 2017).

The tenacity and resilience of the *p*-value as an index of significance is remarkable, despite the long-lasting criticism and discussion about its misuse and misinterpretation (Anderson, Burnham, & Thompson, 2000; Cohen, 1994; Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Finch et al., 2004; Gardner & Altman, 1986). This endurance might be informative on how such indices, and the accompanying heuristics applied to 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, 1994; 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 improvements are needed. For instance, one possible advance could focus on improving the understanding of the values being used, for instance, through a new, simpler, index. Bayesian inference allows making intuitive probability statements of an effect, as opposed to the less straightforward mathematical definition of the *p*-value, that contributes 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 expectation), resulting in a distribution of possible values for the parameters, called the posterior distribution. The description of the posterior distribution (e.g., through its centrality, dispersion, etc.) allows to draw conclusions from Bayesian analyses.

Bayesian "significance" testing indices could be roughly grouped into three overlapping categories: Bayes factors, posterior indices and Region of Practical Equivalence (ROPE)-based indices. Bayes factors are a family of indices of relative evidence of one model over another (e.g., the null *vs.* the alternative hypothesis; Jeffreys, 1998; Ly, Verhagen, & Wagenmakers, 2016). Aside from having a straightforward interpretation ("given the observed data, is the null hypothesis of an absence of an effect more, or less likely?"), they allow to quantify the evidence in favor of the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). However, its use for parameters description in complex models is still a matter of debate (Heck, 2019; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), being highly dependent on the specification of priors (Etz, Haaf, Rouder, & Vandekerckhove, 2018; Kruschke & Liddell, 2018). On the contrary, "posterior indices" reflect objective characteristics of the posterior distribution, for instance the proportion of strictly positive values. They also allow to derive legitimate statements that indicate the probability of an effect falling in a given range similar to the misleading conclusions related to frequentist confidence intervals. Finally, ROPE-based indices are related to the redefinition of the null hypothesis from the classic point-null hypothesis to a range of values considered negligible or too small to be of any practical relevance (the Region of Practical Equivalence - ROPE; Kruschke, 2014; Lakens, 2017; Lakens, Scheel, & Isager, 2018), usually spread equally around 0 (e.g., [-0.1; 0.1]). The idea behind this index is that an effect is almost never exactly zero, but instead can be very tiny, with no practical relevance. It is interesting to note that this perspective unites significance testing with the focus on effect size (involving a discrete separation between at least two categories: negligible and non-negligible), which finds an echo in recent statistical recommendations (Ellis & Steyn, 2003; Simonsohn, Nelson, & Simmons, 2014; Sullivan & Feinn, 2012).

Despite the richness provided by the Bayesian framework and the availability of multiple indices, no consensus has yet emerged on which ones to be used. Literature 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.

Methods

Data Simulation

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

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

Indices

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

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

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

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

ROPE (95%). The *ROPE (95%)* refers to the percentage of the 95% Highest Density Interval (HDI) that lies within the ROPE. As suggested by Kruschke (2014), the Region of Practical Equivalence (ROPE) was defined as range from -0.1 to 0.1 for linear regressions and its equivalent, -0.18 to 0.18, for logistic models (based on the $\pi/\sqrt{3}$ formula to convert log odds ratios to standardized differences; Cohen, 1988). Although we present the "95% percentage" because of the history of this index and of its widespread use, the reader should note that this value was recently challenged due to its arbitrary nature (McElreath, 2018).

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

Bayes factor (*vs.* 0). The Bayes Factor (*BF*) used here is based on prior and posterior distributions of a single parameter. In this context, the Bayes factor indicates the degree by which the mass of the posterior distribution has shifted further away from or closer to the null value (0), relative to the prior distribution, thus indicating if the null hypothesis has become less or more likely given the observed data. The *BF* was computed as a Savage-Dickey density ratio, which is also an approximation of a Bayes factor comparing the marginal likelihoods of the model against a model in which the tested parameter has been restricted to the point-null (Wagenmakers et al., 2010).

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

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

Data Analysis

In order to achieve the two-fold aim of this study; 1) comparing Bayesian indices and 2) provide visual guides for an intuitive understanding of the numeric values in relation to a known frame of reference (the frequentist *p*-value), we will start by presenting the relationship between these indices and main sources of variance, such as sample size, noise and null hypothesis (true if absence of effect, false if presence of effect). We will then

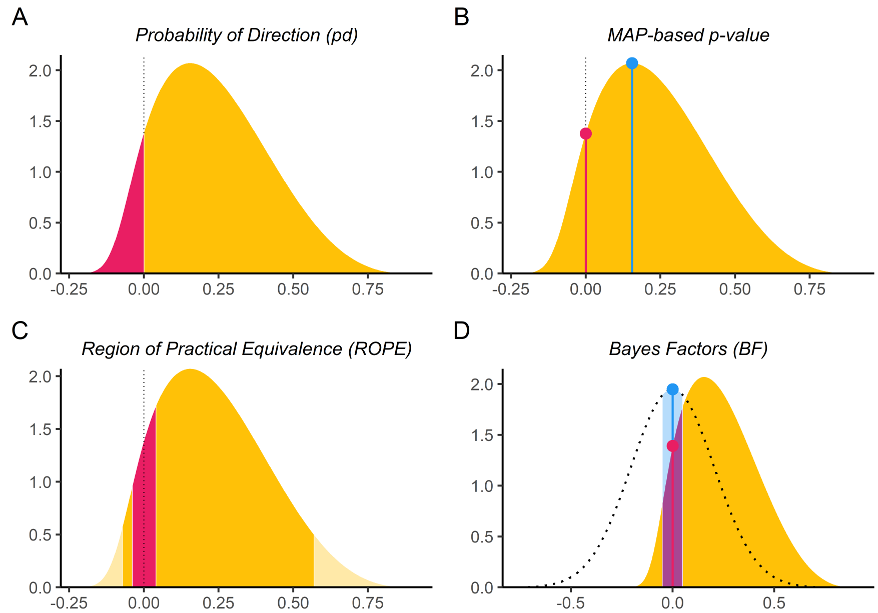

Figure 1. Bayesian indices of effect existence and significance. (A) The Probability of Direction (*pd*) is defined as the proportion of the posterior distribution that is of the median's sign (the size of the yellow area relative to the whole distribution). (B) The MAP-based *p*-value is defined as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

compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds (.05, .01, .001). Finally, we will show the mutual relationship between three recommended Bayesian candidates. Taken together, these results will help us outline guides to ease the reporting and interpretation of the indices.

In order to provide an intuitive understanding of values, data processing will focus on creating clear visual figures to help the user grasp the patterns and variability that exists when computing the investigated indices. Nevertheless, we decided to also mathematically test our claims in cases where the graphical representation begged for a deeper investigation. Thus, we fitted two regression models to assess the impact of sample size and noise, respectively. For these models (but not for the figures), to ensure that any differences between the indices are not due to differences in their scale or distribution, we converted all indices to the same scale by normalizing the indices between 0 and 1 (note that *BF*s were transformed to posterior probabilities, assuming uniform prior odds) and reversing the *p*-values, the MAP-based *p*-values and the ROPE indices so that a higher value corresponds to stronger "significance".

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

Results

Impact of Sample Size

Figure 2 shows the sensitivity of the indices to sample size. The *p*-value, the *pd* and the MAP-based *p*-value are sensitive to sample size only in case of the presence of a true effect (when the null hypothesis is false). When the null hypothesis is true, all three indices are unaffected by sample size. In other words, these indices reflect the amount of observed

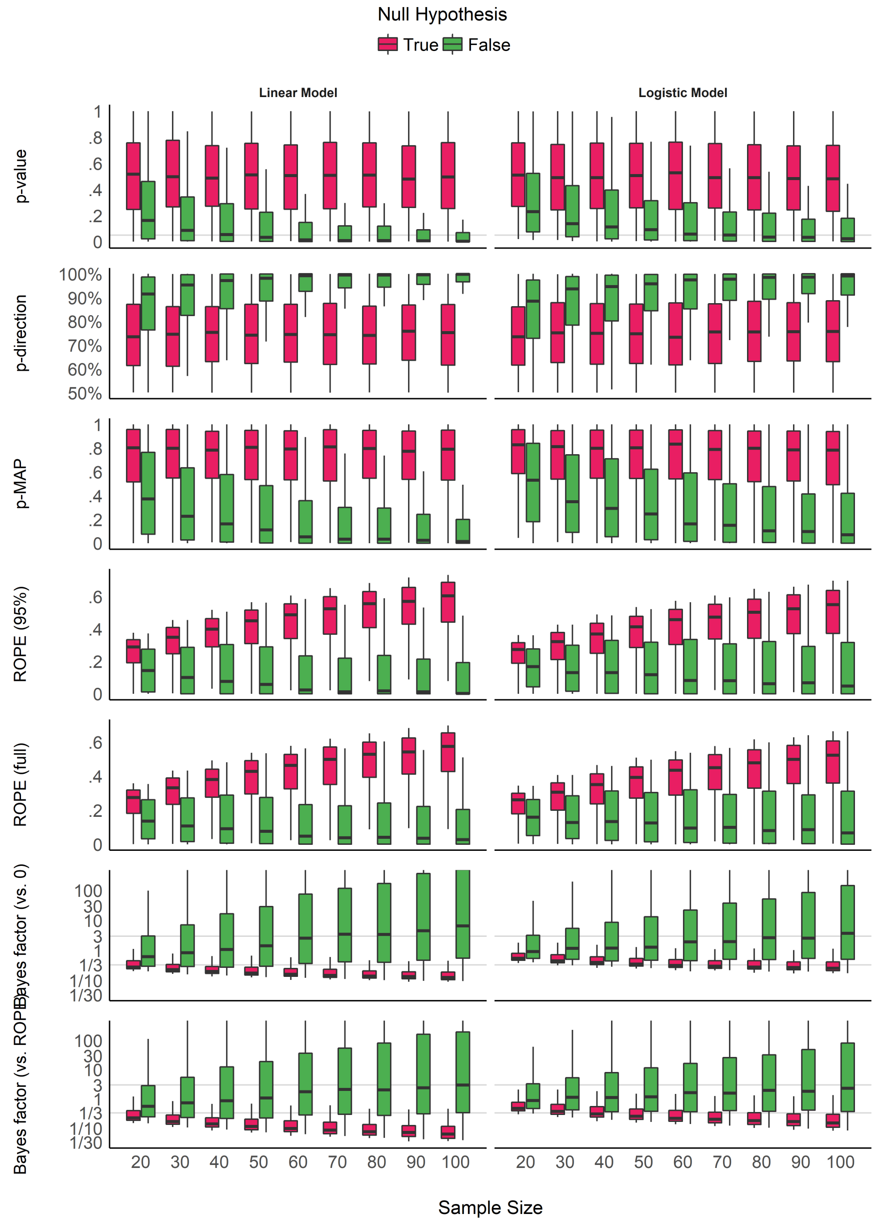

Figure 2. Impact of Sample Size on the different indices, for linear and logistic models, and when the null hypothesis is true or false. Grey vertical lines for *p*-values and Bayes factors represent commonly used thresholds.

evidence (the sample size) for the presence of an effect (i.e., against the null hypothesis being true), but not for the absence of an effect. The *ROPE* indices, however, appear as strongly modulated by the sample size when there is no effect, suggesting their sensitivity

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 / Presence of Effect	Linear Models / Absence of Effect	Logistic Models / Presence of Effect	Logistic Models / 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

to the amount of evidence for the absence of effect. Finally, the figure suggests that *BFs* are sensitive to sample size for both presence and absence of true effect.

Consistently with **Figure 2**, the model investigating the sensitivity of sample size on the different indices suggests that *BF* indices are sensitive to sample size both when an effect is present (null hypothesis is false) and absent (null hypothesis is true). *ROPE* indices are particularly sensitive to sample size when the null hypothesis is true, while *p*-value, *pd* and MAP-based *p*-value are only sensitive to sample size when the null hypothesis is false, in which case they are more sensitive than *ROPE* indices. These findings can be related to the concept of consistency: as the number of data points increases, the statistic converges toward some “true” value. Here, we observe that *p*-value, *pd* and the MAP-based *p*-value are consistent only when the null hypothesis is false. In other words, as sample size increases, they tend to reflect more strongly that the effect is present. On the other hand, *ROPE* indices appear as consistent when the effect is absent.

Finally, *BFs* are consistent both when the effect is absent and when it is present, and *BF (vs. ROPE)*, compared to *BF (vs. 0)*, is more sensitive to sample size when the null hypothesis is true, and *ROPE (full)* is overall slightly more consistent than *ROPE (95%)*.

Impact of Noise

Figure 3 shows the indices' sensitivity to noise. Unlike the patterns of sensitivity to sample size, the indices display more similar patterns in their sensitivity to noise (or magnitude of effect). All indices are unidirectional impacted by noise: as noise increases, the observed coefficients decrease in magnitude, and the indices become less "pronounced" (respectively to their direction). However, it is interesting to note that the variability of the indices seems differently impacted by noise. For the *p*-values, the *pd* and the ROPE indices, the variability increases as the noise increases. In other words, small variation in small observed coefficients can yield very different values. On the contrary, the variability of BFs decreases as the true effect tends toward 0. For the MAP-based *p*-value, the variability appears to be the highest for moderate amount of noise. This behavior seems consistent across model types.

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

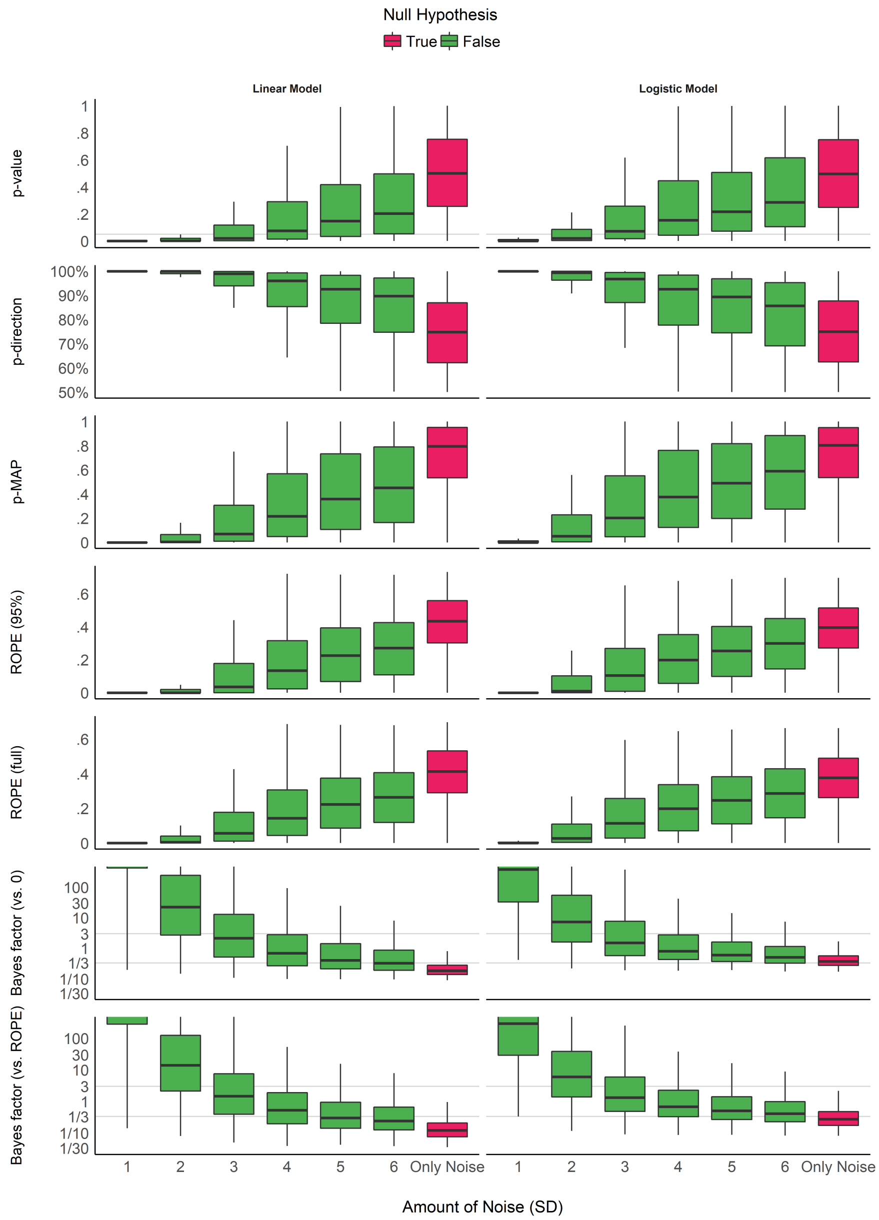

Figure 3. Impact of Noise. The noise corresponds to the standard deviation of the Gaussian noise that was added to the generated data. It is related to the magnitude the parameter (the more noise there is, the smaller the coefficient). Grey vertical lines for *p*-values and Bayes factors represent commonly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

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).

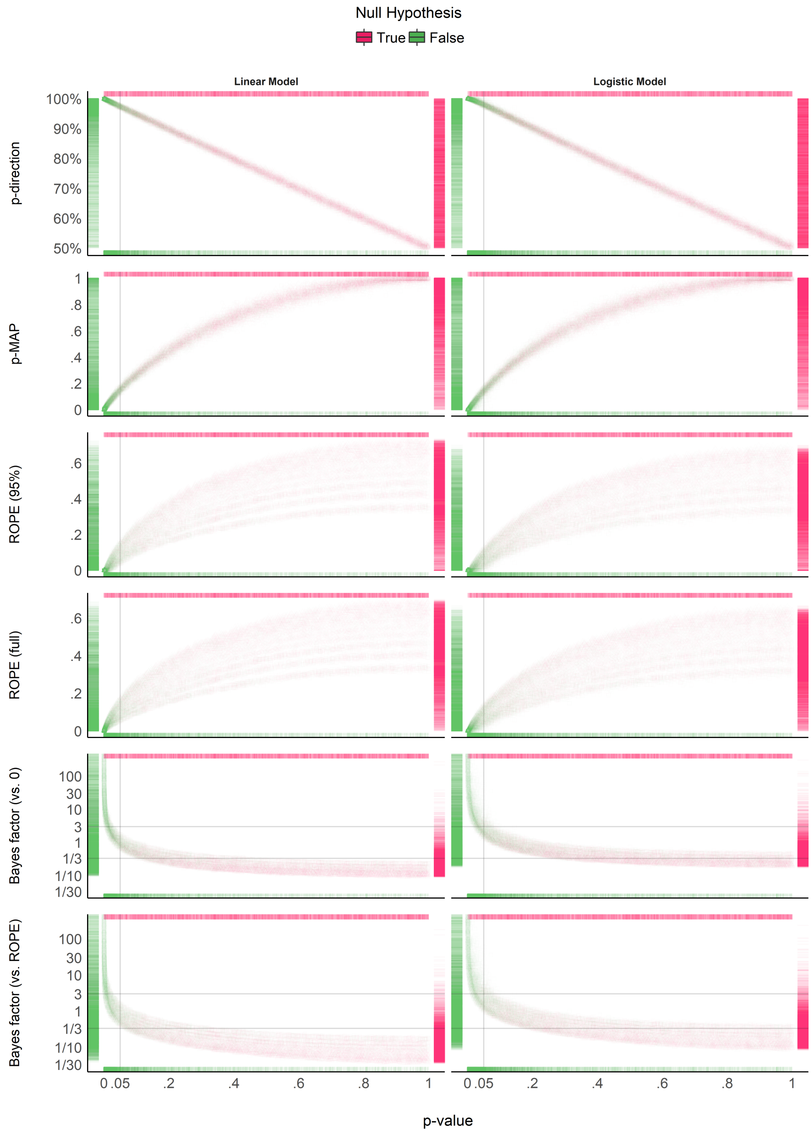

Figure 4. Relationship with the frequentist *p*-value. In each plot, the *p*-value densities are visualized by the marginal top (absence of true effect) and bottom (presence of true effect) markers, whereas on the left (presence of true effect) and right (absence of true effect), the markers represent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

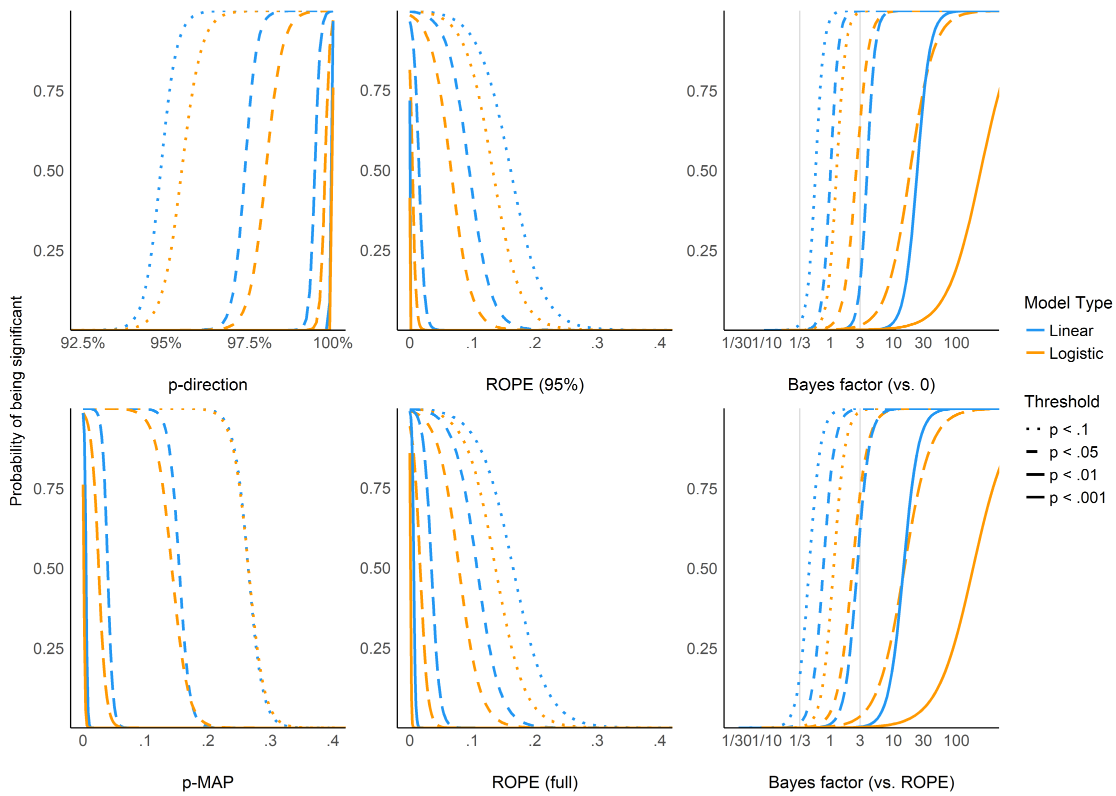

Figure 5. The probability of reaching different *p*-value based significance thresholds (.1, .05, .01, .001 for solid, long-dashed, short-dashed and dotted lines, respectively) for different values of the corresponding Bayesian indices.

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 (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.

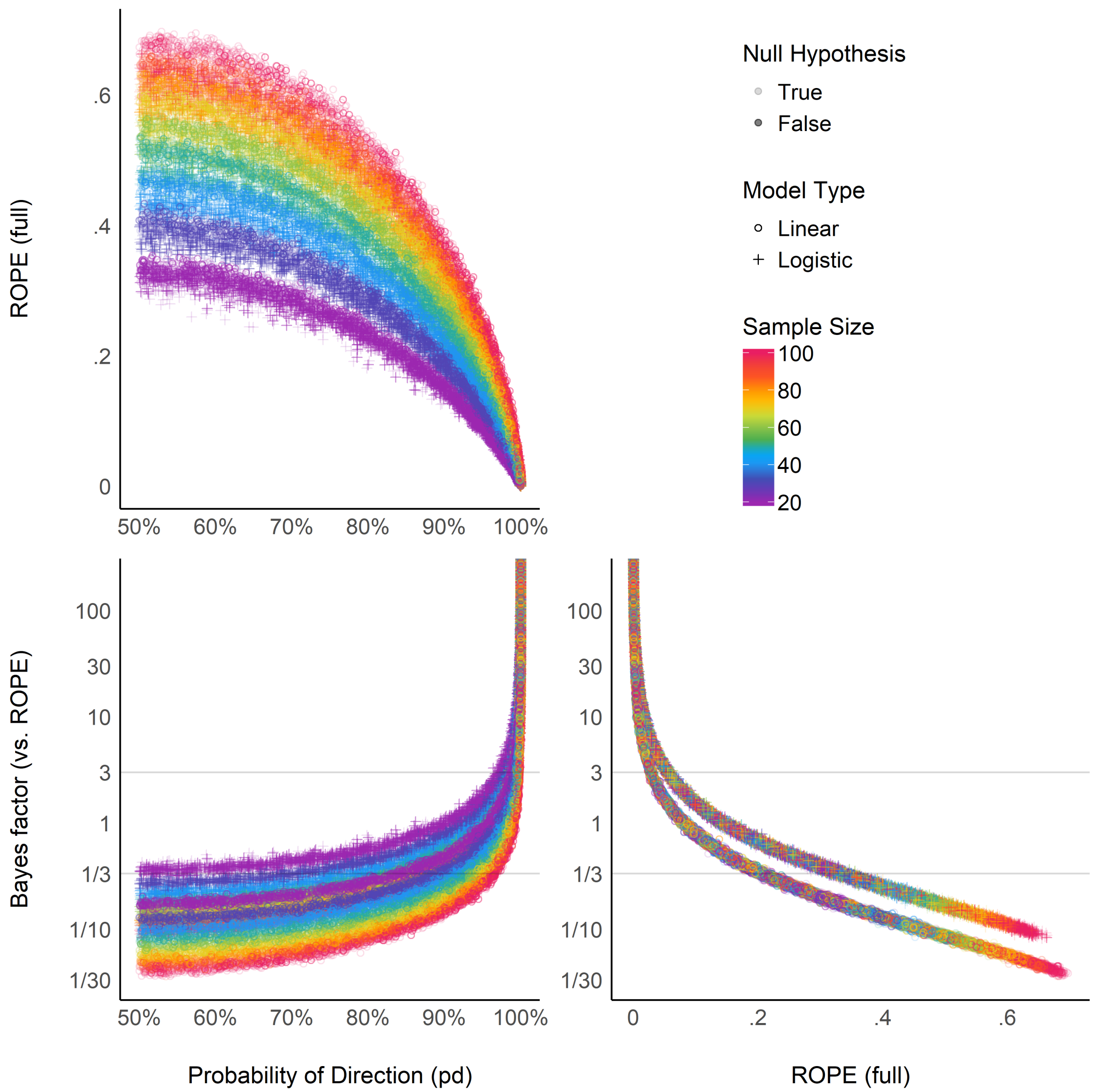

Figure 6. Relationship between three Bayesian indices: The Probability of Direction (*pd*), the percentage of the full posterior distribution in the ROPE, and the Bayes factor (*vs.* ROPE).

Discussion

Based on the simulation of linear and logistic models, the present work aimed to compare 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.

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 only to the amount of evidence for the alternative hypothesis (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* regarding 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, ROPE seems particularly suited to provide evidence in favor of the null hypothesis. Consistent with this, combining Bayes factors with 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. Bayes factors, however, appear to have 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 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 preclude, however, that a change in the general paradigm of significance seeking and "p-hacking" is necessary, and that Bayesian indices are fundamentally different from the frequentist *p*-value, rather than mere approximations or equivalents.

Critically, while the purpose of these indices was solely referred to as *significance* until now, we would like to emphasize the nuanced perspective of existence-significance testing as a dual-framework for parameter description and interpretation. The idea supported here is that there is a conceptual and practical distinction, and possible dissociation to be made, between an effect's existence *and* its significance. In this context, *existence* is simply defined as the consistency of an effect in one particular direction (i.e., positive or negative), without any assumptions or conclusions as to its size, importance, relevance or meaning. It is an objective feature of an estimate (tied to its uncertainty). On the other hand, *significance* would be here re-framed following its original literally definition such as "being worthy of attention" or "importance". An effect can be considered significant if its magnitude is higher than some given threshold. This aspect can be explored, to a certain extent, in an objective way with the concept of *practical equivalence* (Kruschke, 2014; Lakens, 2017; Lakens et al., 2018), which suggests the use of a range of values assimilated to the absence of an effect (ROPE). If the effect falls within this range, it is considered to be non-significant *for practical reasons*: the magnitude of the effect is likely to be too small to be of high importance in real-world scenarios or applications. Nevertheless, *significance* also withholds a more subjective aspect, corresponding to its contextual meaningfulness and relevance. This, however, is usually dependent on the literature, priors, novelty, context

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.

or field, and thus cannot be objectively or neutrally assessed using a statistical index alone.

While indices of existence and significance can be numerically related (as shown in our results), the former is conceptually independent from the latter. For example, an effect for which the whole posterior distribution is concentrated within the [0.0001, 0.0002] range would be considered to be positive with a high level of certainty (and thus, *existing* in that direction), but also not significant (i.e., too small to be of any practical relevance). Acknowledging the distinction and complementary nature of these two aspects can in turn enrich the information and usefulness of the results reported in psychological science (for practical reasons, the implementation of this dual-framework of existence-significance testing is made straightforward through the *bayestestR* open-source package for R; Makowski et al., 2019). In this context, the *pd* and the MAP-based *p*-value appear as indices of effect existence, mostly sensitive to the certainty related to the direction of the effect. ROPE-based indices and Bayes factors are indices of effect significance, related to the magnitude and the amount of evidence in favor of it (see also a similar discussion of statistical significance vs. effect size in the frequentist framework; e.g., Cohen, 1994)

The inherent subjectivity related to the assessment of significance is one of the practical limitations of ROPE-based indices (despite being, conceptually, an asset, allowing for contextual nuance in the interpretation), as they require an explicit definition of the non-significant range (the ROPE). Although default values have been reported in the literature (for instance, half of a "negligible" effect size reference value; Kruschke, 2014), it is critical to reproducibility and transparency that the researcher's choice is explicitly stated (and, if possible, justified). Beyond being arbitrary, this range also has hard limits (for instance, contrary to a value of 0.0499, a value of 0.0501 would be considered non-negligible if the range ends at 0.05). This reinforces a categorical and clustered perspective of what is by essence a continuous space of possibilities. Importantly, as this range is fixed to the scale of the response (it is expressed in the unit of the response), ROPE indices are sensitive to changes in the scale of the predictors. For instance, negligible results may change into non-negligible results when predictors are scaled up (e.g. reaction times expressed in seconds instead of milliseconds), which one inattentive or malicious researcher could misleadingly present as "significant" (note that indices of existence, such as the *pd*, would not be affected by this). Finally, the ROPE definition is also dependent on the model type, and selecting a consistent or homogeneous range for all the families of models is not straightforward. This can make comparisons between model types difficult, and an additional burden when interpreting ROPE-based indices. In summary, while a well-defined ROPE can be a powerful tool to give a different and new perspective, it also requires extra caution on the paets of authors and readers.

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

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

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

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

After all the criticism regarding the frequentist *p*-value, it may appear contradictory to suggest the usage of its Bayesian empirical equivalent. The subtler perspective that we support is that the *p*-value is not an intrinsically bad, or wrong, index. Instead, it is its misuse, misunderstanding and misinterpretation that fuels the decay of the situation into the crisis. *Interestingly, the proximity between the* pd *and the* p-*value follows the original definition of the latter (Fisher, 1925) as an index of effect* existence *rather than* significance *(as in "worth of interest"; Cohen, 1994).* Addressing this confusion, the Bayesian equivalent has an intuitive meaning and interpretation, contributing to making more obvious the fact that all thresholds and heuristics are arbitrary. In summary, the mathematical and interpretative transparency of the *pd*, and its conceptualization as an index of effect existence, offer valuable insight into the characterization of Bayesian results, and its practical proximity with the frequentist *p*-value makes it a perfect metric to ease the transition of psychological research into the adoption of the Bayesian framework.

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

Reporting Guidelines

How can the current observations be used to improve statistical good practices in psychological science? Based on the present comparison, we can start outlining the following guidelines. As *existence* and *significance* are complementary perspectives, we suggest using at minimum one index of each category. As an objective index of effect 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 aid in 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% may be relevant reference points (i.e., easy-to-remember values) for the *pd*. 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 ($BF_{ROPE} = 3.44$) [*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 to be 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 to be small ($Std.Median = -0.29$) [*standardized coefficient*] and significant (0.82% in $ROPE$) [*ROPE (full)*]".

Data Availability

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

Ethics Statement

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

Author Contributions

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

Conflict of Interest Statement

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

Acknowledgments

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

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management*, 912–923.

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 1–7.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … others. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of 'playing the game' it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17.

Cohen, J. (1988). Statistical power analysis for the social sciences.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

De Santis, F. (2007). Alternative bayes factors: Sample size determination and discriminatory power assessment. *Test*, *16*(3), 504–522.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Ellis, S., & Steyn, H. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (p-values): Research note. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, *12*(4), 51–53.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, 2515245918773087.

Etz, A., & Vandekerckhove, J. (2016). A bayesian perspective on the reproducibility project: Psychology. *PloS One*, *11*(2), e0149794.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*(2), 119–126.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., … Goodman, O. (2004). Reform of statistical inference in psychology: The case ofMemory & cognition. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 312–324.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver; Boyd.

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746–750.

Gelman, A. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16–23. https://doi.org/10.1177/0146167217729162

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014).

Bayesian data analysis. (Third edition). Boca Raton: CRC Press.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2019). Rstanarm: Bayesian applied regression modeling via Stan. Retrieved from http://mc-stan.org/

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

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

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

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

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

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

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

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

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

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

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

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963.

Lüdecke, D., Waggoner, P., & Makowski, D. (2019). Insight: A unified interface to access information from model objects in r. *Journal of Open Source Software*, *4*(38), 1412. https://doi.org/10.21105/joss.01412

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

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019). bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Marasini, D., Quatto, P., & Ripamonti, E. (2016). The use of p-values in applied research: Interpretation and new trends. *Statistica*, *76*(4), 315–325.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

McElreath, R. (2018). *Statistical rethinking*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315372495

Mills, J. A. (2017). Objective bayesian precise hypothesis testing. *University of Cincinnati*

[Original Version: 2007].

Mills, J. A., & Parent, O. (2014). Bayesian mcmc estimation. In *Handbook of regional science* (pp. 1571–1595). Springer.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150–152. https://doi.org/10.1038/506150a

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robert, C. P. (2014). On the jeffreys-lindley paradox. *Philosophy of Science*, *81*(2), 216–232.

Robert, C. P. (2016). The expected demise of the bayes factor. *Journal of Mathematical Psychology*, *72*, 33–37.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part iv: Parameter estimation and bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Spanos, A. (2013). Who should be afraid of the jeffreys-lindley paradox? *Philosophy of Science*, *80*(1), 73–93.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282.

Szucs, D., & Ioannidis, J. P. (2016). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *BioRxiv*, 071530.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Lee, M., Rouder, J., & Morey, R. (2019, August). Another statistical paradox. Retrieved from http://www.ejwagenmakers.com/submitted/AnotherStatisticalParadox.pdf

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … others. (2018). Bayesian inference for psychology. Part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for bayesian hypothesis testing in psychological science. *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions*, 123–138.

Wasserstein, R. L., Lazar, N. A., & others. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.