Indices of Effect Existence and Significance in the Bayesian Framework

# Abstract

There is now a general agreement that the Bayesian statistical framework is the right way to go for psychological science. However, its flexible nature is both its power and its weakness, for there is little agreement about what indices should be computed or reported. This lack of a consensual index for the existence of an effect, such as the frequentist *p*-value, further contributes to the unnecessary opacity that many non-familiar readers perceive in Bayesian statistics. Thus, this study describes and compares several indices of effect existence, provide intuitive visual representation of the “behaviour” of such indices in relationship with common sources of variance such as sample size and frequentist significance. The results contribute to the development of an intuitive understanding of the values that researchers report and allow to draw recommendations for Bayesian statistics description, critical for the standardization of scientific reporting.

# Introduction

The Bayesian framework is quickly gaining popularity among psychologists and neuroscientists (Andrews & Baguley, 2013). Reasons to prefer this approach are reliability, better accuracy in noisy data, better estimation for small samples, less proneness to type I error, the possibility of introducing prior knowledge into the analysis and, critically, 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 misuse of which has been shown to critically contribute to the reproducibility crisis of psychological science (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Szucs & Ioannidis, 2016). In response, there is a general agreement that the generalization and utilization of the Bayesian framework is a 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).

While the probabilistic reasoning put forth by the Bayesian framework is pervading most of data science aspects, it is particularly well established for statistical modelling. This facet, on which psychological science massively rely, could roughly be grouped into two soft-edged categories: predictive and structural modelling. Although a statistical model can (often) serve both purposes, predictive modelling is devoted to building and finding the best model that accurately predicts a given outcome. It is centered around concepts such as fitting metrics, predictive accuracy and model comparison. At the extremum of this dimension lie deep learning models, used for their strong predictive power, sometimes at the expense of Human readability (these models have been often referred to as “black-boxes”, emphasizing the difficulty to appraise their internal functioning; Burrell, 2016; Castelvecchi, 2016; Snoek, Larochelle, & Adams, 2012). On the other side, psychologists often use simpler models (for instance, models based on the general linear framework) to explore their data. Within this framework, the goal switches from building the best model, to understanding the parameters inside the model. In reality, the pipeline often starts with predictive modelling (“what is the best model for describing the world, *i.e.*, the observed variable”) and then seemingly transition to structural modelling (“given this model of the world, how do the effects, *i.e.*, the model parameters, influence the outcome”), the conclusions of which often rely on some index of effect “significance”.

It is important to note that the importance and focus on significance assessment might differ across fields. For instance, in applied physics or engineering, the goal of a study is often to precisely *estimate* (quantify) a given effect and its size (the question being, for example, whether a given coefficient is closer to 7 or to 9, or whether a difference is moderate or big). The Bayesian framework has demonstrated its superiority in addressing this type of questions, with its probabilistic framework leading to more accurate estimations and allowing to quantify the inherent uncertainty associated with them. On the contrary, in other fields, such as experimental psychology, the implicit aim has often been focused on *significance*: “are these two conditions different?”, “is there (any) correlation between these two variables?”. In other words, focus has been on being able to conclude whether or not a given parameter affects substantially the outcome with little care to its precise quantification.

Despite a general agreement on its defects and recurring waves of attacks occurring for decades (Gardner & Altman, 1986; anderson2000null; finch2004reform; fidler2004editors; Cohen, 2016), the tenacity and resilience of the *p*-value as an index of significance is remarkable, being still widely used and taught (although some journals have taken a radical step by banning them, *e.g.*, Gill (2018); expecting to shift the perspective to effect size estimation). This endurance might be informative on how such indices, and the accompanying heuristics applied to interpret them (*e.g.*, *.05*, *.01* and *.001*), 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 (*e.g.*, in medical settings). In spite of the statistical perspective, in which the absence of any hard mathematical thresholds or discrete categories naturally leads to the desire for a nuanced and subtle view, decisionners might often seek a threshold value to rationalize their acts and decisions. This practical need for significance assessment might be one of the mechanism fostering the wide adoption of the frequentist *p*-value and its arbitrary interpretation clusters. Unfortunately, these heuristics have severely rigidified, with meeting significance threshold becoming a goal unto themselves rather than a tool for understanding the data (Cohen, 2016; Kirk, 1996). This is particularly problematic given that *p*-values and NHST can only be used to reject the null hypothesis and cannot be used to accept the null as true (Wagenmakers, 2007), thus making “discovery of effects” the de-facto focus of scientific research.

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 (as opposed to the obscure mathematical definition of the *p*-value that contributes to its common misinterpretation). Another improvement could be found in providing an intuitive (*e.g.*, visual) understanding of the behaviour 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.

So what does the Bayesian framework offer as alternatives for the *p*-value? Bayesian testing indices could be roughly grouped into three overlapping categories: Bayes factors, posterior indices and 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). They provide many advantages over the *p*-value by having a straightforward interpretation as well as allowing to quantify evidence in favour of the null hypothesis (Dienes, 2014; Jarosz & Wiley, 2014). Nonetheless, its use for parameters description in complex models is still a matter of debate (Heck, 2019; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), and its use is highly dependent on the specification of priors of both compared models (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. While the simplicity of their computation and interpretation is an asset, it also means they are limited in the information that they provide. Importantly, Bayes factors and posterior indices are both the “natural, direct, and unavoidable consequence of Bayes’ rule” (Rouder, Haaf, & Vandekerckhove, 2018, p. 106). 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]). It is interesting to note that this perspective unites Bayesian indices 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).

Curiously, despite the richness provided by the Bayesian framework and the availability of multiple indices, no consensus has yet emerged on the ones to be used. On the contrary, the literature continues to bloom in a raging debate, often polarized between proponents of the Bayes factor as the supreme index and its detractors (*e.g.*, Spanos, 2013; Robert, 2014; robert2016expected; Wagenmakers, Lee, Rouder, & Morey, 2019), with strong theoretical arguments being developed on both sides. Unfortunately, no practical, empirical and direct comparison between these indices has ever 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 the 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 “behaviour” in relationship with major sources of variance, such as sample size and noise. 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, in turn, prevent the crystallization of the possible heuristics and categories derived from such indices, as has unfortunate occurred in the used of *p*-values.

Thus, based on the simulation of multiple 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 “behaviour” of such indices in relationship with sample size, noise and effect presence, as well as their relationship to frequentist *p*-value (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

The simulations of datasets suited for linear or logistic regression, we’re 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 perfect separated by *x* (in the case of no effect, a *y* variable was created that was independent of *x*). Finally, a Gaussion noise was added to the *x* variable (the error).

The simulation aimed at modulating the following characteristics: *data type* (*i.e.*, used for linear or logistic regression), *Sample size* (from 20 to 100 by steps of 10), *“true” effect* (original regression coefficient from which data is drawn prior to noise addition, 1 - presence of effect or 0 - absence of effect) and *noise* (Gaussian noise applied to the predictor with SD uniformly spread between 0.666 and 6.66, with 1000 different values). We generated a dataset for each combination of these characteristics, resulting in a total of 36000 (2 \* 2 \* 9 \* 1000) datasets. The code used for generation is available on Github (<https://github.com/easystats/easystats/tree/master/publications/makowski_2019_bayesian/data>). Please 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. The Bayesian models had default, mildly informative priors (normal distribution with mean 0 and SD 1) over the parameter of interest, and were fitted using MCMC (4 chains of 2000 iterations, half of which used for warm-up). For all of the simulated models, we computed seven indices related to the effect of *x*. These are available and described in details in the *bayestestR* R package **(BAYESTESTR CITATION)**.

### Frequentist *p*-value

Based on the frequentist regression, this index 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 **(BAYESTESTR CITATION)**.

### MAP-based *p*-value

The *MAP-based p-value* (also referred to as the “Bayesian *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%) and ROPE (full)

The *ROPE (95%)* refers to the percentage of the 95% HDI that lies within the ROPE. As suggested by Kruschke (2014), the Region of Practical Equivalence (ROPE) was defined as ranging from -0.1 to 0.1 for linear regressions and its equivalent, -0.18 to -0.18, for logistic models (based on the formula to convert log odds ratios to standardized differences; Cohen, 1988). The *ROPE (full)* refers to the percentage of the whole posterior distribution that lies within the ROPE.

### Bayes factor (*vs.* 0) and Bayes factor (*vs.* ROPE)

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(s), relative to the prior distribution, thus indicating if the null hypothesis has become less or more likely given the observed data. We created two indices corresponding to two definitions for the null. In the case of testing against a point null (0), a Savage-Dickey density ratio was computed, 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). We also computed a *BF* against the range of negligible values (defined here same as for the ROPE indices), by comparing the prior and posterior odds of the parameter falling within vs. outside the ROPE (see *Non-overlapping Hypotheses* in Morey & Rouder, 2011).

## Data Analysis

The aim of this study is two-fold: 1) compare between Bayesian indices of effect existence and significance, 2) provide visual guides for an intuitive understanding of the numeric values in relation with a known frame of reference (the frequentist NHST framework). Thus, we will start by 1) presenting the relationship between these indices and main sources of variance, such as effect existence, sample size and noise. 2) Compare Bayesian indices with the frequentist *p*-value and its commonly used thresholds (.05, .01, .001). Taken together, these results will help us to outline numeric guides to ease the reporting and interpretation of the indices.

# Results

## Impact of Sample Size

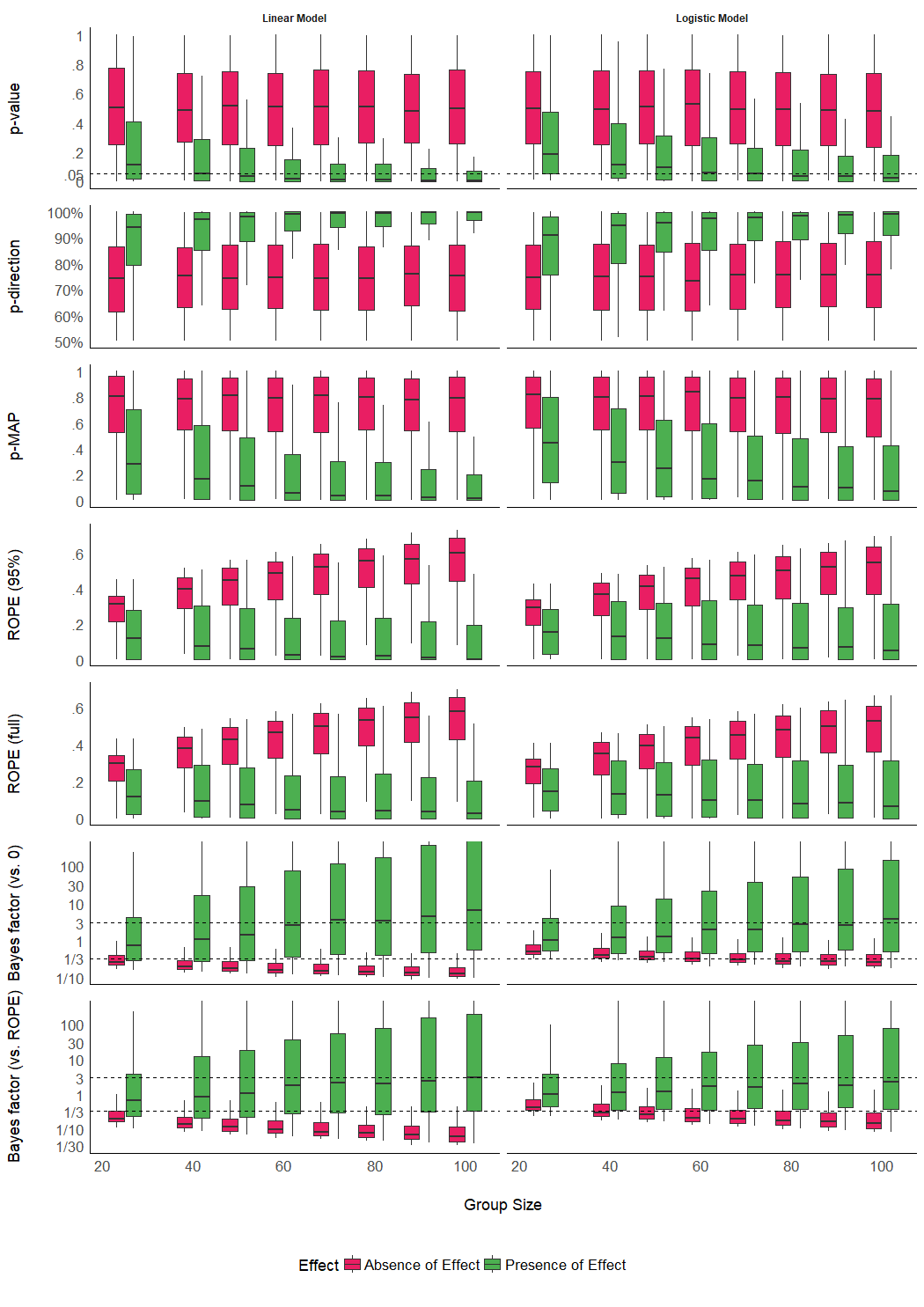


Figure 1. Impact of Sample Size. Dashed lines for p-values and Bayes factors represent commonly used thresholds.

> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.  
> Can't calculate log-loss.

Table 1. Sensitivity to sample size.

|  |  |  |
| --- | --- | --- |
| Type | Parameter | Coefficient |
| Linear | BF\_ROPE\_log | 0.019 |
| Linear | BF\_log | 0.028 |
| Linear | p\_value | 0.087 |
| Linear | p\_direction | 0.091 |
| Linear | p\_MAP | 0.120 |
| Linear | ROPE\_95 | 0.162 |
| Linear | ROPE\_full | 0.168 |
| Logistic | BF\_ROPE\_log | 0.007 |
| Logistic | BF\_log | 0.014 |
| Logistic | p\_value | 0.091 |
| Logistic | p\_direction | 0.092 |
| Logistic | p\_MAP | 0.139 |
| Logistic | ROPE\_95 | 0.164 |
| Logistic | ROPE\_full | 0.170 |

The analysis suggests that ROPE-based indices are the most sensitive to sample size, followed by the MAP-based *p*, the *pd*, the *p*-value. The BF (vs. 0) and the BF (vs. ROPE) are the less sensitive to sample size. This pattern is evident across model types. Note that the *p*-value and *pd* are only affected by sample size when an effect is present, but *not* in the absence of an effect, whereas the ROPE-based indices and the Bayes factors are affected by sample size

## Impact of Noise

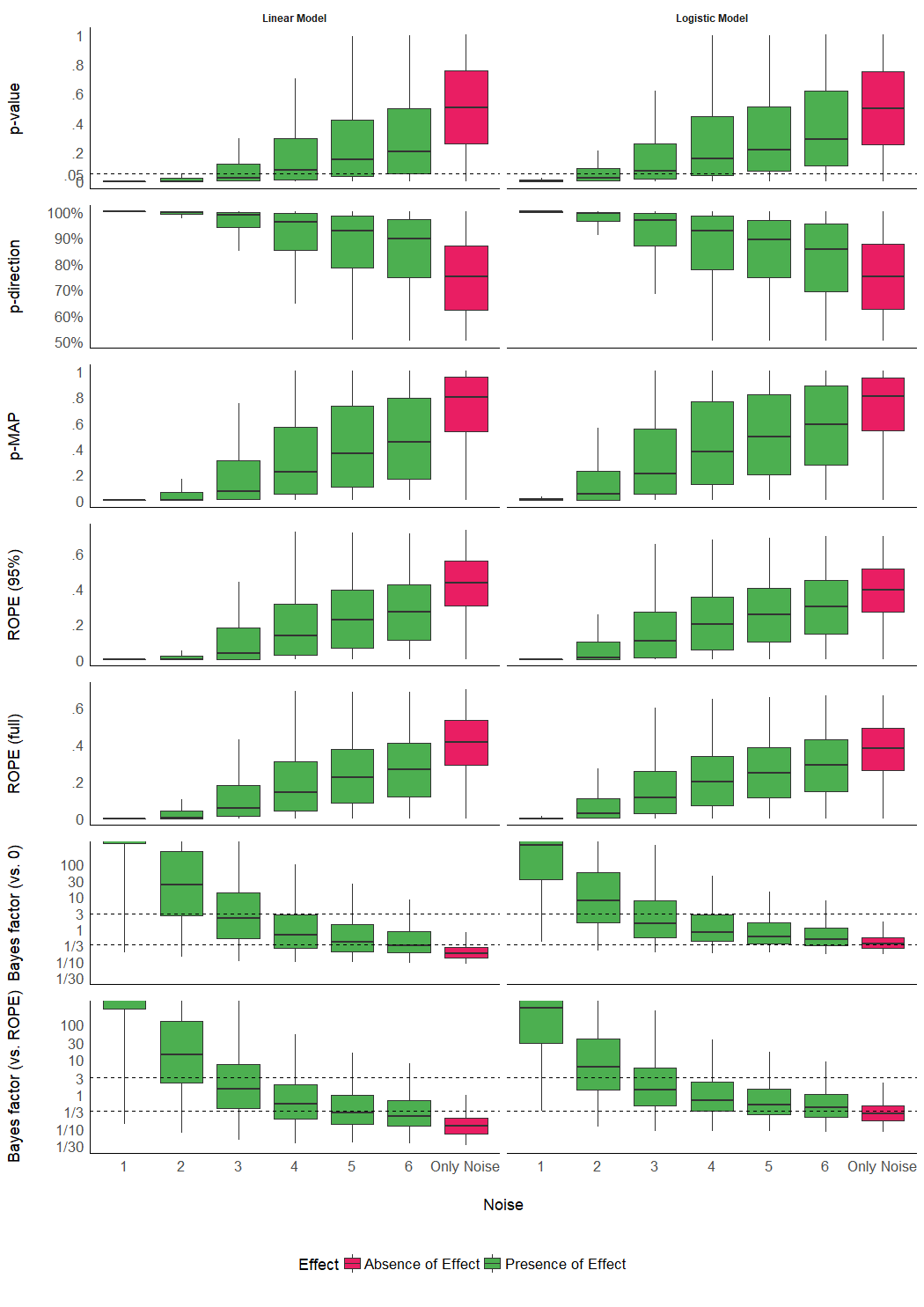


Figure 2. Impact of Noise. Dashed lines for p-values and Bayes factors represent commonly used thresholds. The scale is capped for the Bayes factors as these extend to infinity (and beyond).

Table 2. Sensitivity to noise.

|  |  |  |  |
| --- | --- | --- | --- |
| Effect | Type | Parameter | Coefficient |
| Absence | Linear | BF\_log | 0.000 |
| Absence | Linear | BF\_ROPE\_log | 0.000 |
| Absence | Linear | ROPE\_95 | 0.005 |
| Absence | Linear | ROPE\_full | 0.005 |
| Absence | Linear | p\_MAP | 0.006 |
| Absence | Linear | p\_value | 0.011 |
| Absence | Linear | p\_direction | 0.011 |
| Absence | Logistic | BF\_log | 0.000 |
| Absence | Logistic | BF\_ROPE\_log | 0.000 |
| Absence | Logistic | ROPE\_95 | 0.001 |
| Absence | Logistic | ROPE\_full | 0.001 |
| Absence | Logistic | p\_MAP | 0.002 |
| Absence | Logistic | p\_value | 0.005 |
| Absence | Logistic | p\_direction | 0.006 |
| Presence | Linear | BF\_log | 0.207 |
| Presence | Linear | BF\_ROPE\_log | 0.218 |
| Presence | Linear | p\_value | 0.320 |
| Presence | Linear | p\_direction | 0.323 |
| Presence | Linear | ROPE\_95 | 0.404 |
| Presence | Linear | ROPE\_full | 0.412 |
| Presence | Linear | p\_MAP | 0.499 |
| Presence | Logistic | BF\_log | 0.119 |
| Presence | Logistic | BF\_ROPE\_log | 0.131 |
| Presence | Logistic | p\_value | 0.376 |
| Presence | Logistic | p\_direction | 0.378 |
| Presence | Logistic | ROPE\_95 | 0.423 |
| Presence | Logistic | ROPE\_full | 0.423 |
| Presence | Logistic | p\_MAP | 0.562 |

## Presence *vs.* Absence of Effect

Table 3. Performance comparison of the indices in predicting the presence of an effect.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model\_Type | Model | AIC | BIC | R2\_Tjur | RMSE | PCP | BF | BF\_log |
| linear | p\_value | 18500 | 18531 | 0.33 | 1.01 | 0.66 |  |  |
| linear | p\_direction | 18461 | 18492 | 0.33 | 1.01 | 0.66 | 3.32e+08 | 19.62 |
| linear | p\_MAP | 16850 | 16882 | 0.40 | 0.97 | 0.70 | Inf | Inf |
| linear | ROPE\_95 | 16358 | 16389 | 0.42 | 0.95 | 0.71 | Inf | Inf |
| linear | ROPE\_full | 16234 | 16265 | 0.42 | 0.95 | 0.71 | Inf | Inf |
| linear | BF\_log | 15032 | 15063 | 0.45 | 0.91 | 0.73 | Inf | Inf |
| linear | BF\_ROPE\_log | 14984 | 15015 | 0.45 | 0.91 | 0.73 | Inf | Inf |
| binary | p\_value | 20486 | 20517 | 0.23 | 1.07 | 0.62 |  |  |
| binary | p\_direction | 20501 | 20532 | 0.23 | 1.07 | 0.62 | 4.78e-04 | -7.65 |
| binary | p\_MAP | 19200 | 19231 | 0.29 | 1.03 | 0.65 | 1.75e+279 | 642.98 |
| binary | ROPE\_95 | 19018 | 19049 | 0.30 | 1.03 | 0.65 | Inf | Inf |
| binary | ROPE\_full | 18953 | 18985 | 0.30 | 1.03 | 0.65 | Inf | Inf |
| binary | BF\_log | 17975 | 18007 | 0.33 | 1.00 | 0.66 | Inf | Inf |
| binary | BF\_ROPE\_log | 17934 | 17965 | 0.33 | 1.00 | 0.67 | Inf | Inf |

For each index and each model type, we fitted a (frequentist) logistic regression to predict the presence (*vs.* absence) of effect, adjusted for noise and sample size. The comparison of the performance of these models (AIC, BIC and Tjur’s R2) revealed a consistent pattern across model type (*i.e.*, similar for linear and logistic models), suggesting that *BF (vs. ROPE)* is the best index to discriminate between the presence and the absence of an effect, followed by *BF (vs. 0)*, *ROPE (full)*, *ROPE (95\_%)*, *p-MAP*, *p-direction* and the frequentist *p*-value. The Bayes factor (against the frequentist *p*-value model, computed via BIC approximation Wagenmakers, 2007), used here as a measure of relative performance, supported this conclusion.

## Relationship with the frequentist *p*-value

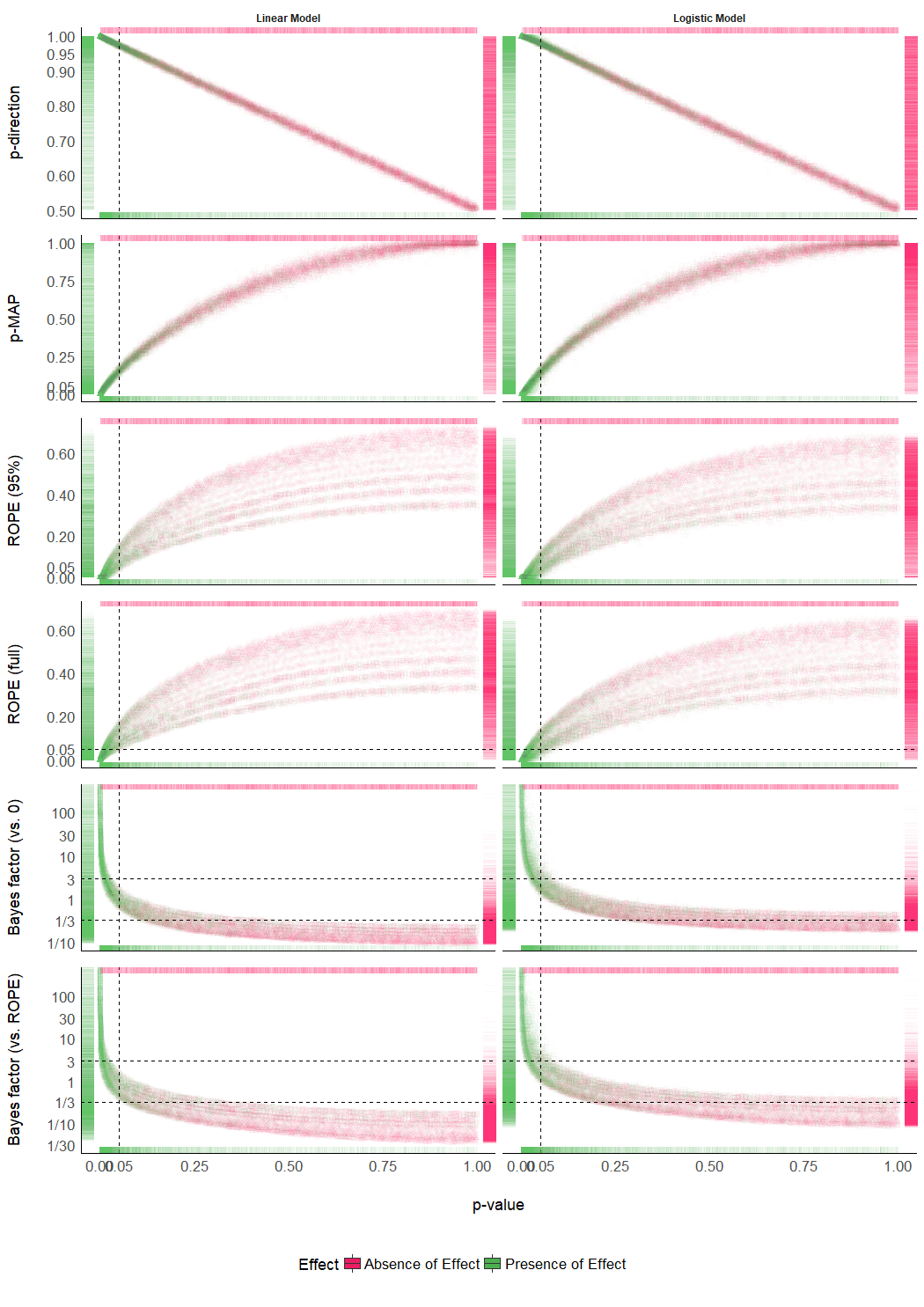


Figure 3. Relationship with the frequentist p-value. In each plot, the p-value densities can be visualised by the marginal top (null effect) and bottom (true effect) markers, whereas on the left (true effect) and right (null effect), they represent the density of the index of interest. Different point shapes, representinng different sample sizes, illustrate its impact on the percentages in ROPE, for which each curve line is associated with one sample size.

**Figure 3** suggests that the *pd* has a 1:1 correspondence with the frequentist *p*-value (through the formula ). *BF* indices still appear as having a strong relationship (although severely non-linear) with the frequentist index, which is to be expected since the *BF* is consistent (i.e., it approaches the “true” bound as sample size increases) both when the null is true and when the alternative is ture (as can be seen in the marginal distributions, marked by colored dashes in the plot margins), whereas the *p*-value is only consistent when the alternative is true, but has a uniform distribution [0-1] when the null is true (Rouder & Morey, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). *ROPE*-based percentages appear to be only weakly related to *p*-values.

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

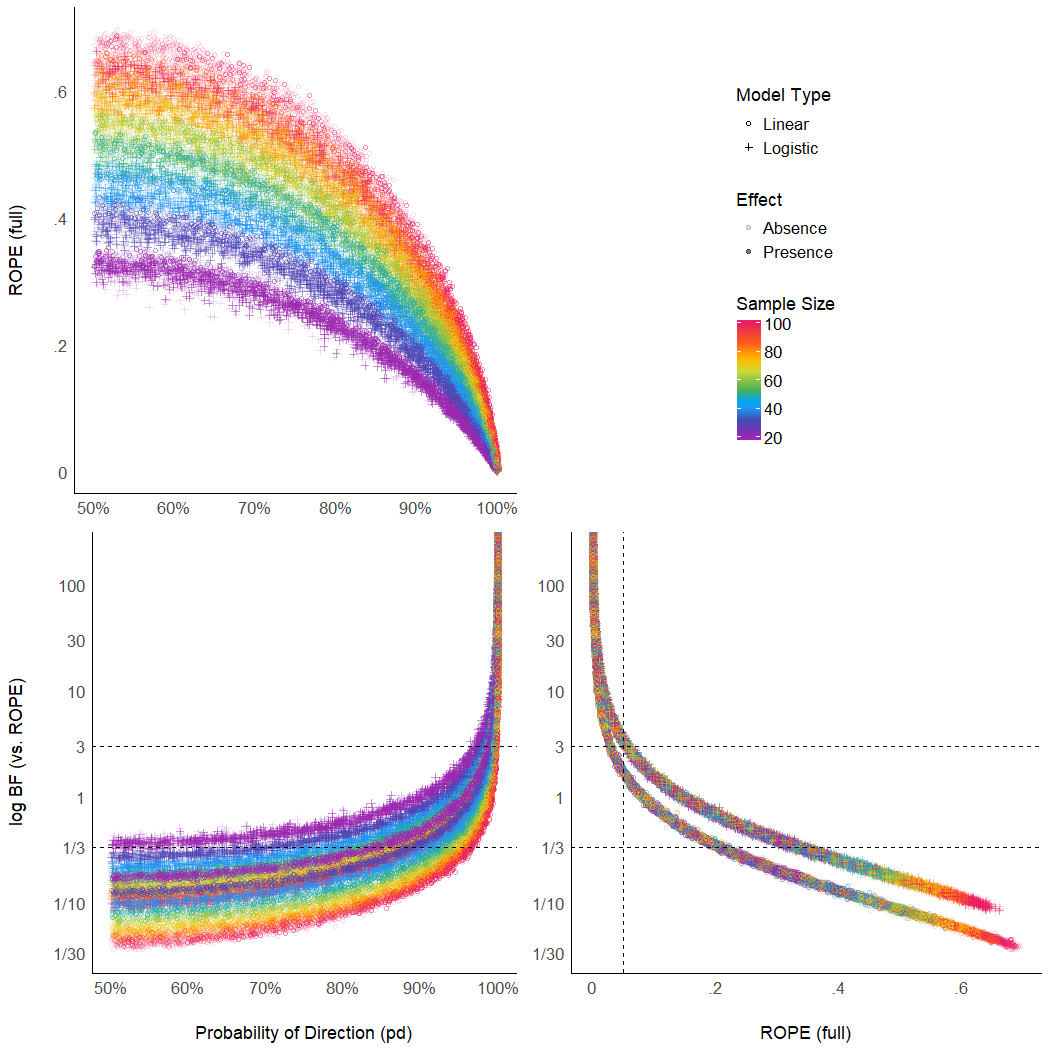


Figure 5. Relationship between three Bayesian indices.

The **Figure 5** suggests that the relationship between the *ROPE (full)* and the *pd* might be strongly affected by the sample size, and the relationship between *BF (vs. ROPE)* and the *pd* might be subject to differences across model types (though see next paragraph). Moreover, the *ROPE (full)* and the *BF (vs. ROPE)* seem very closely related within the same model type. These results reflect *ROPE (full)* and *BF (vs. ROPE)*’s consistency both when the null is true and when the alternative is true, where the *pd*, being equivalent to the *p*-value, is only consistent when the null is true.

Note also the similarity between *ROPE (full)* and *BF (vs. ROPE)*; this is expected as mathematically, the relationship between *BF (vs ROPE)* and *ROPE (full)* is:

Where is the *prior* odds of falling within the ROPE. Since this odds-ratio is dependent only on the shape of the prior distributions and the definition of the range of the ROPE, it is a constant that *normalizes* the posterior ROPE odds relative to the prior ROPE odds. This is also why *ROPE (full)* and *BF (vs. ROPE)* were found to have overall the same sensitivity in discriminating between null and true effects (but different criterion for decision making - not discussed here).

# Discussion

Table 4. 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 favouring 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 favouring 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 with 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 with the probability mass has shifted away from or towards 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. | A ROPE range needs to be arbitrarily defined. Sensitive to the scale (the unit) of the predictors. Sensitive to selection of prior distribution shape, location and scale. |

Based on the simulation of multiple linear and logistic models, the present work aimed at comparing several Bayesian indices of effect existence, provide visual representations of the “behaviour” of such indices in relationship with important sources of variance such as sample size, noise, presence of effect, as well as comparing them with the well-known and widely used frequentist *p*-value and its arbitrary interpretation heuristics.

The results tend to suggest that indices could be separated into two categories. Consistency talk.

Noise.

Discrimination analysis (see **Table 4**). Performance talk.

MSB: Looking at the models, we can talk about how p-values have some discriminatory power, but how the other indices are much better (besides pd)

MSB: interval BFs discriminate better between null (or negligible) and true (un-negligible) effects (Morey & Rouder, 2011; Rouder & Morey, 2012)

What is the point of comparing Bayesian indices with the frequentist *p*-value, especially after having pointed out to its many flaws? Indeed, while this comparison may seem counter-intuitive or wrong (as the 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 the sectarism and isolation caused by a dogmatic approach to a framework. This does not preclude, however, that a change in the general paradigm of significance seeking in necessary, and that Bayesian indices are fundamentally different from the frequentist *p*, rather than mere approximations or equivalents.

Critically, while the purpose of these indices was solely termed as *significance* until now, we would like to emphasize the nuanced perspective of the existence-significance testing as a dual-framework for parameters 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 *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 (“being worthy of attention; importance”), which a neutral approach would link with the concept of effect size. An effect can be considered significant if its magnitude is higher than a 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 absence of effect (the ROPE). If the effect falls within this range, it is considered as 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. 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 or field, and thus cannot be objectively or neutrally assessed with a statistical index.

Interestingly, the weight of one or the other aspect of the EXIT framework (Effect eXistence and sIgnificance Testing) might depend on the question at hand. For instance, in a study exploring the effects of a new treatment, the initial focus might be *existence*: how much are we certain that the effect is beneficial and not harmful? In a further step, however, the researcher might become interested in *significance*: is this effect large enough to be of any interest (for instance in relationship with cost-benefit trade-off)? Note that indices of significance and existence are conceptually independent. For example, an effect for which the whole posterior distribution is concentrated within the [0.0001, 0.0002] range would be considered as positive with a high certainty (and thus, *existing* in a that direction), but also not significant (*i.e.*, too small to be of any practical relevance). Acknowledging the distinction and complementary 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 EXIT is made straightforward through the *bayestestR* open-source package for R (**BAYESTESTR CITATION**).

Critically to the aim of that paper, the EXIT dual-perspective spontaneously stems out from the probabilistic nature of the Bayesian framework, which allows these two aspects of parameters assessment to coexist and yet be neatly delineated. Moreover, the distinction between *existence* and *significance* is also supported by the empirical data presented in this paper, in regards to the sensitivity to the indices to the amount of evidence (sample size). 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. On the other hand, ROPE-based indices and Bayes factors are effect of significance (related to the magnitude and the amount of evidence in favour of it). Thus, an effect will be comprehensively reported if, beyond its estimation (with a point estimate, such as the median, and an index of uncertainty, such as the 89% Credible Interval; McElreath, 2018), it presents one index of existenc,e and contextually test and discuss its significance and relevance in regards to theoretically justified characteristics.

The inherent subjectivity related to the assessment of significance is one of the practically limitation the ROPE-based indices (although being, conceptually, an asset, allowing for contextual nuance in the interpretaion), as they require an explicit definition of the non-significant range (the ROPE). Although default values were reported in the litterature (for instance, half of a “negligible” effect size reference value; Kruschke, 2014), it is critical for the reproducibility and transparency that the researcher’s choice is explicitly stated (and, if possible, justified). Beyond being arbitrary, this range also has hard bounds (for instance, contrary to a value of 0.0499, a value of 0.0501 would be considered as non-negligible). 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 outcome response (in is expressed in the unit of the outcome), these indices are sensitive to changes in the scale of the predictors. In other words, as the ROPE represents a fixed portion of the response’s scale, it is dependent on the scale of the predictor. For instance, in the case of a simple linear regression, for which the median of the coefficient of *x* on *y* is of 0.02 and falls within the ROPE (being not significant), simply multiplying *x* values by 100 would result in a coefficient with a meidan of 0.02 \* 100 = 20, which would fall outside of the rope (which range is fixed to *y*), that one inatentive or malicious researcher could misleadingly present as “significant” (note that indices of existence, such as the *pd*, would not be affected). Finally, the ROPE definition is also dependent on the model type, and selecting a consistent or homogenous range for all the families of models is not straightforward. This, in turn, can make comparisons between model types difficult, and an additional burden when interpretating 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 from the authors and the 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 bypass the usage 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.

Our results underline Bayes factor as the best predictor of the presence of an effect. Moreover, its easy interpretation in terms of odds in favour, or against, one or the other hypotehis makes it a compelling index for communication. Nevertheless, one of the main critique of Bayes factors, which is also underlined in our results, is its sensitivity to priors (which shows here through its sensitivity to model types, as priors for logistic and linear models are different). Moreover, while the BF against a ROPE appears as even better than the BF against a point-null, it also carries all the limitations related to the ROPE specification mentioned above. Thus, we recommend using Bayes factors (preferentially *vs* a ROPE) if the user has explicitly specified informative priors.

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 to assess size or importance and is not enable to give information in favour of the null. In other words, a high *pd* suggests the presence of an effect but a small *pd* does not give us any information about how much the null hypothesis is plausible, suggesting that this index can only be used to eventually “reject the null” (but not accepting it, which is consistent with the interpretation of the frequentist *p*-value). On the contrary, the BFs (and to some extent the percentage in ROPE, although being bounded to 0% and 100%) continue increasing or decreasing as the evidence becomes stronger (more datapoints), in both directions.

Much of these strengths also apply to the MAP-based *p*-value. Altough possibly showing some superiority in terms of sensitivity as compared to the *pd*, it also present 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 might impact the MAP-based *p*-value with unknown results. Additionaly, the *pd* 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 might appear as counter-intuitive to suggest the usage of its Bayesian empirical equivalent. The more subtle 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, in our opinion, the decay of the situation into the crisis. Interestingly, the proximity between the *pd* and the the *p*-value suggests that the latter is more an index of effect *existence* than *significance* (*i.e.*, “worth of interest”). Addressing this confusion, the Bayesian equivalent has an intuitive meaning and interpretation, making also obvious the fact that all thresholds and heuristics are arbitrary. Additionally, its mathematical and interpretative transparency of the *pd*, and its conceptualization as an index of effect existence, offers a 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.

# Reporting Guidelines

How these observations can be used to improve statistical good practices in psychological science? Importantly, before being able to draw a definitive conclusion about the qualities of these indices, further studies need to 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.

Nevertheless, based on the present comparison, we can start outlining the following guidelines. As the two aspects of the EXIT framework, *existence* and *significance*, are complementary, we suggest using at minimum one index of each category. Thus, in order to assess the existence and significance of effects within a regression model, we recommend to report, at minimum, the *pd* as an objective index of effect existence and the *BF (vs. ROPE)* as an index of significance. The former for its simplicity of interpretation, its robustness and its numeric proximity to the well-known frequentist *p*-value, and the latter for its ability to discriminate between presence and absence of effect (De Santis, 2007), and the information it provides related to relative evidence of the size of the effect.

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

“The effect of X has a probability of 92.14% [*pd*] of being negative (Median = -0.03, 89% CI [-0.05, 0.01]), moderately not-significant (BFROPE = 0.29) [*BF (vs. ROPE)*] and can be considered as very small (Std. Median = -0.09) [*standardized coefficient*].”

# Supplementary Materials

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>.

# Acknowledgments

This study was made possible by the development of the **bayestestR** package, itself part of the [*easystats*](https://github.com/easystats/easystats) ecosystem, an open-source and collaborative project created to facilitate the usage of R. Thus, we would like to thank the [council of masters](https://github.com/orgs/easystats/people) of easystats, all other padawan contributors, as well as the users.

# References

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.

Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20.

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. (2016). The earth is round (p<. 05). In *What if there were no significance tests?* (pp. 69–82). Routledge.

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.

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.

Gill, J. (2018). Comments from the new editor. *Political Analysis*, *26*(1), 1–2. <https://doi.org/10.1017/pan.2017.41>

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.

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.

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: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

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.

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

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.

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.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 2951–2959.

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.