Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks

Christopher Harms[1,2] & Daniël Lakens[2]

[1] University of Bonn, Germany

[2] Eindhoven University of Technology, Eindhoven, the Netherlands

Author Note

All data for the example study, along with analysis scripts for R (R Core Team, 2017), JASP (JASP Team, 2018), and jamovi (jamovi project, 2018), and the scripts used to create the figures in this article are provided in an OSF repository at https://osf.io/wptju/.

Correspondence concerning this article should be addressed to Christopher Harms, Kaiser-Karl-Ring 9, 53111 Bonn, Germany. E-mail: christopher.harms@uni-bonn.de

Abstract

The investigation of 'null effects' is important for cumulative knowledge generation in science. To draw informative conclusions from null-effects, researchers need to move beyond the incorrect interpretation of non-significant results in a null-hypothesis significance test as evidence for the absence of an effect. We explain how to statistically evaluate null-results using equivalence tests, Bayesian estimation, and Bayes factors. A worked example demonstrates how to apply these statistical tools, and interpret the results. Finally, we explain how no statistical approach can actually prove that the null-hypothesis is true, and briefly discuss the philosophical differences between statistical approaches to examine null-effects. The increasing availability of software and online tools to perform equivalence tests, Bayesian estimation, and Bayes factors make it timely and feasible to move beyond traditional null-hypothesis tests, and allow researchers to draw more informative conclusions about null-effects.

Keywords: equivalence testing, hypothesis, bayes factors, bayesian estimation

Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks

Most scientific research questions are stated in order to demonstrate the prediction that an effect or a difference exists. Does this drug work? Is there a difference between participants treated with antidepressants and patients going to psychotherapy? Common practice is to analyse the resulting studies using null hypothesis significance testinyg (NHST), for example by performing a t-test or a Mann-Whitney-U-test, and to conclude that there is a difference between a control and a treatment group when a difference of zero can be statistically rejected.

There are three scenarios in which the opposite research question, demonstrating the absence of an effect or difference, might be of interest:

1. Especially in clinical research, it might be important to know if a cheaper or shorter treatment works just as well as a more expensive or longer treatment. Studies designed to answer such questions investigate non-inferiority (e.g., people in one group do not score worse than people in another group) or the statistical equivalence of different treatments (e.g., people in one group score the same as people in another group).
2. We might design a study that has the goal to demonstrate the absence of an effect because we aim to falsify theoretical predictions about the presence of a difference.
3. Even when we do not explicitly aim to test the absence of a theoretically predicted effect, we should be prepared to observe a non-significant finding in any study we perform. Either when examining a novel hypothesis, or when performing a study that was designed to replicate a previous finding, we should be able to statistically evaluate null-results.

In all three cases statistical tools need to be applied that can provide an answer to the question whether we should believe, or act as if, a meaningful effect is absent. As Earp (2017) has laid out in his editorial, there is increasing attention to the fact that "null results" need to be published in order to have a coherent scientific body of results. Non-significant results

are to be expected, even when examining a true effect, and publication bias (not submitting or publishing non-significant resuls) will inflate effect size estimates in the literature. By using statistical approaches that allow researchers to evaluate null-results, researchers will be able to learn more from their data, and publication bias can perhaps be mitigated.

Researchers might want to know if a null-hypothesis is true, and therefore be interested in "proving the null". However, there are no statistical techniques that can unconditionally answer the question whether or not the null-hypothesis is true. As we will see below, statistical techniques that allow researchers to evaluate null results only allow conclusions about the null-hypothesis in relation to some specified alternative hypothesis. The null-hypothesis can not be statistically evaluated in complete isolation. Furthermore, it is impossible in empirical research to "prove" a prediction, since theories and predictions are inherently probabilistic in an inductive empirical science. Rare events will happen, and thus the absence of an effect is always concluded based on a certain probability of making an error, or given a certain level of certainty. The aim of the present article is to give an overview of statistical methods suited to investigate "null effects", and explain how to translate the statistical results from these methods into valid conclusions about the prediction that is tested. We provide a hypothetical example that is analyzed using four different methods, discuss how to interpretat the results (as well as possible misinterpretations), and briefly explain the inferential frameworks these different methods are based on.

Investigating "Null Effects"

It is common practice in empirical research to rely almost exclusively on null-hypothesis significance testing (NHST) to investigate the presence of an effect. Because a null-hypothesis test can only reject the null (i.e. commonly the hypothesis of "no effect"), it cannot be used to inform us about the absence of an effect. When we observe a non-significant effect (e.g., $p > \alpha$, where α is the level of significance chosen ahead of data-collection), all we can conclude is that, assuming the true effect size in the population is

zero, the observed data was not sufficiently different from zero to reject the null hypothesis without in the long run being wrong more often than a desired error rate. This does not rule out the possibility that the true effect size in the population differs from zero. In fact, one can argue that it is most likely that the true effect size is unequal to zero. Except for very rare questions, an effect size of exactly zero (as assumed in the null hypothesis) is implausible (see "crud factor", Meehl, 1990). For hypothesis testing, however, it is a useful model for comparison.

The experiment might have had relatively low power to detect the true effect size, or – equivalently – a high probability of making a Type 2 error (not rejecting the null-hypothesis when a true effect is present in the population). Null-hypothesis significance testing answers a specific question (i.e., can we reject the null-hypothesis?). When another question is of interest (i.e., can we conclude a meaningful effect is absent?), other statistical techniques should be used. Several statistical techniques have been developed to allow researchers to draw meaningful inferences about null-effects. Here, we will discuss equivalence testing, Bayesian estimation (i.e., the ROPE procedure) and Bayesian hypothesis testing (using Bayes factors). We will demonstrate these different approaches using a fictional dataset from an imaginary study.

Imagine you want to investigate whether mindfulness meditation has an effect on lower back pain (LBP), which is an increasingly common problem among desk-working adults. In a fictional study patients with lower back pain are recruited and randomly assigned to either an eight week mindfulness meditation class (treatment group) or an eight week acupuncture therapy (active control group). At the time of inclusion in the study and after the eight week study period self-reported lower back pain intensity is measured on a 100mm Visual Analogue Scale (VAS) (Abdel Shaheed, Maher, Williams, Day, & McLachlan, 2016; Machado et al., 2015). The dependent variable to be analyzed is the difference between the VAS scores at the end and start of the study. The mean change over the eight week period between the

treatment group and the acupuncture group is examined using a two-sample t-test.[1]

The sample size of the study is determined based on an *a priori* power analysis. Based on a discussion with experts in the field, the smallest effect size of the treatment that is still deemed worthwhile is Cohen's $d = 0.30$, and the study is designed to have a high probability of observing a statistically significant effect, if there is a true effect at least as large as this smallest effect size of interest. Assuming it is relatively easy to get people to enroll in the study and the design can be designed to be informative, the alpha level is set to 0.01 and the desired power for the smallest effect size of interest is set at 90%.[2] This means that if there is a true effect of $d = 0.30$ or larger, we have at least 90% chance of observing a significant effect (in the long run). Based on the desired error rates, the power analysis indicates 332 patients per group should enroll in the study.

The fictional measurements collected from 664 participants are visualised in Figure 1. The mean change in self-reported lower back pain intensity on the 100mm VAS over the eight week period (and standard deviations) are -2.30 (14.77) in the Meditation group and -0.39 (15.13) in the control group.

Traditional Significance Test

A common first question in experiments where participants are randomly assigned to two conditions is to examine whether we can statistically reject a difference between the groups that is exactly zero. This null hypothesis can be examined by performing a t-test with a significance level of $\alpha = 0.01$. The two-sample Welch's t-test (which does not assume equal variances) yields $t(661.63) = 1.64$, $p = .101$. The p-value is not statistically significant, which means the observed difference in the data is not extreme enough to reject the

[1] The study design and analysis plan used herein is simplified for illustrative purposes. Practitioners might in reality consider a multilevel analysis to better account for different sources of variation (e.g. Hayes, 2006). The general recommendations in this paper apply also in more complex models.

[2] Ideally, the alpha level is set based on a cost-benefit analysis of Type 1 and Type 2 errors, see Lakens et al. (2018a).

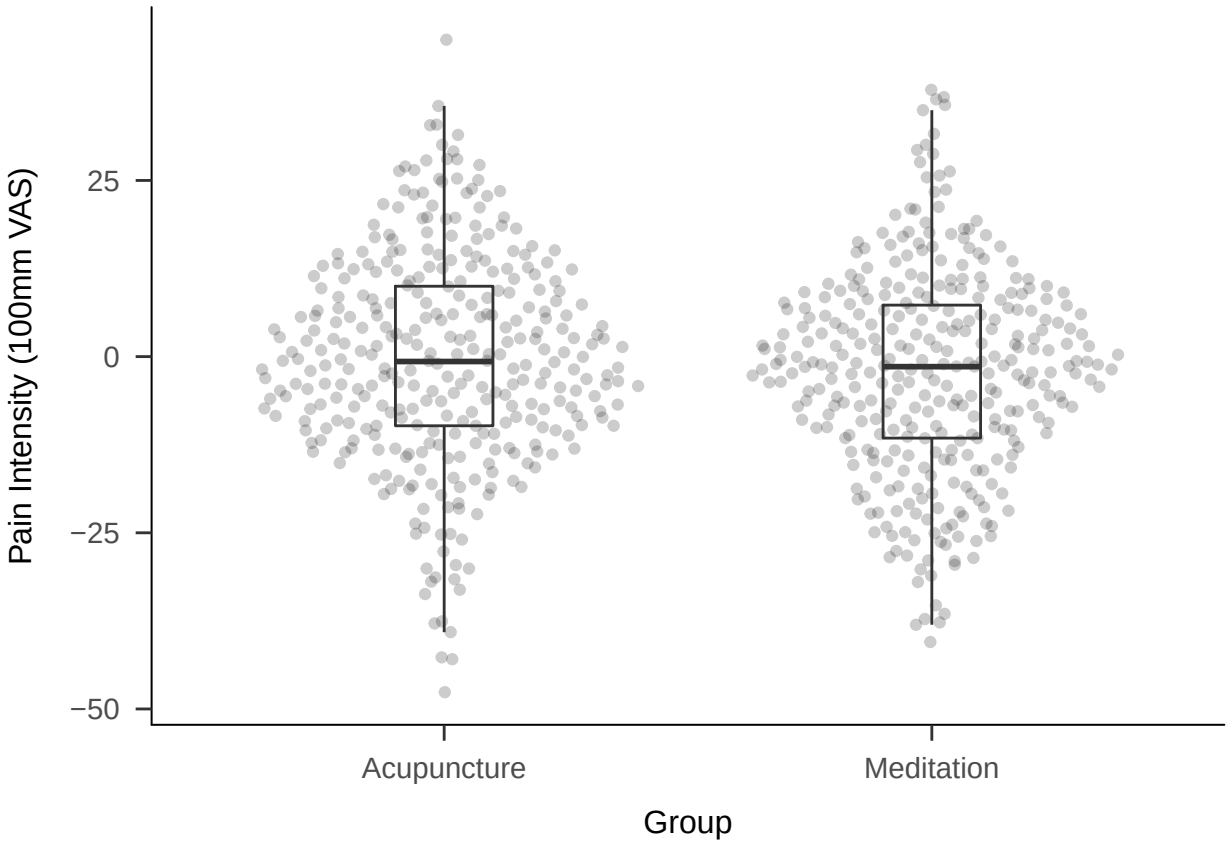

Figure 1. Plot for the data of the imaginary study. Each dot represents a single case. Box plot shows median and 25% and 75% quartiles. Y-axis is dependent variable, i.e. Pain Intensity after either 8 weeks of meditation class or after 8 weeks of acupuncture therapy.

hypothesis that the changes in pain scores in both groups are the same. A non-significant test result does *not* mean that the null hypothesis is true. Non-significant results simply indicate that the data are not surprising if we assume there were no differences between the groups. This might be because there is no true difference between the two groups, in which case a non-significant effect is expected with a frequency of 0.99. But it is also possible that there is a difference, but due to chance, it was not observed, which should happen 10% of the time if the true effect size is $d = 0.30$ (and more often if the difference between groups in the population is smaller than $d = 0.30$).

A null hypothesis significance test cannot distinguish between the conclusion that an observed difference is too small to be considered meaningful, or an inconclusive result (i.e.,

the effect is not statistically different from zero, but also not statistically smaller than any effect you care about). This often leads researchers to believe non-significant results are not informative. While a non-significant result in a null-hypothesis significance test *per se* does not allow us to decide between the absence of a meaningful effect, or an inconclusive result due to low power, the data might be informative when analyzed with statistical tests that do allow researchers to draw more useful conclusions about null-effects.

In the past researchers were advised to interpret non-significant results by performing a sensitivity analysis, and report an effect size the study had high power to detect. For example, if a study had 90% power to detect an effect of $d = 0.30$, researchers might conclude that if there is an effect, it would most likely be smaller than $d \geq 0.30$. This is referred to as the "power approach" (Meyners, 2012; Schuirmann, 1987). Based on the absence of a significant effect, researchers would conclude that it is unlikely that a true effect as large or larger than a specific size is present. However, this "power approach" is superseded by the development of equivalence tests (Meyners, 2012), and the "power approach" is no longer recommended.

Equivalence Tests

There is no statistical procedure that can confirm that the difference between two groups is exactly zero (beyond sampling the entire population, and finding that the observed difference or effect is exactly 0). However, it is possible to test whether an effect is close enough to zero to reject the presence of a meaningful difference. In this approach, researchers need to specify the difference that is considered too small to be meaningful, or the smallest effect size of interest (SESOI). The SESOI is in clinical domains also referred to as the "minimal clinically important difference" (MCID). A statistical test (very similar to the traditional t-test) is performed that examines whether we can statistically reject the presence of a difference as extreme, or more extreme, as the smallest difference we care about. If we can reject the presence of a difference (with a desired alpha level) we can act as

if the difference is *practically equivalent* to zero. This procedure is known as *equivalence testing* (Rogers, Howard, & Vessey, 1993).

For clinical scenarios in which pain intensity is measured using a 100mm VAS in patients with lower back pain, a difference of 9mm is considered to be a minimal clinically important difference, which is the difference of 9mm is when patients indicate that they subjectively feel "slightly better" instead of "equal" (Wandel et al., 2010). Note that this is only one approach to determine a smallest effect size of interest, and other justifications for a smallest effect size of interest are possible (Lakens, Scheel, & Isager, 2018b). Ideally, the SESOI should be informed by theory and previous research (such as meta-analyses or systematic reviews). The SESOI needs to be determined before collecting the data (similar to decisions about the sample size, the alpha level, and the desired statistical power). An informative study should be designed to have sufficient power both (i) to detect an effect that exceeds the SESOI and (ii) to demonstrate equivalence to zero or another specific value (thus rejecting the smallest effect size of interest).

One way to test for equivalence is to perform the Two One-Sided Tests (TOST) procedure. A lower (Δ_L) and upper (Δ_U) equivalence bound is specified (e.g., a difference of -9mm or 9mm on a 100mm VAS). A first one-sided test is performed to examine whether we can reject effects *smaller* than $\Delta_L = -9$mm, and a second one-sided test is performed to test whether we can reject effect *larger* than $\Delta_U = +9$mm. If both one-sided tests are significant, we can reject the presence of a difference of ± 9mm or larger, and conclude the observed effect is statistically equivalent, given the equivalence bounds that were chosen.

Lakens (2017) created an R-package (TOSTER) and a spreadsheet to perform equivalence tests for t-tests, correlations, proportions, and meta-analyses. Performing an equivalence test (again using Welch's t-test) on our fictional data, with an α-level of 0.01, yields a significant result ($t_1(661.63) = 6.11$, $p < .001$; $t_2(661.63) = -9.40$, $p < .001$). The result is vizualized in Figure 2, where the 98% confidence interval is plotted and compared to the equivalence bounds of -9mm and $+9$mm. The width of the confidence interval is $1 - 2\alpha$

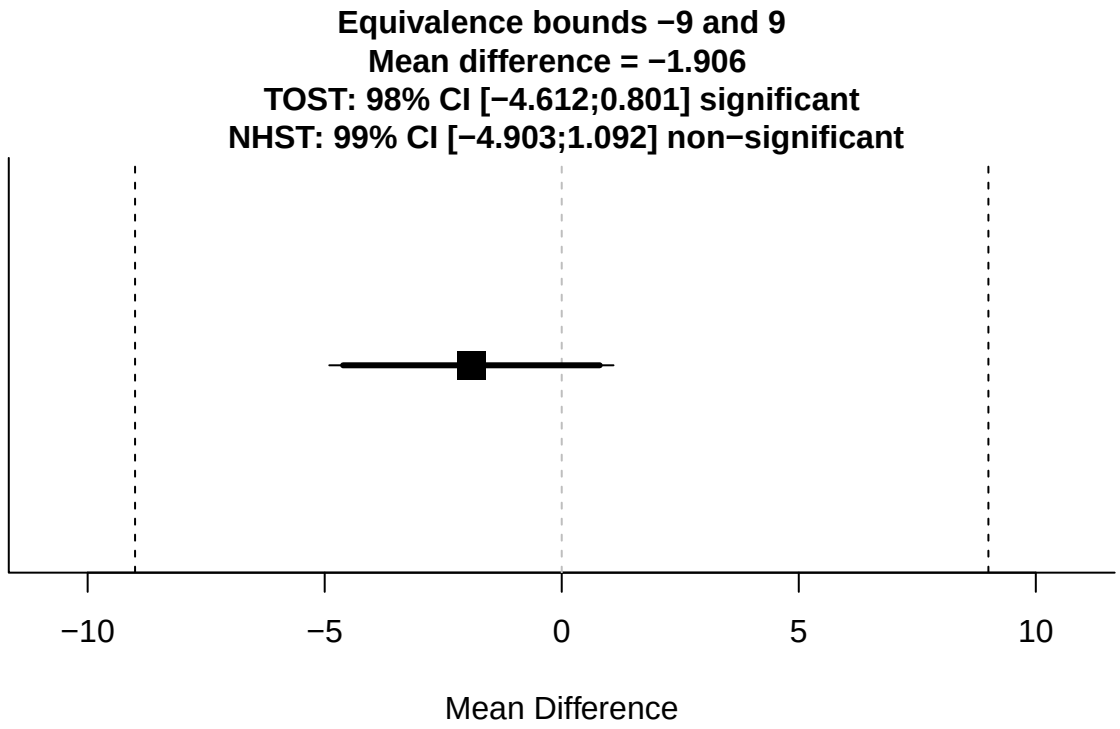

Figure 2. Visual representation of the equivalence test. Plotted is the confidence interval for the mean difference between the two groups. Bold interval is the 98% interval used for the two one-sided test along with the thin 99% interval used for the traditinal significance test against the null hypothesis of no difference. The equivalence test is significant, i.e. the 98% confidence interval lies well within the equivalence bounds of −9mm and +9mm and we can conclude that there is significant equivalence between the two groups.

since two one-sided tests are performed, both of which need to be significant to conclude equivalence, (Rogers et al., 1993). Using a Neyman-Pearson approach to statistical inferences, in which the goal is to make dichotomous decisions while controlling error rates at a desired level, we can *act* as if the difference between the two groups is smaller than the minimal clinically important difference of ±9mm, without being wrong too often in the long run.

The present example represents the case of a non-significant result that is equivalent to zero. It should be noted, that the equivalence testing approach also allows for significant *and* equivalent outcomes: If sample size increases and the estimated effect size is significantly different from zero, while also significantly smaller than an equivalence bound, both the

traditional significance test as well as the equivalence test will be significant – indicating a significant but equivalent result.

Accessible introductions and explanations to equivalence testing are available (e.g. Lakens, 2017; Lakens et al., 2018b; Meyners, 2012) and can be performed in R, using a spreadsheet (Lakens, 2017), or using the free software jamovi. We provide scripts for R (R Core Team, 2017) and jamovi (jamovi project, 2018) to reproduce the analyses and results in this paper as supplemental material.

Bayesian estimation

Where Frequentist statistics, which underly null-hypothesis significance tests and equivalence tests, allow controlling error rates in the long run, without ever being able to say whether the conclusion made for any single study is one of these errors, Bayesian statistics allows researchers to make statements about the probability of a hypothesis, given the observed data. This is due to a different definition of the term "probability". The debate about which definition of probability is "correct" has lead to a debate among statisticians and philosophers of science that has been going on for many decades. Luckily, researchers don't need to choose a side (unless they want to), because both approaches can be used side-by-side when analysing data. Excellent introductions to Bayesian statistics from an applied perspective on statistics can be found in McElreath (2016) or Kruschke (2014).

Bayesian statistics is best understood in the context of statistical modelling. A statistical model is a mathematical description of the probability of data. In Bayesian statistics a model consists of three different parts. The first part is called a *prior distribution*: For each parameter we choose a probability distribution that describes expectations about possible parameter values. This prior can be understood as our "belief" before seeing the data (hence the *prior*). This terminology already highlights the distinction between the frequentist and the Bayesian understanding of probability: While frequentist consider "probability" as a statement about long-term frequencies of events, Bayesians think of

"probability" as a "degree of belief". This subjective interpretation is easily explained – and very intuitive to some – but not without criticism. Even among Bayesians there is disagreement about the subjective nature of the prior. Gelman (2011) provides one accessible commentary on this debate.

As the second part of a Bayesian model, we take the observed data into account through a *likelihood function*, and calculate a posterior distribution through the use of Bayes' theorem. In mathematical notation this is

$$P(\theta|Data) = \frac{P(Data|\theta) \cdot \pi(\theta)}{P(Data)}$$

where $\pi(\theta)$ is the prior distribution for our parameter θ, and $P(Data|\theta)$ is the likelihood function of the model. $P(\theta|Data)$ is the *posterior distribution* of the parameter after seeing the data (i.e. the conditional probability of the parameter values given the observed data). The posterior distribution is thus – analogous to the prior distribution – our belief about different parameter values for θ *after* having seen the data. When moving from a prior to a posterior distribution credibility is realocated from the prior distribution to a posterior distribution that represents credibility informed by both the prior information and the data. The term $P(Data)$ in the denominator is a normalizing constant in order for the posterior $P(\theta|Data)$ to be a proper probability distribution. We will later refer to it in the section about Bayes factors as the *marginal likelihood* of the model (since it is the likelihood marginalized over all parameter values), also called *model evidence*.

Kruschke (2013) introduced a pre-defined Bayesian model that can be used to draw inferences about the estimated differences between two independent groups. This procedure provides researchers with a simple and easy-to-use test to evaluate the data in a Bayesian estimation framework. As such, the approach provides samples from the posterior distribution, which can be used to make inferences about the data. Importantly, the whole posterior distribution provides the statistical inference, even if only summaries are presented (Kruschke & Liddell, 2017). Even when summaries of the posterior distribution are given, such as means, standard deviations, or credibility intervals, the entire Bayesian posterior

distribution is available, and provides more information than frequentist statistics. One way to summarise the posterior distribution is to provide intervals of parameter values that are considered to be most credible. In Bayesian statistics Highest Density Intervals (HDI) are commonly used. For example, a 89% Highest Density Interval contains the values which, based on the statistical model used (including the prior distribution), are considered the 89% most credible. These can be calculated using the "BEST" R-package (Kruschke & Meredith, 2017) or a web-app (Bååth, 2012).

In our imaginary study where we compare an 8-week meditation class to patients receiving acupuncture we find a 95% Highest Density Interval (HDI) of $[-4.24; 0.32]$ for the difference in pain intensity between the two conditions. This means that the 95% most credible values for the difference in means, given our model, which incorporates both the prior information and the observed data, lie between -4.24mm and 0.32mm. Figure 3 visualizes this result.

Differences between the confidence interval reported above and the Bayesian HDI are to be expected. First, since we are using a 95% interval for the HDI and a 98% confidence interval in the equivalence test, the HDI is expected to be narrower than the confidence interval reported above. More generally, the prior affects the width and location of the HDI in Bayesian estimation. If the priors that are used for the model are not uniform, differences between an HDI and a confidence interval are to be expected. With sufficient data, the data will outweigh the prior, but with smaller amounts of data, it can be advisable to explore the impact of different priors on the inference. In the BEST model, the priors are not uniform but chosen to have minimal impact on the inferences, so even if only few data points are available, the prior should not have too much influence on the results.

The posterior distribution can be used to answer several other questions as well. Besides the HDI, we can find the most credible value for the difference between the two groups, which would be the posterior mode, or *Maximum A Posteriori estimate* (MAP), which is -1.81 (and differs slighty from the frequentist estimate of the difference due to the

prior). When one aims to make a dichotomous decision about parameter values based on the posterior distribution, Kruschke and Liddell (2017) propose to define a *Region of Practical Equivalence* (ROPE) which is identical to setting equivalence bounds based on a smallest effect size of interest as laid out above. The ROPE procedure uses the following decision rule (see Kruschke, in press):

> If the 95% HDI of the [parameter's posterior distribution] falls completely outside the ROPE than reject the null value, because the 95% most credible values of the parameter are all not practically equivalent to the null value. If the 95% HDI of the [parameter's posterior distribution] falls completely inside the ROPE then "accept" the null value for practical purposes, because the 95% most credible values of the parameter are practically equivalent to the null value. Otherwise remain undecided.

By comparing the 95% HDI with the *region of practical equivalence* (ROPE) between $\Delta_L = -9$mm and $\Delta_U = +9$mm, based on the same equivalence bounds as before, researchers can conclude equivalence when the HDI lies within the region of practical equivalence (or between the equivalence bounds). Because the 95% HDI ($[-4.24; 0.32]$) lies well within those bounds (as can be seen in Figure 3), we declare a difference of exactly zero to be accepted for practical purposes based on the decision rule above. We do not, however, accept or reject any other specific value within the ROPE. In the vocabulary of Bayesian statistics, using a decision rule on a posterior distribution of a single model does not constitute "hypothesis testing". The term "Bayesian hypothesis testing" refers strictly to the use of Bayes factors for model selection, which will be considered in the section of the paper.

The Bayesian ROPE procedure is quite similar to equivalence tests, but there are several important dinstinctions. In the Bayesian approach we can make statements about which values we believe are most credible, based on the data and the model, while in frequentist statistics we are make dichotomous decisions based on long-run error rates. Frequentist statistics is concerned with frequencies of events in the long run. Hypothesis

tests as outlined above aim to control rates of errors at pre-specified levels. The width of a confidence interval for example is directly related to the chosen α level. In the Bayesain approach, on the other hand, no statements about rates of decision errors can be made without additional assumptions and analyses. Kruschke and Liddell (2017) use the 95% interval as a convention related to the 5% significance level, but the width of the HDI is arbitrary, should only be seen as a useful summary of the complete posterior distribution, and is *not* related to the 5% Type 1 error rate of the confidence interval.[3]

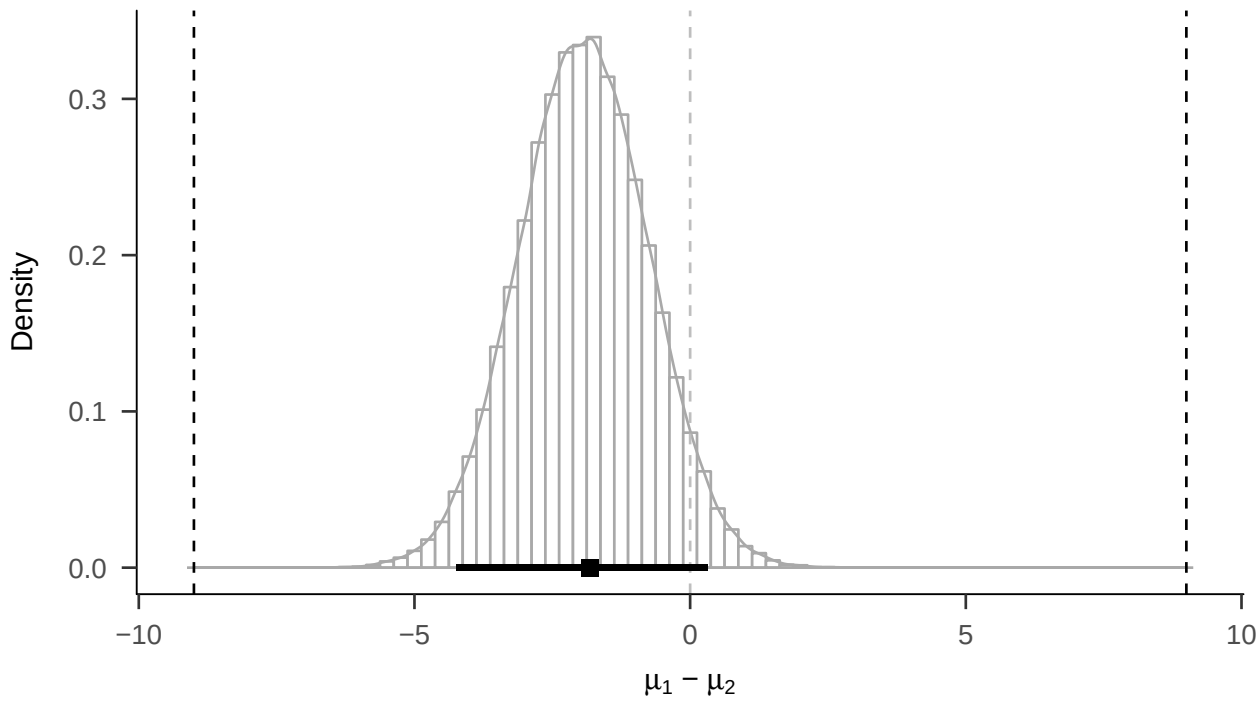

Figure 3. Histogram with superimposed density estimate of samples from posterior distribution for the Bayesian t-test model (Kruschke, 2013). Thick bar is the 95% Highest Density Interval, indicating the 95% most credible values for the mean difference between the two groups. The square in the interval is the *Maximum A Posteriori* estimate, i.e. the most credible value from the posterior distribution.

[3]Note, however, that in some practical cases frequentist confidence intervals and Bayesian credibility intervals yield the same range of values (Albers, Kiers, & Ravenzwaaij, 2018).

Bayesian hypothesis testing with Bayes factors

The ROPE procedure uses Bayesian statistics to estimate the parameter values that are most credible and then uses a decision rule to accept or reject specific values. Bayesian statistics can also be used to directly test two competing models. Hypothesis testing can be considered as a special case of model selection, where two specific hypotheses are expressed in terms of competing models. One way to perform this type of model selection in Bayesian statistics (or Bayesian hypothesis testing) is to compare the *marginal likelihoods* of two models M_0, the null model, and M_1, the alternative model, and quantify the relative *model evidence* in terms of a ratio:

$$\mathrm{BF}_{01} = \frac{P(Data|M_0)}{P(Data|M_1)}$$

This ratio is called a *Bayes factor* and allows statements about relative model evidence. A Bayes factor of $\mathrm{BF}_{01} = 4.2$ can be interpreted as "the data provide 4.2 times more evidence for M_0 than for M_1."[4] Bayes factors indicate by what amount the relative belief in the models should shift according to rational Bayesian belief updating:

$$\underbrace{\frac{P(M_0|Data)}{P(M_1|Data)}}_{\text{Posterior Odds}} = \underbrace{\frac{\pi(M_0)}{\pi(M_1)}}_{\text{Prior Odds}} \times \underbrace{\frac{P(Data|M_0)}{P(Data|M_1)}}_{\text{Bayes factor}}$$

The most common approaches to calculating Bayes factors model the null-hypothesis as a point, with an alternative model that distributes the probability of the true value across a range of possible value. That is generally similar to frequentist hypothesis testing, where the null hypothesis is commonly also a point hypothesis of exactly zero. For Bayes factors that closely resemble traditional statistical tests, the models are distinguished by different prior distributions for a parameter (usually a test statistic).

Defining a reasonable alternative model is an important part of calculating a Bayes

[4] The subscript in BF_{01} specifies the relative evidence for the null compared to the alternative, but a Bayes Factor can also be expressed as the relative evidence for the alternative compared to the null, or $\mathrm{BF}_{10} = 1/4.2 = 0.24$.

factor. There are different ways in which the alternative model can be specified. One way is to use researchers' beliefs or expectations of theoretical predictions. Another way would be to use data observed in previous studies to inform the alternative model (Harms, 2016; Verhagen & Wagenmakers, 2014).

Figure 5(D) illustrates the two models compared when calculating a Bayes factor. In the figure M_0 is represented by a point-null hypothesis and M_1 is represented by a distribution that assumes small effect sizes are more likely than large effect sizes, but which is not very restrictive and assigns probabilities to a wide range of possible values.

A common criticism on Bayes factors is that they are much more sensitive to the specification of the prior than Bayesian model estimation. In a Bayesian estimation framework (such as the ROPE procedure) the data quickly overwhelm the prior, so the prior has a very limited effect on the final statistical inference. For Bayes factors, on the other hand, priors have much more weight and thus need to justified carefully before looking at the data. We would caution against the use of "default" priors when calculating Bayes factors (see Dienes, 2014), which are a compromise between general expectations about effect sizes and useful mathematical properties (e.g. Rouder, Speckman, Sun, Morey, & Iverson, 2009), but should only be chosen if they actually reflect a useful alternative model given the research question. Moreover, Bayes factors – very much like p-values – do not convey information about the magnitude of an effect or the uncertainty in its estimation. See Kruschke and Liddell (2018) for further criticism on Bayes factors.

Bayes factors can be used to examine null effects by quantifying the relative evidence in the data for a null-model compared to an alternative model. In our calculation of a Bayes factor for our hypothetical data we wanted the prior for the alternative model to represent our expectation about the presence of a true effect. If our 8-week meditation class reduces pain intensity in a 100mm VAS scale compared to the active control condition, we expect it to be similar in size to other non-pharmaceutical interventions. Hoffman, Papas, Chatkoff, and Kerns (2007) performed a meta-analysis of different psychological interventions on pain

intensity in patients with chronic lower back pain, and provided an estimated meta-analytical effect size of $d = 0.62$ (95% CI: $[0.25; 0.98]$) when comparing the effect of cognitive-behavioral therapy (CBT) against a waiting list condition. Therefore, we calculate a Bayes factor based on the expectation that a mindfulness meditation intervention might have similar effect size.

We specify an alternative model with a normal prior distribution centered on 0.62 with a standard deviation of 0.37 (calculated from the confidence interval): $M_1 : \ \delta \sim \mathcal{N}(0.62, 0.37)$. The M_1 model is compared against the null model M_0 with a prior that has its point mass at 0 (i.e. a point null hypothesis).

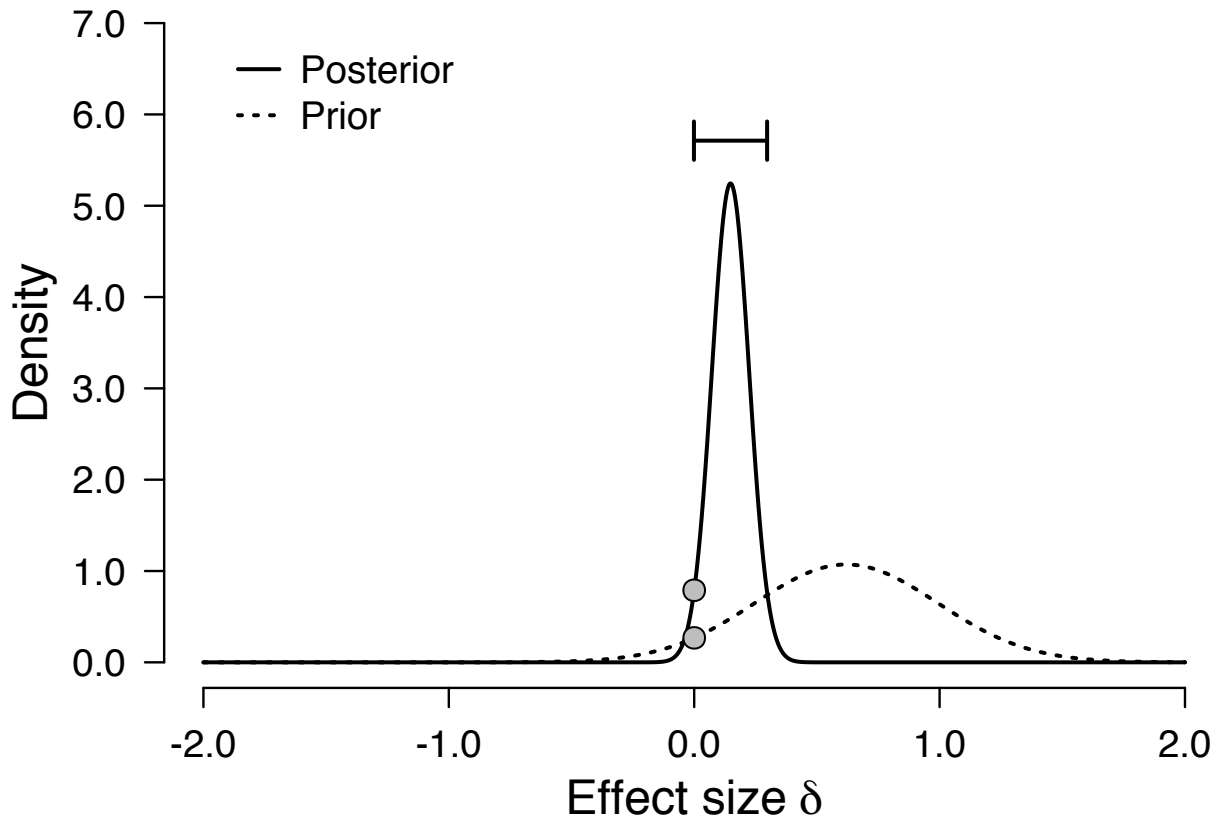

Figure 4. Visual representation of the Bayes factor as Savage-Dickey ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010): The Bayes factor can be understood as the ratio between the posterior and the prior at $\delta = 0$.

A Bayes factor for a t-test yields $\mathrm{BF}_{01} = 2.95$ (calculuated using R, or JASP following the formula given by Gronau, Ly, & Wagenmakers, 2017). We can thus conclude that the

data is 2.95 times more in favour of the null model compared to the informed alternative model that we specified. The Bayes factor can be represented visually as in Figure 4: It shows the ratio between the prior and posterior distribution at $\delta = 0$, the point of interest for the null hypothesis. This ratio is called the *Savage-Dickey ratio* (Wagenmakers et al., 2010). Although Bayes Factors can be interpreted as continuous measures of model evidence, thresholds for interpreting Bayes factors have been proposed by Jeffreys (1961) which might be useful for researchers who begin to report and interpret Bayes factors. A Bayes factor of 1 indicates the data are equally likely under both models. Bayes factors between 1 and 3 constitute mere "anecdotal" evidence, which is considered "worth not more than a bare mentioning" (Jeffreys, 1961, Appendix B). Thus, although the data support the null model over the alternative model, there is no good reason to conclude in favor of either model. Stronger model evidence would be desirable, which means more data must be collected (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).

The difference between the result of Bayes factor analysis, the equivalence test, and the ROPE procedure reported earlier has several reasons. Most importantly, the questions that were asked differed across the tests. The equivalence test sought to reject an effect specified by and upper and lower equivalence bounds of ±9mm (see Figure 5(B)), and the ROPE procedure examined wether the 95% HDI fell within the region of practical equivalence (Figure 5(C)). The Bayes factor investigated whether the data was more in line with a null model or amodel specified based on expectations derived from previous studies. Researchers need to be aware of the precise question they want to ask from the data and the method they use to do answer their question. In order to draw informative inferences from the data, it is crucial that a statistical test is selected in which alternative hypotheses are defined that answer a question of interest.

The Bayes factor tells us how much our belief in the null model versus the alternative model should change. It does not, however, directly tell us how likely the null hypothesis is because it is a relative measure. As can be seen in the equation above, to calculate the

posterior odds of the two competing hypotheses, a researcher needs to combine the Bayes factor with prior probabilities for the two hypotheses. There is rarely an objective answer to this question, and researchers are free to hold different beliefs. If we feel that the two models are equally likely *a priori*, i.e. the prior odds are 1:1, then the Bayes factor would be equal to the posterior odds. If, on the other hand, we feel that the null hypothesis is four times more likely than the alternative hypothesis (before seeing any data from the study) and the Bayes factor is $\mathrm{BF}_{01} = 2.95$, we should believe that the null model is about 11.78 (4 times 2.95, with a small difference due to rounding) more likely than the alternative after seeing the data. Since different researchers can have different beliefs about the prior odds of two hypotheses, Bayes factors are commonly reported, without a reference to prior or posterior odds, and the reader is assumed to update their own priors. The Bayes factor contains the neccessary information to make an inference, but cannot be used on its own to reach a conclusion. Prior odds are a necessary part of the inferential method when using Bayes factors.

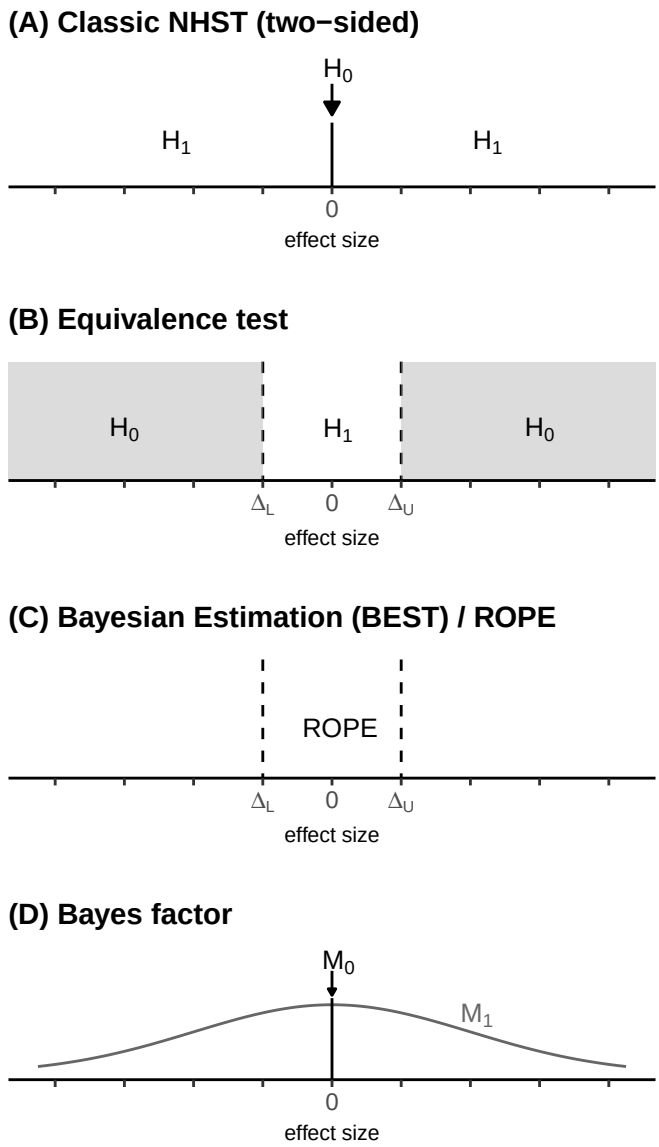

Figure 5. Illustration of the different hypotheses under investigation (adapted from Lakens et al., 2018b). **(A)** The classic two-sided significance testing aims to reject a point null hypothesis (here an effect size of exactly zero). **(B)** In equivalence test, the H_0 of no equivalence is tested (grey region), so the white area is the rejection region. **(C)** For the Bayesian estimation approach, the 95% highest density interval of the posterior is compared against the Region of Practical Equivalence (ROPE) between Δ_L and Δ_U. **(D)** For the Bayes factor, two models are compared that differ in their prior distributions: The M_0 prior is a point mass of 1 at an effect size of 0, the alternative model M_1 is here plotted as a Normal distribution as an example.

Discussion

There are good reasons to want to test whether meaningful effect sizes or theoretically predicted differences are absent in data that have been collected to examine a hypothesis. In recent years, statistical techniques such as equivalence testing, Bayesian estimation, and Bayesian hypothesis tests have become more widely available through open source software tools such as R (R Core Team, 2017), jamovi (jamovi project, 2018), and JASP (JASP Team, 2018), and accessible introductions with detailed examples (Kruschke, 2014; Lakens et al., 2018b; McElreath, 2016). These statistical tools allow researchers to move beyond merely testing whether the null hypothesis can be rejected in null-hypothesis significance tests. These complementary statistical approaches invite researchers to more carefully consider and specify which effect sizes they predict when there is a true effect. In planning a study and performing statistical inferences, researchers should more explicitly consider the fact that the null hypothesis could be true. A statistical evaluation of the observed data should allow for informative conclusions about this possibility. This implies that an informative study should be designed to allow for inferences about both presence *and absence* of a meaningful effect. Doing so would also, hopefully, prevent the common mistake to interpret a p-value larger than the alpha level (e.g., $p > .05$) as the absence of an effect.

In the context of clinical trials, the repeated use of equivalence and non-inferiority tests can have negative effects on the conclusions derived from such research. That is, if sampling and measurement error are large and the equivalence region is rather wide, repeated studies comparing non-inferiority of different treatments or doses might favor treatments which are ineffective or even harmful (Aberegg, Hersh, & Samore, 2018; Everson-Stewart & Emerson, 2010). A phenomenon that has been termed "bio-creep". The prevalence of bio-creep is a matter of ongoing research; Beryl and Vach (2011) come to the conclusion, that it is not a major cause of concern in practice (at least on average). Awareness of the issue is nevertheless important and should even more underline the need for conducting direct replication studies, meta-analyses and the use of adequate statistical tools.

Possible Misconceptions

Probability is not intuitive, and every statistical technique runs the risk of being misinterpreted. The techniques discussed in this article have great potential to improve statistical inferences, but it is important to prevent misinterpretations. When performing a null-hypothesis significance test, a non-significant result can not be used to conclude a meaningful effect is absent. To conclude this, one has to specify and test against whichever effect one defines to be "meaningful". An equivalence test can be used to statistically *reject* effects as large or larger than the smallest effect size of interest, with a long-term error rate. It can *not* be used to conclude the effect is exactly 0, or to reject *any* effect. Even when an equivalence test is statistically significant, it is possible that there is an effect - it is just smaller than what was deemed meaningful when the study was designed. For this reason, conclusions based on equivalence tests must always specify the equivalence bounds that are used, and it is recommended to combine equivalence tests with null-hypothesis significance tests (which can also help to identify effects that are significant *and* equivalent, or practically insignificant differences). Thus, a statement such as "the difference was statistically equivalent to zero" is wrong, but a correct interpretation is "we could reject effect sizes more extreme than the equivalence bounds of -0.4 and 0.4".

When calculating the posterior distribution in Bayesian statistics, a prior is combined with the observed data. Any statements about the posterior distribution are not just based on the data, but also conditional on the model. The model includes the prior distributions which can be chosen more freely. They may represent a researchers beliefs prior to the data, but can also be used to regularise estimates or incorporate information from previous studies. It is thus important to explicitly state the model setup and provide a justification for the choice of prior distributions. As a consequence, the 95% HDI of the posterior distribution does *not* include parameter values that are likely to be *true*. No statistical procedure, including confidence intervals, will be able to provide this. Instead, the 95% HDI contains the parameter values which are most *credible*, given the data and the model (including the

priors). Finally, when calculating Bayes factors, it is important to realize that they provide relative evidence for two specificied models. A Bayes factor can indicate strong support for a null model relative to an alternative model, but both models can be wrong (and are perhaps even likely to be wrong, as "all models are wrong, but some are useful", Box & Draper, 1987, p. 424). The Bayes factor gives a relative indication which of the two models are more in line with the data (or "less wrong" according to Box).

Differences between Inferential Frameworks

All statistical methods give rise to probabilistic inferences. Rare events happen, and unlikely outcomes can be observed. Probabilistic methods can never be used to know with certainty that an effect is present or absent. Thus, none of the statistical techniques presented in this paper are capable of *proving* the null. After analyzing their data, researchers might be tempted to conclude "there was no effect", but none of the statistical approaches discussed here allow for such a conclusion. It is important to understand the questions that the different statistical techniques described in this article provide an answer to.

Equivalence tests are used to make dichotomous conclusions to guide behavior, while controlling error rates in the long run. The goal of such a test is to reject the presence of effects large enough to matter, without being wrong too often. Any single study might lead to an incorrect conclusion, but theories that are correct should make predictions that are confirmed with expected error rates in lines of research. Although single studies are never sufficient to draw strong conclusions in science, this idea is especially central in frequentist statistics.

Bayesian statistics focuses more strongly on quantifying beliefs or making statements about credibility. In the case of Bayesian estimation, the focus lies on allocating credibility to parameter values (such as effect sizes or differences between groups), which can result in statements about degrees of belief. In the case of Bayes factors, the focus lies on quantifying

the rational change in belief in a null-model or an alternative model, which is also termed *statistical evidence* (Morey, Romeijn, & Rouder, 2016). Although there are many different flavors of Bayesian statistics, a strength of these approaches lies in drawing conclusions that incorporate pre-existing information in statistical inferences. Whether quantified beliefs or any other statistical inference corresponds with reality depends on how accurate model assumptions are. This is relevant for Bayesian models and the chosen prior distributions as well as for model assumptions in frequentist statistics.

In Bayesian estimation the prior can be used to regularise parameter estimates. Especially in small samples and in more complex models, this avoids overfitting the data and can lead to better estimates for out-of-sample inferences and predictions (Gelman et al., 2013, Chapter 14.6). While the perceived subjectivity of priors has been critised by frequentist analysts, with increasing amounts of data, the prior has less influence on the statistical inference. Finally, the Bayesian approach to statistical modelling is very versatile and can be used even in very complex models. Bayesian hierarchical or multilevel models are particularly useful in clinical research, for example, when using clustered samples or repeated measurements (Gelman et al., 2013, Chapter 5; Goldstein, Browne, & Rasbash, 2002; Turner, Omar, & Thompson, 2001).

Conclusion

Null hypothesis significance testing has been critised because it is often misused and misunderstood (e.g. Wasserstein & Lazar, 2016). Null-hypothesis tests limit researchers to asking the question whether the null-hypothesis can be rejected. By adding statistical techniques such as equivalence testing, Bayesian estimation, and Bayes factors to ones repertoire, researchers can substantially improve the inference they can draw from null-effects. Being able to demonstrate the absence of effects is important in all major approaches to philosophy of science (Fidler, Thorn, Barnett, Kambouris, & Kruger, 2018). When researchers only publish scientific findings that statistically reject null effects, the scientific

literature is biased, which hinders the accumulation of scientific knowledge (Kühberger, Fritz, & Scherndl, 2014; Locascio, 2017). By using statistical approaches that can provide informative conclusions about null effects, researchers might not be able to "prove the null", but they can substantially improve their statistical inferences about null-effects.

Conflicts of Interest

None.

Author Contributions

The general idea for the manuscript was jointly developed by CH and DL. CH drafted the introduction, the example, and the discussion section and created the analysis scripts. DL provided critical comments and extended the discussion section. CH and DL collaboratively revised the manuscript for final submission.

References

Abdel Shaheed, C., Maher, C. G., Williams, K. A., Day, R., & McLachlan, A. J. (2016). Efficacy, Tolerability, and Dose-Dependent Effects of Opioid Analgesics for Low Back Pain. *JAMA Internal Medicine*, *176*(7), 958. doi:10.1001/jamainternmed.2016.1251

Aberegg, S. K., Hersh, A. M., & Samore, M. H. (2018). Do non-inferiority trials of reduced intensity therapies show reduced effects? A descriptive analysis. *BMJ Open*, *8*(3), e019494. doi:10.1136/bmjopen-2017-019494

Albers, C. J., Kiers, H. A. L., & Ravenzwaaij, D. van. (2018). *Credible Confidence: A pragmatic view on the frequentist vs Bayesian debate*. doi:10.17605/OSF.IO/WE9H6

Bååth, R. (2012). Bayesian Estimation Supersedes the t-test (BEST) - online. Website. Retrieved from http://www.sumsar.net/best_online/

Beryl, P., & Vach, W. (2011). Is there a danger of "biocreep" with non-inferiority trials? *Trials*, *12*(Suppl 1), A29. doi:10.1186/1745-6215-12-S1-A29

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. doi:10.3389/fpsyg.2014.00781

Earp, B. D. (2017). The need for reporting negative results - a 90 year update. *Journal of Clinical and Translational Research*, *3*(S2), 1–4. doi:10.18053/jctres.03.2017S2.001

Everson-Stewart, S., & Emerson, S. S. (2010). Bio-creep in non-inferiority clinical trials. *Statistics in Medicine*, *29*(27), 2769–2780. doi:10.1002/sim.4053

Fidler, F., Thorn, F. S., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect. *Advances in Methods and Practices in Psychological Science*. doi:10.17605/OSF.IO/4GA56

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, *2*, 67–78. Retrieved from

http://www.rmm-journal.com/htdocs/volume2.html

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third Edit.). Boca Raton, FL: CRC Press.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Multilevel modelling of medical data. *Statistics in Medicine*, *21*(21), 3291–3315. doi:10.1002/sim.1264

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian T-Tests. *ArXiv E-Prints*. Retrieved from https://arxiv.org/abs/1704.02479

Harms, C. (2016). A Bayes Factor for Replications of ANOVA Results. *ArXiv E-Prints*.

Hayes, A. F. (2006). A Primer on Multilevel Modeling. *Human Communication Research*, *32*(4), 385–410. doi:10.1111/j.1468-2958.2006.00281.x

Hoffman, B. M., Papas, R. K., Chatkoff, D. K., & Kerns, R. D. (2007). Meta-analysis of psychological interventions for chronic low back pain. *Health Psychology*, *26*(1), 1–9. doi:10.1037/0278-6133.26.1.1

jamovi project. (2018). jamovi (Version 0.8) [Computer Software]. Retrieved from https://www.jamovi.org

JASP Team. (2018). JASP (Version 0.8.6) [Computer software]. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1961). *Theory of Probability* (3. ed., p. 447 p.). Oxford: Clarendon Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (in press). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*. doi:10.17605/OSF.IO/S5VDY

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective.

Psychonomic Bulletin & Review. doi:10.3758/s13423-016-1221-4

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. doi:10.3758/s13423-017-1272-1

Kruschke, J. K., & Meredith, M. (2017). *BEST: Bayesian estimation supersedes the t-test*. Retrieved from https://CRAN.R-project.org/package=BEST

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, *9*(9), e105825. doi:10.1371/journal.pone.0105825

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2018a). Justify your alpha. *Nature Human Behaviour*. doi:10.1038/s41562-018-0311-x

Lakens, D., Scheel, A. M., & Isager, P. M. (2018b, February). Equivalence testing for psychological research: A tutorial. PsyArXiv. Retrieved from psyarxiv.com/v3zkt

Locascio, J. J. (2017). Results blind science publishing. *Basic and Applied Social Psychology*, *39*(5), 239–246. doi:10.1080/01973533.2017.1336093

Machado, G. C., Maher, C. G., Ferreira, P. H., Pinheiro, M. B., Lin, C.-W. C., Day, R. O., ... Ferreira, M. L. (2015). Efficacy and safety of paracetamol for spinal pain and osteoarthritis: systematic review and meta-analysis of randomised placebo controlled trials. *BMJ*, *350*(mar31 2), h1225–h1225. doi:10.1136/bmj.h1225

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195. doi:10.2466/PR0.66.1.195-244

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*(2),

231–245. doi:10.1016/j.foodqual.2012.05.003

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. doi:10.1016/j.jmp.2015.11.001

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:10.1037/0033-2909.113.3.553

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. doi:10.3758/PBR.16.2.225

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi:10.1037/met0000061

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi:10.1007/BF01068419

Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine*, *20*(3), 453–472. doi:10.1002/1097-0258(20010215)20:3<453::AID-SIM803>3.0.CO;2-L

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. doi:10.1037/a0036731

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. doi:10.1016/j.cogpsych.2009.12.001

Wandel, S., Juni, P., Tendal, B., Nuesch, E., Villiger, P. M., Welton, N. J., ... Trelle, S. (2010). Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis. *BMJ*, *341*(sep16 2), c4675–c4675. doi:10.1136/bmj.c4675

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108