Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests

Daniël Lakens, PhD

Eindhoven University of Technology

D.Lakens@tue.nl

Neil McLatchie, PhD

Lancaster University

Peder M. Isager, MPhil

Eindhoven University of Technology

Anne M. Scheel, MSc

Eindhoven University of Technology

Zoltan Dienes, PhD

Sussex University

Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, IPO 1.33, PO Box 513, 5600 MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl.

Abstract

Researchers often conclude an effect is absent when a null-hypothesis significance test yields a non-significant *p*-value. However, it is neither logically nor statistically correct to conclude an effect is absent when a hypothesis test is not significant. We present two methods to evaluate the presence or absence of effects: Equivalence testing (based on frequentist statistics) and Bayes factors (based on Bayesian statistics). In four examples from the gerontology literature we illustrate different ways to specify alternative models that can be used to reject the presence of a meaningful or predicted effect in hypothesis tests. We provide detailed explanations of how to calculate, report, and interpret Bayes factors and equivalence tests. We also discuss how to design informative studies that can provide support for a null model or for the absence of a meaningful effect. The conceptual differences between Bayes factors and equivalence tests are discussed, and we also note when and why they might lead to similar or different inferences in practice. It is important that researchers are able to falsify predictions or can quantify the support for predicted null-effects. Bayes factors and equivalence tests provide useful statistical tools to improve inferences about null effects.

Keywords: Bayesian statistics, Frequentist statistics, hypothesis testing, TOST, falsification.

Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-013.

Author Note

All materials required to reproduce the analyses reported in this article are available at https://osf.io/67znq/ and https://github.com/Lakens/BF_TOST

Researchers are often interested in the presence or the absence of a predicted effect. Theories often predict there will differences between groups (e.g., older versus younger individuals), or correlations between variables. If such predicted patterns are absent in the data, the study fails to support the theoretical prediction. Other times, theories might predict the absence of an effect. In both these cases, it is important for researchers to base their conclusion that the data they have collected are in line with the absence of an effect on solid statistical arguments.

How can we conclude an effect is absent based on a statistical test of a hypothesis? All too often, non-significance (e.g., $p > .05$, for the conventional alpha level of 5%) is used as the basis for a claim that no effect has been observed. Unfortunately, it is not statistically or logically correct to conclude the absence of an effect when a non-significant effect has been observed (e.g. Dienes, 2016; Altman & Bland, 1995; Rogers, Howard, & Vessey, 1993). As an extreme example to illustrate the problem, imagine we ask two young individuals and two older individuals how trustworthy they would rate an interaction partner who did not reciprocate in a trust game. When we compare the trustworthiness ratings between these two groups in a statistical test, the difference turns out to be exactly zero, and there is no reason to conclude trustworthiness ratings differ between younger and older individuals. But is it really enough data to conclude the absence of an age difference in trustworthiness ratings? And if this is not enough data, what would be?

To conclude the absence of an effect, we need to quantify what 'an effect' would look like. It might be tempting to state that anything that is not zero qualifies as an effect, but this approach is problematic. First, this definition includes tiny effects (e.g., a correlation of $r = 0.00001$) which is practically impossible to distinguish from 0, because doing so would require billions of observations. Second, theories should ideally predict effects that fall within a specified range. Effects that are too small or too large should not be taken as support for a

theoretical prediction. For example, fluid cognitive abilities decline more rapidly in old age than crystallized abilities (Ritchie et al., 2016). A decline of 1 standard deviation in reaction time differences (which measure fluid cognitive abilities) from age 70 to age 80 in healthy adults could be a large, but valid prediction, while the same prediction would be implausibly large for verbal ability tasks. Finally, some effects are *practically* insignificant, or too small to be deemed worthwhile. For example, if a proven intervention exists to accelerate the rehabilitation process after a hip fracture, a new intervention that requires similar resources might only be worthwhile if it leads to a larger effect than the current intervention.

If we are interested in the absence of an effect, or want to falsify our predictions regarding the presence of an effect, it is essential to specify not just what our data would look like when the null hypothesis is true, but to also specify what the data would look like when the alternative hypothesis is true. By comparing the data against both models, we can draw valid conclusions about the presence *and* the absence of an effect. Researchers should always aim to design studies that yield informative information when an effect is present, as well as when an effect is absent (see the section on how to justify sample sizes in the discussion). We present two methods for evaluating the presence or absence of effects. One approach is based on frequentist statistics and known as equivalence testing (Schuirman, 1987), or more generally as inference by confidence intervals (Westlake, 1972). Researchers specify equivalence bounds (a lower and an upper value that reflect the smallest effect size of interest in a two-sided test) and test whether they can reject effects that are deemed large enough to be considered meaningful. The second method is based on Bayesian statistics and is known as the Bayes factor (Jeffreys, 1939). The Bayes factor measures the strength of evidence for one model (e.g., the null hypothesis) relative to another model (e.g., the alternative hypothesis); it is the amount by which one's belief in one hypothesis versus another should change after having collected data .

Bayes factors and equivalence tests give answers to slightly different questions. An equivalence test answers the question 'Can I reject the presence of an effect I would consider interesting, without being wrong too often in the long run?' A Bayes factor answers the question 'given the data I have observed, how much more (or less) likely has the alternative model become, compared to the null model?' The choice between frequentist and Bayesian statistics is sometimes framed as an ideological decision. Specifically, should one be interested in quantifying evidence (Bayesian) or in controlling error rates in the long run (frequentist)? Here, we present both approaches as research questions one might want to ask. We will focus on how to ask and answer both questions, and discuss when both questions are sensible.

Testing predictions using Bayes factors and equivalence tests

Most researchers are used to specifying and testing a null model, which describes what the data should look like when there is no effect. Both Bayes factors and equivalence tests additionally require researchers to specify an alternative model which describes what the data should look like when there is an effect. Bayes factors provide a continuous measure of relative support for one model over another model. Each model represents the probability of effect sizes assuming the hypothesis is correct (before the data have been taken into account), and is known as the *prior model.* After collecting data, a Bayes factor of e.g. 5 suggests that the data are 5 times more likely given the alternative hypothesis than given the null hypothesis, and a Bayes factor of 0.2 (1/5) suggests the data are five times more likely given the null hypothesis than the alternative hypothesis. Whether or not this should lead one to believe the null hypothesis is now more likely to be true than the alternative hypothesis depends on one’s prior belief in either hypothesis. When testing whether people can predict the future, a Bayes factor of 5 in favor of the alternative model might increase your belief in

precognition somewhat, but you might still think the probability of precognition is extremely low (see e.g. Dienes, 2008). Typically, in psychology or gerontology, one can ignore these prior probabilities of the theories (which can vary between people), and simply communicate the Bayes factor, which represents the evidence provided by the data, and let readers apply the Bayes factor to update their individual prior beliefs.

A common approach when calculating a Bayes factor is to specify the null hypothesis as a point (e.g., a difference of exactly zero), while the alternative model is a specification of the probability distribution of the effect a theory would predict. Specifying the two models is a scientific, not a statistical question, and requires careful thought about the research question one is asking (as we illustrate in the examples in this article). A Bayes factor provides a continuous measure of how much more likely the data are under the alternative hypothesis compared to the null hypothesis. A Bayes factor of 1 means the data are equally likely under both models, Bayes factors between 0 and 1 indicate the data are relatively more likely under the null hypothesis, and Bayes factors larger than 1 indicate the data are relatively more likely under the alternative hypothesis. For a more detailed discussion of Bayes factors, see Dienes (2014), Kass and Raftery (1995), or Morey, Romeijn, and Rouder (2016).

Equivalence tests allow researchers to reject the presence of effects as large or larger than a specified size while controlling error rates. To perform an equivalence test, researchers first have to determine a 'smallest effect size of interest', the smallest effect they deem meaningful. We then use this effect size to set a lower equivalence bound Δ_L (in a negative direction) and an upper equivalence bound Δ_U (in a positive direction). Next, we simply perform two one-sided significance tests against each of these equivalence bounds to examine whether we can reject the presence of a meaningful effect. This approach reverses the question that is asked in a null-hypothesis significance test: Instead of examining whether we can reject an effect of zero, an equivalence test examines whether we can reject effects that

are as extreme or more extreme than our smallest effect size of interest (in technical terms, each bound serves as a null hypothesis in a one-sided test, and a region around zero becomes the alternative hypothesis, see Figure 1). If we can reject both equivalence bounds (i.e., the first one-sided test shows that the effect in our data is significantly *larger* than Δ_L, *and* the second one-sided test shows that it is significantly *smaller* than Δ_U), then we can conclude that the effect is equivalent (see Figure 1). In other words, we can reject the hypothesis that the true effect is as extreme as or more extreme than our smallest effect size of interest. This approach is known as the 'two one-sided tests' approach (TOST), and can be seen as an improved version of the 'power approach' (Meyners, 2012) where researchers report which effect size they had high power to detect (see Example 4). The TOST approach is equivalent to examining whether a 90% confidence interval (for alpha = 0.05) around the effect falls between Δ_L and Δ_U, and concluding equivalence if the 90% confidence interval does not contain either equivalence bound (Westlake, 1972, for a related Bayesian procedure, see Kruschke, 2011, 2018). If we conclude that the true effect lies between the bounds whenever this procedure produces a significant result, we will not be wrong more often than 5%[1] of the time. For a more extensive introduction to the TOST procedure, see Meyners (2012), Rogers et al. (1993), or Lakens (2017).

Specifying alternative models

Most researchers are used to testing hypotheses using null-hypothesis significance tests where the null model is typically an effect of zero, and the alternative model is 'anything else' (see Figure 1). Specifying an alternative hypothesis in more detail might be a challenge

[1] We use an alpha level of 0.05 in all examples. The confidence interval corresponding to one-sided tests is $1 - 2 \times \alpha$. Thus, for an alpha level of 0.05, a 90% confidence interval is used.

at first. Both Bayes factors and equivalence tests require researchers to think about the size of the effect under the alternative hypothesis. Bayes factors additionally require the specification of the shape of the distribution instead of simply specifying the alternative hypothesis as a point. Because theories typically allow a range of effect sizes, it is common to specify the alternative hypothesis as a distribution of effect sizes, some of which may be more plausible than others. Because the information gained from performing these tests depends on how the alternative hypothesis is specified, the justification for the alternative hypothesis is an essential part of calculating and reporting equivalence tests and Bayes factors. In other words, the answer we get always depends on the question we ask. It is therefore important to clearly specify and justify the question that is asked, whenever reporting statistical tests.

One possible way to justify the alternative model is to base it on a question related to previous studies. For example, a Bayes factor can test whether the new data you have collected are more likely if the null hypothesis is true, or if the data are more likely under an alternative model specified using the effect observed in an earlier study. An equivalence test examines whether we can reject the effect size that was observed in the earlier study, or, as a more stringent test, an effect size the earlier study had a statistical power of 33% or 50% to detect (Lakens, Scheel, & Isager, 2018; Lenth, 2007; Simonsohn, 2015). Such a test would either conclude that the new data support the original claim, or that effect sizes as found in an earlier study can be rejected. An additional benefit of Bayes factors and equivalence tests is that the results can be inconclusive (e.g., the observed Bayes factor is close to 1, or the null-hypothesis significance test and the equivalence test are both non-significant), which would indicate that the performed study was not sensitive enough to distinguish between the null and the alternative model.

Another possible way to justify the alternative model is to determine a smallest effect size of interest. Sometimes a smallest effect size of interest can be specified based on

objective criteria, such as when the minimal clinically important difference for a measure has been determined (Jaeschke, Singer & Guyatt, 1989). Other times, researchers might be able to justify the smallest effect size of interest on the basis of a cost-benefit analysis. Provided this cost-benefit analysis is reasonable, rejecting the presence of a meaningful effect, or providing strong support for the null hypothesis, would then suggest that an effect is too small to be worth the resources required to reliably study it. Finally, researchers might suggest that the sample sizes that are used in a specific research area, or that can be reasonably collected, make it interesting to ask if we can reject effect sizes that could be studied reliably using such resources. Rejecting the presence of an effect that can be examined with the current resources suggests that researchers need to increase the resources invested in examining a specific research question.

Some simple considerations allow convenient ways of specifying the alternative model for Bayes factors (see Dienes, 2014). An alternative model indicates the plausibility of different effect sizes (e.g., differences between population means) given the theory. What effect size is most plausible given past research? A relevant previous study or meta-analysis might provide an indication of the effect that can be expected, and an alternative model based on a normal distribution centered at the expected effect can be used (see Figure 2). However, when building on effect sizes from the published literature, publication bias and 'follow-up bias' (Albers & Lakens, 2018) often leads to inflated effect sizes. Therefore, a useful assumption in such cases may be that smaller effects are more plausible than larger ones. We can represent this assumption about the plausibility of different effect sizes by a normal distribution with a mean of zero and a standard deviation that sets the rough scale of the effect. As the mean of the distribution is zero, smaller effects are more likely than bigger ones. If the theory predicts effects in a positive direction, we remove effects below zero, and are left with a half-normal distribution (Dickey, 1973; see Figure 2). When modeled as a

normal distribution the implied plausible maximum effect is approximately twice the standard deviation. Often this is a good match to scientific intuitions; and when it is not, this feature turns out not to greatly affect results (Dienes, 2017). Dienes (2014) suggests setting the standard deviation for a normal (or half-normal) distribution to the predicted effect. Bayes factors also depend on the model of the null hypothesis, which can be specified as flexibly as the model of the alternative hypothesis. Here, the null hypothesis is always specified as a point-null hypothesis, so we do not further describe it in the examples.

In four detailed examples we illustrate different ways to specify alternative models. These examples are also used to explain how to calculate, report, and interpret Bayes factors and equivalence tests. The specific examples were chosen to demonstrate different ways of specifying the alternative model for Bayes factors (e.g., based on past research, the measures used in the study, effect sizes considered interesting by the authors elsewhere in their study) and equivalence tests (e.g., based on the smallest effect that a previous study could have detected, the smallest clinically relevant effect size, the smallest effect sizes considered interesting by the authors elsewhere in their study). The four examples also demonstrate a range of conclusions that can be drawn, both for Bayes factors (e.g., evidence for H0, evidence for H1, inconclusive evidence) and equivalence tests (e.g., reject meaningful effect sizes, fail to reject meaningful effect sizes). Note that our re-analyses are based on the statistics reported in each paper and are rounded to one or two digits after the decimal, meaning that the values we report may sometimes differ slightly from those that would result from analyzing the raw data. We provide reproducible R scripts for all of the examples as well as instructions for how to calculate Bayes factors and equivalence tests from summary statistics or raw data in online calculators (Dienes, 2008; Rouder, Speckman, Sun, Morey, & Iverson, 2009), simple spreadsheets (Lakens, 2017), the new statistical point-and-click software packages JASP (JASP team, 2018) and jamovi (jamovi project, 2018), and in R (R

core team, 2017) at https://osf.io/67znq/ (all of these software solutions are free, and most are open source).

Following these examples, we also show how to design informative studies that can provide support for the absence of a meaningful effect. The conceptual differences between Bayes factors and equivalence tests are discussed, but we also note how they often (but not always) lead to comparable inferences in practice.

Example 1: Emotion regulation preference in older vs. younger adults.

Martins, Sheppes, Gross, and Mather (2016) explored the relationship between age and emotion regulation by testing participants' preference for distraction vs. reappraisal while viewing images of varying affective intensity. Contrary to their prediction, they did not find a difference in the proportion of trials in which younger (n = 32) rather than older men (n = 32) preferred the distraction strategy for negative affective images (M_{young} = 0.34, SE_{young} = 0.03; M_{older} = 0.32, SE_{older} = 0.03), $t(62) = 0.35$, $p = .73$, $d = 0.09$. They conclude their discussion by stating that they *"...found no age differences in regulation preferences in negative contexts..."*. However, a non-significant *t*-test is not sufficient to support the conclusion that there are no age differences. We can calculate a Bayes factor and perform an equivalence test to formally examine whether we can conclude the absence of a difference.

Bayes Factors. To calculate a Bayes factor we must first specify a model of the data assuming there is a true effect. We will give two examples, one based on no prior knowledge other than the scale that is used, and one model based on results from a previous study by Scheibe, Sheppes, and Staudinger (2015). This example will illustrate that using vague alternative models, based on the limits of the scale, will allow quite large effect sizes, and

how an alternative model based on more specific scientific information is typically a more interesting question to ask.

One relatively objective way to specify a model for the data if there is a true effect would be to consider the range of possible results based on the scale used in the research. The maximum possible effect when calculating proportions is a difference of 1 (i.e., if all of the older adults prefer distraction in 100% of the trials, and all younger adults prefer reappraisal in 100% of the trials). Of course, this extreme outcome is very unlikely, and if there is a true effect, smaller differences should be more plausible. We can model this prior belief about smaller differences being more likely than larger differences by using a half-normal distribution, with a standard deviation of 0.5. In such a model, the plausibility of differences is distributed across a wide range of possible outcomes, but smaller effects are considered more plausible than larger effects.

Calculating a Bayes factor, based on the observed data that expresses the relative support for an alternative model (specified as a half-normal distribution with a standard deviation of 0.5) over a point-null hypothesis, yields $B_{H(0, 0.5)} = 0.13$. Note that subscripts for alternative models with half-normal distribution are reported as $B_{H(0, S)}$, where 'H' indicates a half-normal, centered at 0, with standard deviation of S, while $B_{N(M, S)}$ indicates a normal distribution with mean M and standard deviation S, following Dienes, 2014. This means that the data are 1/0.13 = 7.69 times more probable under the null model than under the alternative model. We conclude that that there is strong evidence for the null hypothesis relative to the (rather vague) alternative model of a difference between groups.

We recommend reporting robustness regions for each Bayes factor that is reported. A robustness region specifies the range of expected effect sizes used when specifying the alternative model that support the same conclusion (e.g. evidence for H1, evidence for H0, or inconclusive outcomes). Robustness regions are reported as *Rob. Reg.* [L, U], and give the

lower and upper effect size for the alternative model that leads to the same conclusion, given a certain Bayes factor threshold. In this article, we consider Bayes factors larger than 3 as support for the alternative, and Bayes factors smaller than ⅓ as support for the null model (cf. Jeffreys, 1939). For this Bayes factor, the robustness region is [0.147, ∞]. The fact that the upper bound of the robustness region goes to infinity indicates that all effects larger than the rough scale of effect used to specify the alternative hypothesis generate the same conclusion.

We can also specify an alternative model that is based on existing information about the effect we are examining - and thus is more relevant to actual inference in the scientific context. Martins and colleagues build on a previous study by Scheibe, Sheppes, and Staudinger (2015) where the same paradigm was used to examine the difference between distraction or reappraisal choices in older and younger participants. Based on this study by Scheibe and colleagues, who reported that 40.5% of young adults chose distraction compared to 48.5% of older adults, we have some reason to expect a mean difference of 0.405 – 0.485 = -0.08. Note this difference is in the opposite direction to the result obtained by Martins and colleagues, who found a mean difference of 0.338 - 0.321 = 0.017. Thus, the mean difference is entered as a negative value to reflect the fact the mean difference went in the opposite direction to that predicted. We can thus choose an alternative hypothesis with a half-normal distribution centered on 0 and a standard deviation of 0.08 (see Figure 3). Now, with this more informed hypothesis, we find that the data provided by Martins and colleagues offers only weak, inconclusive evidence for the null hypothesis, $B_{H(0,\ 0.08)} = 0.42$, *Rob. Reg.* [0, 0.189]. If Martins and colleagues wish to obtain strong evidence for either hypothesis, they need to collect more data.

Equivalence test. To perform the equivalence test we must start by specifying the smallest effect size of interest. In their previous study, Scheibe and colleagues (2015) did not

explicitly state which effect size they were interested in. In this case, one way to decide upon a smallest effect size of interest is to assume that the authors were only interested in effects that could have yielded a significant result, and then look at the effect sizes that could have been statistically significant given the sample size they collected. For any specific sample size and alpha level, a critical test value can be calculated, and test statistics larger than this value will yield significant *p*-values. Because Scheibe et al. (2015) collected 77 participants in total, and used an alpha level of .05, the critical *t*-value is 1.99. This critical *t*-value can be transformed into a 'critical standardized effect size' of Cohen's $d = 0.45$. Only effects larger than 0.45 (or smaller than -0.45) would have been statistically significant in this study. If we assume that sample size in this study was chosen, at least implicitly, based on effect sizes deemed interesting by the researchers who designed this study, we can set the smallest effect size of interest to an absolute effect size of $d = 0.45$. It might of course be that the authors did not think about their sample size at all, and would be interested in smaller effect sizes than $d = 0.45$. In other words, our assumption might be wrong, which highlights the important responsibility of authors to specify their smallest effect size of interest. As De Groot (1969) noted: "Anyone publishing a hypothesis should therefore indicate in particular how crucial experiments can be instituted that may lead to the refutation or abandonment of the hypothesis."

Now that the smallest effect size of interest has been determined (based on the study by Scheibe et al., 2015), we can proceed by reanalyzing the results from Martins et al, (2016) with an equivalence test against equivalence bounds of $d = -0.45$ and $d = 0.45$ using the two one-sided tests procedure. The first one-sided test indicates that we can reject effects as small as or smaller than $d = -0.45$ (or, in raw scores, a difference of -0.088), $t(62) = 2.17$, $p = 0.017$. However, the second test shows we cannot reject effects as large or larger than $d = 0.45$ (in raw scores, 0.088), $t(62) = -1.47$, $p = 0.074$. Both one-sided tests need to be significant to

conclude equivalence, so given the observed data and the alpha level we decided on, we cannot conclude that the effect is statistically equivalent. It is common to report an equivalence test by only providing the one-sided test with the higher *p*-value (if this test is significant, so is the other). So, we would conclude: Based on equivalence bounds of $d = -.45$ and $d = 0.45$, we cannot reject effect sizes that we still consider meaningful, $t(62) = -1.47$, $p = 0.074$. Because the effect was also not statistically different from 0 in a traditional null-hypothesis test (as reported by Martin and colleagues), the result is inconclusive. We can neither conclude that the effect is different from zero, nor that the effect is too small to matter. We need to collect more data to draw conclusions about the presence of an effect, or the absence of a meaningful effect (or both).

Discussion. Martins and colleagues did not observe a statistically significant difference between younger and older men in their choice of distraction as a method for emotion regulation. What can we conclude based on the Bayes factor and equivalence test? The equivalence test shows that based on the current data, we cannot reject the presence of effects as extreme as $d = \pm 0.45$ or more extreme. Whether or not effects of $d = \pm 0.45$ are interesting should be discussed by researchers in this field. If smaller effects are deemed interesting, larger studies need to be performed to draw conclusions. The Bayes factor we calculated for an uninformed alternative model suggest that the data provides stronger relative support for a null model than for a model that predicts effects up to a difference in proportions of 1. However, there is not enough evidence to prefer a null model over a more informed alternative model that predicts smaller effects[2]. That is, based on the best estimate

[2] Taking the ratio of the two Bayes factors in the current example, we see that the data were 0.42/0.07 = 6 times more probable under the more informed hypothesis than under the less informed hypothesis. Bayes factors calculated for different alternative hypotheses can be

of which effect sizes would be reasonable (based on related earlier research), the data are non-evidential. We would tentatively answer the question about whether an effect is absent as follows: We cannot reject effect sizes that are still deemed interesting (d = 0.45) and there is no reason to interpret the data as strong relative evidence for a null model, compared to an alternative model informed by previous findings. Thus, it seems prudent to suspend judgment about the absence of an effect until more data is available.

Example 2: Comparing self-reported chronic pain in two age groups.

Shega, Tiedt, Grant, and Dale (2014) studied the relationship between self-reported chronic pain and other indicators of decreased quality of life in a sample of 2902 older adults (from the National Social Life, Health, and Aging Project). Pain intensity was measured using a 7-point Likert scale (0 = no pain, 6 = the most pain imaginable). They report non-significant changes in reported pain across age groups (age 62-69: M = 2.03, SE = 0.084, n = 1020; age 70-79: M = 1.98, SE = 0.057, n = 1015; age >79: M = 2.14, SE = 0.102, n = 554; p = .254). Based on the large sample size, we can assume that the effect size is accurately estimated to be close to zero, but to test for the absence of an effect, we need to calculate Bayes factors or perform equivalence tests.

Bayes factors. Shega et al. (2014) do not explicitly state that they have a directional hypothesis (i.e., they are interested both in whether older adults experience higher or lower pain intensity than younger adults). Past research suggests that the experience of pain peaks between the age of 50-65 and then plateaus for the remaining years of life (Gibson & Lussier, 2012). One could therefore model the alternative such that small effects around zero are

---

compared in such a way when they have been calculated using the same data and against the same model of H0 ($B_{H1/H0} / B_{H2/H0} = B_{H1/H2}$).

considered most plausible, and with effects in either direction considered increasingly less plausible. Thus, an alternative model based on a normal distribution centered on zero would be appropriate.

However, the model of the alternative hypothesis depends on what question we want to answer. Researchers have extensively studied the clinical importance of pain ratings (see e.g. Dworkin et al., 2008). Reductions in pain of approximately 10-20% were reported by Dworkin et al. to be noticeable, and reductions of approximately 40% were judged to be meaningful. Twenty percent corresponds to a difference of 1.21 points on a seven point scale. If the question is whether clinically important pain differences occur between different age groups, we can model the alternative hypothesis as a half-normal with an SD of 1.21. Specifying the alternative in this way allows us to ask the question if the observed data provide more relative evidence for a model that should be expected when a noticeable difference exists than for a null model.

For the three comparisons in Shega and colleagues, and assuming the authors were interested in effects around the size of noticeable pain differences, we obtain strong evidence for the null hypothesis when comparing participants aged 62-69 with participants aged 70-79, $B_{H(0,\ 1.21)} = 0.128$, *Rob. Reg.* [0.454, ∞], and with participants older than 79, $B_{H(0,\ 1.21)} = 0.24$, *Rob. Reg.* [0.878, ∞]. Here, the evidence in favour of the null hypothesis suggests that the data are 1 / 0.128 = 7.81 and 1 / 0.24 = 4.17 times more probable assuming the null hypothesis is true than assuming the alternative hypothesis is true, meaning that the Bayes factors provide at least some evidence for the null hypothesis relative to the alternative hypothesis. In contrast, we find only weak evidence for the null hypothesis when comparing participants aged 70-79 with those older than 79, $B_{H(0,\ 1.21)} = 0.37$, *Rob. Reg.* [0, 1.361]. The robustness regions indicate that the Bayesian inferences are robust to a broad range of models that could be used to specify H1. In the one case where we find weak evidence for the null

hypothesis, the conclusion would only change if an effect less than the minimal clinically relevant effect size was specified for the prior distribution.

Equivalence test. In order to perform the two one-sided tests (TOST) procedure we need to specify equivalence bounds based on a smallest effect size of interest. When examining the minimal clinically important difference, researchers estimate the smallest change in pain ratings that leads to a noticeable change on other clinically relevant scales. For example, Kelly (2001) reports that the smallest effect size that leads to an individual to report feeling "a little better" or "a little worse" is 12 mm (95% CI [9; 12] on a 100-mm visual analogue scale of pain intensity (this is very similar to the 10% difference argued as just noticeable by Dworkin et al., 2008, cited above). To be conservative, we can use a 9 mm difference as the smallest effect size of interest (because smaller differences are not clinically meaningful), which corresponds to a difference of approximately 0.55 points on a 7-point scale (used by Shega and colleagues).

We can now perform equivalence tests for the differences in self-reported pain between the three age groups reported by Shega, Tiedt, Grant, and Dale (2014). Because we perform three tests, we will use Bonferroni correction to control the type-I error rate and adjust the alpha level to 0.05/3 = .017 for each comparison (for an explanation of corrections for multiple comparisons, see Armstrong, 2014). We will begin with the largest difference reported for people in the age of 70-79 and those older than 79 (based on the means, sample sizes, and standard deviations reported earlier). With an alpha of 0.017 and equivalence bounds set to ±0.55 (expressed as a raw mean difference), both one-sided tests (against a difference of -0.55 and 0.55, respectively) are significant, $t(904.22) = 3.30$, $p < .001$, which means we can reject the presence of an effect that is large enough to be clinically meaningful (see Figure 4). Note that we report Welch's *t*-test, as indicated by the fractional degrees of

freedom, because sample sizes are unequal, and standard deviations can be assumed to be unequal as well (see Delacre, Lakens, & Leys, 2017). The same conclusion holds for the difference between participants in the age-range of 62-69 and 70-79, $t(1791.71) = -4.88$, $p < .001$, and for the difference between participants in the age range of 62-69 and older than 79, $t(1246.34) = 3.30$, $p < 0.001$.

Discussion. Shega and colleagues examined whether there were differences in self-reported chronic pain across age groups. When we analyze their data with Bayes factors, we see consistent support for a null model compared to an alternative model that is specified based on 'clinically important' differences (as reported by Dworkin et al., 2008). When we analyze the data with equivalence tests, we find that we can reject the presence of effect sizes large enough to be 'just noticeable' (as reported by Kelley, 2001). Thus, we can conclude that it seems unlikely that there are substantial differences in self-reported pain across age-groups. Where the Bayesian model for the alternative was based on the distribution of effect sizes observed in Dworkin et al. (2008), the equivalence bounds were based on work by Kelley (2001), establishing a single effect size that represents the minimal clinically relevant difference. The two justifications for the alternative model differ slightly, and illustrate how researchers can use different justifications when quantifying their alternative models. Justifications for alternative models should be transparent, and are always open for debate.

Example 3: Correlating Big Five openness with Eriksonian Ego integrity.

Westerhof, Bohlmeijer, and McAdams (2017) studied the relationship between concepts of ego integrity and despair from Erikson's theory of personality and the factors of the Big Five model of personality traits. They predicted that the Big Five trait 'openness' (as measured by the NEO-FFI) should be related to ego integrity (as measured by the

Northwestern Ego Integrity Scale), and they conclude that this hypothesis is supported by a significant positive correlation (r = 0.14, p = .039). They also report a nonsignificant correlation between openness and despair (r = 0.12, p = .077). Note the diametrically opposed conclusions drawn from these results, although the difference between the two correlations is very small (r = 0.14 and r = 0.12). Can the conclusion that there is no relationship between openness and despair be statistically justified?

Bayes Factors. The authors' willingness to interpret r = .14 as evidence for a relationship between openness and ego integrity provides an approximate scale of the effect size that they would count as evidence for a relationship between openness and despair. Since it is preferable to define models in raw effect sizes, we transform these values into raw effects by calculating r x *SD1/SD2* and obtain b = 0.19 for r = .14 and b = 0.20 for r = .12, respectively ($SD_{openness}$ = 0.6, $SD_{ego\ integrity}$ = *0.8*, $SD_{despair}$ = 1.0). We can model the alternative hypothesis for a correlation between openness and despair using a half-normal distribution with a standard deviation of b = .19 (r = .14) (see Figure 5). A Bayes factor for the observed correlation of b = .20 (r = .12) yields weak support for the alternative hypothesis, $B_{H(0, .19)}$ = 2.98, *Rob. Reg.* [0, 3.024]. We should not be too quick to interpret the nonsignificant result as evidence for the null hypothesis. In fact, the data offer weak evidence for an alternative model similar to an expected distribution for significant effects that are reported by the authors. The appropriate response would be to suspend judgment and recruit more participants.

Equivalence test. In this case, we know that the authors are willing to treat a significant correlation of r = .14 as support for their theoretical prediction, so we might assume the authors consider correlations of r = .14 large enough to be meaningful. Note that

we again draw inferences about researchers' theoretical predictions based on the results and conclusions they report. It would be a great improvement to current research practices if authors would explicitly state which effect sizes they consider to be too small to be meaningful (and provide a good reason for that judgment). We can perform the two one-sided tests procedure for correlations (which relies on Fisher's *z* transformation), to formally test whether we can reject the presence of an effect as large as or larger than $r = .14$ for the correlation between openness and despair. Not surprisingly given the observed correlation of $r = .12$, we cannot reject the presence of effect sizes as large of larger than $r = .14$: The equivalence test is not significant, $z = -0.30$, $p = 0.38$. We cannot conclude the absence of meaningful effects if we consider effects of $r = 0.14$ meaningful. Note that it is possible to observe an effect of $r = 0.12$ and reject effects of $r = 0.14$, but the required sample size to detect such small differences would be extremely large (to achieve 80% power for such a test, more than 10,000 observations are required). To reject the presence of small effects, large samples are needed, such that the 90% confidence interval around the observed effect becomes extremely narrow.

Discussion. The Bayes factor suggests there is no reason to treat a correlation of $r = .12$ as evidence for the absence of an effect. As the current data provides inconclusive evidence for either hypothesis, more data are needed to reach a conclusion. The equivalence test shows that we can certainly not reject effect sizes of $r = .14$, which had been interpreted as evidence for the presence of an effect for other correlations. Given that we can neither reject the null nor the smallest effect size of interest, the results are inconclusive.

Example 4. No Short-Term Differences in Memory After Reward Cues

Spaniol, Schain, and Bowen (2013) examined whether anticipating a reward would enhance long-term memory formation equally well in older and younger individuals. They

found support for this prediction in two studies. They also tested the hypothesis that an effect of reward cues should be absent when a recognition task was presented after only a short delay. They concluded in Experiment 2: “Second, neither age group showed an effect of reward on memory at the short delay.” There was no statistically significant difference in the recognition hit rate in the short delay condition for trials where low reward or high reward stimuli were presented for younger participants (n = 32, M = 0.76, SD = 0.17, and M = 0.77, SD = 0.14, respectively) or older participants (n = 32, M = 0.75, SD = 0.12, and M = 0.76, SD = 0.12, respectively).

The authors were fully aware that a non-significant result does not allow one to conclude the absence of an effect. Therefore, in Experiment 1, Spaniol and colleagues (2013, p. 731) write how an interaction effect “failed to reach significance […] even though the power to detect a medium-sized interaction was high”. The authors rely on what is known as the ‘power approach’ to conclude a meaningful effect was absent. In the power approach, a non-significant *p*-value is used to argue for the absence of an effect that a study had high power to detect. For example, if a study had 99% power to detect a medium effect size, and no significant test result is observed, researchers using the power approach would feel justified in concluding the absence of a medium effect, because a population effect of medium size would almost certainly have yielded a significant *p*-value in the experiment. Meyners (2012) explains that this approach, although it was common and even recommended by the Food and Drug Administration in the 1980s, should no longer be used. One important reason why equivalence tests are preferable is that even practically insignificant differences will be statistically significant in very large samples. When using equivalence tests, researchers can instead conclude that such effects are significant *and* equivalent (i.e., statistically different from zero, but also too small to matter). In addition, Bayes factors can show that a study with low statistical power for interesting effect sizes provides evidence for

H0 relative to H1, or that a high-powered non-significant result provides no evidence for H0 relative to H1 (Dienes & McLatchie, 2018).

Bayes factor. We can compare the difference scores for hit rates in the memory task between low- and high-reward trials for older and younger people (i.e., we are examining the interaction effect between reward and age). The scale of the effect expected under the alternative hypothesis when assessing the impact of high versus low rewards on recognition following long and short delays can be inferred from Spaniol et al.'s first experiment. In Experiment 1, low-reward or high-reward stimuli were presented for younger participants (M = 0.54, SD = 0.18, and M = 0.61, SD = 0.16, respectively) and older participants (M = 0.61, SD = 0.15, and M = 0.64, SD = 0.14, respectively). Thus, the obtained difference between high and low reward stimuli between younger and older adults in Experiment 1 was: (0.61 - 0.54) - (0.64 - 0.61) = 0.04. This provides an approximate scale of effect size that is expected under the alternative hypothesis fo r Experiment 2. For the short delay conditions, the resulting Bayes factor provides only weak evidence for the null hypothesis, $B_{H(0,\ 0.04)} = 0.44$, *Rob. Reg.* [0, 0.054].

Equivalence test. When presenting a non-significant result, the authors discuss the statistical power they had to detect a medium effect size (Cohen's d = 0.50), which corresponds to a raw score of 0.039. If we assume that this is the smallest effect size they considered interesting, we can set the equivalence bounds to $d = \pm 0.5$, or mean differences of -0.039 and 0.039. When we calculate a 90% confidence interval around the mean difference for hit rates in the memory task between low and high reward trials for older and younger people, it ranges from -0.033 to 0.033, which does not overlap with the equivalence bounds (see Figure 6). We can thus conclude that the difference scores for younger and older

participants do not themselves differ more than what we consider a 'medium' effect size. The two one-sided tests against the equivalence bounds both give $t(62) = 1.98$, $p = 0.026$ (the tests are identical when symmetrical bounds are used and the observed effect is exactly zero). It should be noted that setting equivalence bounds based on the benchmarks proposed by Cohen (1988) is not considered best practice in equivalence testing (see Lakens et al., 2018). While in this example we have assessed equivalence with respect to what the authors claim to be interested in, we recommend that researchers specify equivalence bounds based on theoretical predictions or practical importance, where possible, and only use benchmarks as a last resort.

Discussion. The reanalysis of the present results using a Bayes factor and the two one-sided tests approach supports the conclusions of Spaniol, Schain, and Bowen (2013). The Bayes factor suggests that the data offers only weak, inconclusive evidence for the null hypothesis, whereas the equivalence test allows us to reject the presence of effects as large or larger than a 'medium' effect.

General discussion

We have provided several detailed examples to illustrate how researchers in the field of gerontology could improve inferences about null effects with Bayes factors and equivalence tests. As we mentioned in the beginning, these calculations can be performed based on summary statistics using free and easy to use software solutions (see https://osf.io/67znq/ for instructions), which in recent years have substantially lowered the barriers to making use of the two methods.

To restrict analytic flexibility and preserve the validity of confirmatory hypothesis tests, it is important to specify the alternative model before looking at the data. We recommend that researchers preregister their hypotheses (preregistration can be done

independently of the publication format of an article; for further resources see https://cos.io/prereg/), and we especially recommend the Registered Reports format (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014), where the study design is peer-reviewed before the data are collected. It is an excellent way to receive feedback from peers about the justification for the alternative model, and whether the question that is asked by calculating a Bayes factor or by performing an equivalence test is considered interesting. Specifying the alternative model before collecting data for hypothesis tests is also important because the sample size required to design an informative study depends, in part, on the alternative model. When testing theories, the values used to specify the alternative model for a Bayes factor, or the smallest effect size of interest for an equivalence test, should be chosen based on reasons internal to the theory, such as effect sizes that are theoretically deemed similar. When the smallest effect size of interest is chosen simply on the basis of resources (i.e., available funds to pay subjects) the statistical inference does not provide grounds for theory testing (for a more detailed discussion, see Lakens et al., 2018).

Justifying sample sizes for equivalence tests and Bayes factors

Researchers should aim to design studies that yield informative results about the presence, and absence, of meaningful effects. It is important that the sample size justification for studies reflects both the possibility that the alternative hypothesis is true, and the possibility that the null hypothesis is true. When designing studies in which one plans to draw inferences based on equivalence tests, one can perform an a-priori power analysis to make sure a study has high statistical power to reject the smallest effect size of interest. When using Bayes factors, one can design informative studies by using the results from previous research to determine the minimum sample size required to obtain sufficient evidence.

Bayes factors. For Bayes factors, one could estimate the sample size required to provide noteworthy evidence for both the null and the alternative hypothesis, where 'noteworthy' depends on the Bayes factor you would like to observe (i.e., Bayes factors larger than three, six, or ten have been recommended as noteworthy evidence; see Jeffreys, 1939; Schönbrodt & Wagenmakers, 2017). The value of such an estimate would be to allow researchers to make an informed decision regarding how many participants might be required for a given study. However, after the data have been collected, inferences depend only on the obtained Bayes factor and not on prior sample size calculations.

Consider a researcher who aims to determine whether frail older adults demonstrate cognitive deficits relative to non-frail older adults. A recent study by Bunce, Batterham, and Mackinnon (2018) measured verbal fluency across frail and non-frail adults with an animal naming task and reported that non-frail adults recalled significantly more names of animals (M = 11.53, SD = 3.52, n = 304) than did frail adults (M = 10.11, SD = 3.20, n = 154), $t(456) = 4.20$, $p < .001$. The mean difference reported by Bunce and colleagues can be used to provide a model of the alternative hypothesis, as the original study provides a rough scale of effect. Thus, the alternative hypothesis can be modeled with a half-normal distribution with a mode of zero and a standard deviation of 11.53 - 10.42 = 1.42. The standard error reported by Bunce and colleagues provides an estimate of the level of noise in the measurement, SE = (M1-M2)/t = 0.338. Given that we have a model of the alternative hypothesis, how many participants would we need to recruit in order to meet the desired level of evidence if the study were to (1) obtain evidence for the alternative (e.g., obtain a mean difference of 1.42) or (2) obtain evidence for the null (e.g., obtain a mean difference of 0)? One can calculate a series of Bayes factors in which the number of participants is varied from 1 to as many participants as the researcher has the resources to recruit. By adjusting the standard error of the data obtained by Bunce and colleagues (which varies as a function of sample size) it is

possible to calculate the number of participants required to achieve a desired level of evidence. Based on the alternative model specified above, one would need to collect a total of 82 participants to provide noteworthy evidence for the alternative hypothesis (here taken to be a Bayes factor larger than 3), and 207 participants to provide noteworthy evidence for the null (here taken to be a Bayes factor smaller than 0.33). In practice, one can also test sequentially by adding additional data to guarantee sufficient evidence (for a discussion, see de Heide & Grunwald, 2018; Dienes, 2016; Schönbrodt & Wagenmakers, 2017).

Equivalence tests. For equivalence tests, a researcher can perform a-priori power analyses to calculate the number of participants that are required to achieve a desired probability of finding an statistically equivalent result, given certain equivalence bounds, the alpha level, and the true effect size (e.g. an effect size of 0). This can be done in R, using the power analysis functions of the TOSTER package (Lakens, 2017), or in an online calculator (http://powerandsamplesize.com/Calculators/). Imagine a researcher who wants to test if a reaction time game improves elderly adults' choice reaction time in a driving simulator. They determine that the smallest reaction time difference they would be interested in is the time it takes a car to travel 1 meter at a speed of 20 mph (32.1869 km/h), which is $1 \text{ m} / \frac{32186.9 \text{ m}}{3600 \text{ s}} = 0.1118 \text{ s}$. Roenker and colleagues (2003) report an average standard deviation of 0.268 seconds for choice reaction times in a driving simulator task in a sample of 55– to 86-year-olds (calculated as the mean of the nine standard deviations for Choice RTs in Table 1 in Roenker et al., 2003). Based on these data, 0.1118 seconds would correspond to a standardized mean difference of $d = \frac{0.1118 \text{ s}}{0.268 \text{ s}} = 0.42$. To be a bit more conservative, the researcher decides to set equivalence bounds at $d = \pm 0.4$. Assuming a true effect of $d = 0$, the TOSTER power analysis for a two-sample *t*-test shows that 272 participants (136 per group) would be needed to reject these bounds with 90% power at an alpha level of 5%. Equivalence

tests to reject effects of a specific size require slightly larger sample sizes than would be required to have the same power to to detect these effect sizes in a null hypothesis test (Lakens, 2017) and the closer the equivalence bounds are to zero, the larger the sample size needed to have high power for the equivalence test.

Should you use Bayes factors or equivalence tests?

Although Bayes factors and equivalence tests ask slightly different questions, they can give converging answers. For example, an equivalence test can show that we can reject the presence of a meaningful effect, while the Bayes factor informs us that the data provide substantially more evidence for the null model than for the alternative model (see Example 2). Despite often leading to similar conclusions, these two approaches differ both on a philosophical level (e.g., how do we define probability?) and on a practical level (e.g., do we want to incorporate prior information in our statistical inferences or not?). There is no reason to limit yourself to asking only a single question from your data: one recommendation is 'a B for every *p*'; reporting a Bayes factor alongside every significance test. If the methods lead to different answers, this is often informative. It can lead one to reflect on the difference between the two approaches, but as long as both tests are used and interpreted correctly, their answers should be interesting regardless of whether Bayes factors and *p*-values agree. Evidence and errors are closely related in practice, and Bayesian and frequentist statistics will often lead to similar conclusions (Jeffreys, 1939). All else being equal, the larger the Bayes factor discriminating the null and alternative model, the lower the error rates in deciding in favor of one or the other (for discussion, see Dienes & McLatchie, 2018).

It is possible to choose one approach over the other. Equivalence tests follow from a Neyman-Pearson perspective on statistical inferences, where the main goal is to accept or reject hypotheses without being wrong too often by controlling type-1 and type-2 error rates.

If a researcher wants to use statistical tests to guide their behavior in the long run, and the smallest effect size of interest is theoretically or practically important, then equivalence testing would naturally suggest itself as a method. If on the other hand the researcher is interested in quantifying relative evidence for two competing models, and the most salient aspect of the predicted effect is its rough scale or its maximum, then the Bayes factor would be a natural method. If reliable prior information is available, or theories make more specific theoretical predictions, Bayesian approaches become increasingly interesting. Remember to always ask yourself if hypothesis testing is appropriate, or whether you might simply want to estimate the size of an effect instead (for a related approach to the TOST procedure, see the ROPE approach by Kruschke, 2011). Note that although we have focused on hypothesis tests in this article, reporting and interpreting effect size estimates is important, and should always accompany hypothesis tests.

Conclusion

Embracing methods that allow us to provide support for the absence of a predicted or meaningful effect has the potential to greatly improve our statistical inferences. It is logically incorrect to conclude the absence of an effect simply on the basis of a non-significant result (e.g., $p > .05$), and we should aim to prevent this common mistake. This will require researchers to specify not just what their data would look like when there is no effect, but also what their data would look like when there is an effect. Quantifying a smallest effect size of interest, or the predictions of a theory, can be a challenge and will require discussions among the researchers in a field. But being able to falsify predictions, or corroborate hypotheses that predict the absence of an effect, is of utmost importance for scientific progress.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ, 311(7003)*, 485. https://doi.org/10.1136/bmj.311.7003.485

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics, 34(5)*, 502-508.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100(3)*, 603–617. https://doi.org/10.1348/000712608X377117

Bunce, D., Batterham, P. J., & Mackinnon, A. J. (2018). Long-term associations between physical frailty and performance in specific cognitive domains. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences.*

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.

de Heide, R., & Grünwald, P. D. (2018). Why optional stopping is a problem for Bayesians. *ArXiv:1708.08278 [Math, Stat]*. Retrieved from http://arxiv.org/abs/1708.08278

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology, 30(1)*. https://doi.org/10.5334/irsp.82

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35, 285-305.

Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific and statistical inference. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781, 1-17. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z. (2017). https://www.youtube.com/watch?v=g9bIfZ4KqCQ&t=5164s

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review, 1–12*.

Dworkin, R. H., Turk, D. C., Wyrwich, K. W., Beaton, D., Cleeland, C. S., Farrar, J. T., ... & Brandenburg, N. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The Journal of Pain, 9(2),* 105-121.

Gibson, S. J., & Lussier, D. (2012). Prevalence and relevance of pain in older persons. *Pain Medicine, 13,* S23-S26.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10(4),* 407–415. https://doi.org/10.1016/0197-2456(89)90005-6

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

JASP Team (2018). JASP (Version 0.8.5) [Computer software].

Jeffreys, H. (1939). *Theory of probability*. Oxford: Clarendon

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

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. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245918771304

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

Lenth, R. V. (2007). Post hoc power: tables and commentary. *Iowa City: Department of Statistics and Actuarial Science, University of Iowa*. Retrieved from https://pdfs.semanticscholar.org/fbfb/cab4b59e54c6a3ed39ba3656f35ef86c5ee3.pdf

Martins, B., Sheppes, G., Gross, J. J., & Mather, M. (2016). Age differences in emotion regulation choice: older adults use distraction less than younger adults in high-intensity positive contexts. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, gbw028.

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference, 26(2),* 231–245. https://doi.org/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.

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

Ritchie, S. J., Tucker-Drob, E. M., Cox, S. R., Corley, J., Dykiert, D., Redmond, P., Pattie, A., Taylor, A. M., Sibbett, R., Starr, J. M., & Deary, I. J. (2016). Predictors of ageing-related decline across multiple cognitive functions, *Intelligence, 59*, 115-126. doi: 10.1016/j.intell.2016.08.007

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. http://dx.doi.org/10.1037/0033-2909.113.3.553

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors, 45(2), 218 - 233*. doi: 10.1518/hfes.45.2.218.27241

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.

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.

Scheibe, S., Sheppes, G., & Staudinger, U. M. (2015). Distract or reappraise? Age-related differences in emotion-regulation choice. *Emotion, 15*, 677–681. doi:10.1037/a0039246

Schönbrodt, F. D., & Wagenmakers, E. J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review, 1-15*.

Shega, J. W., Tiedt, A. D., Grant, K., & Dale, W. (2014). Pain measurement in the National Social Life, Health, and Aging Project: presence, intensity, and location. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 69,* S191-S197.

Simonsohn, U. (2015). Small Telescopes Detectability and the Evaluation of Replication Results. *Psychological Science, 26(5),* 559–569. https://doi.org/10.1177/0956797614567341

Spaniol, J., Schain, C., & Bowen, H. J. (2013). Reward-enhanced memory in younger and older adults. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 69(5),* 730-740.

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. https://doi.org/10.1016/j.cogpsych.2009.12.001

Westerhof, G. J., Bohlmeijer, E. T., & McAdams, D. P. (2017). The relation of ego integrity and despair to personality traits and mental health. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 72(3),* 400-407.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61(8),* 1340–1341.

Figures

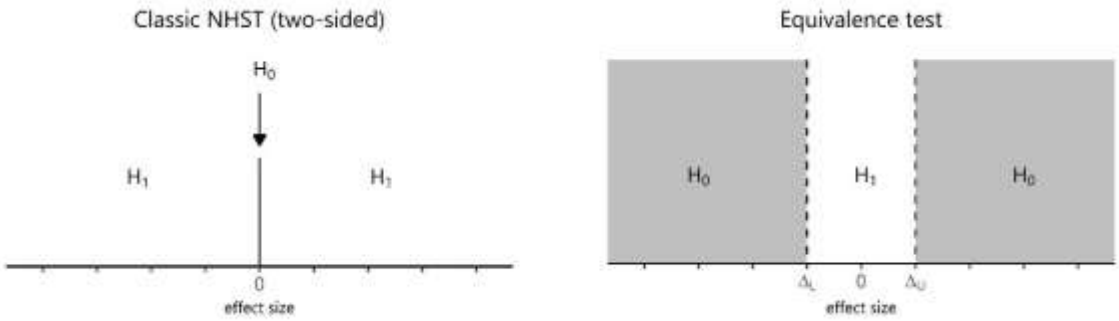

Figure 1. Visualization of null hypothesis (H0) and alternative hypothesis (H1) for a null-hypothesis significance test (left), which tests whether the hypothesis that an effect is equal to 0 can be rejected, and for an equivalence test (right), which tests whether the hypothesis that an effect as extreme as or more extreme than Δ_L or Δ_U can be rejected.

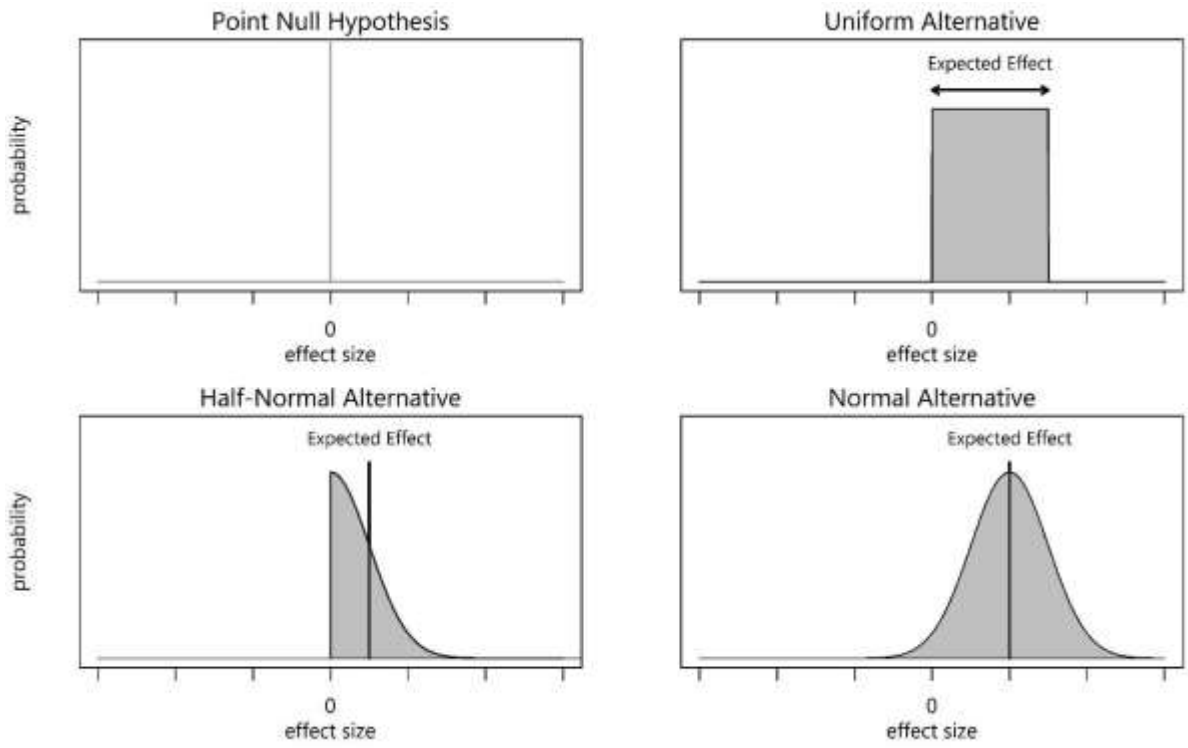

Figure 2. Commonly used distributions to model hypotheses for Bayes factors. *Top Left.* The point null hypothesis predicts that 0 is the only plausible value. *Top Right.* The uniform distribution models all values within an interval as equally plausible. Values outside of the interval are considered not considered possible. *Bottom Left.* The half-normal models a directional prediction where smaller values are more plausible than larger values. *Bottom Right.* The full normal models the expected value as the most plausible value, with effects in either direction considered increasingly less plausible.

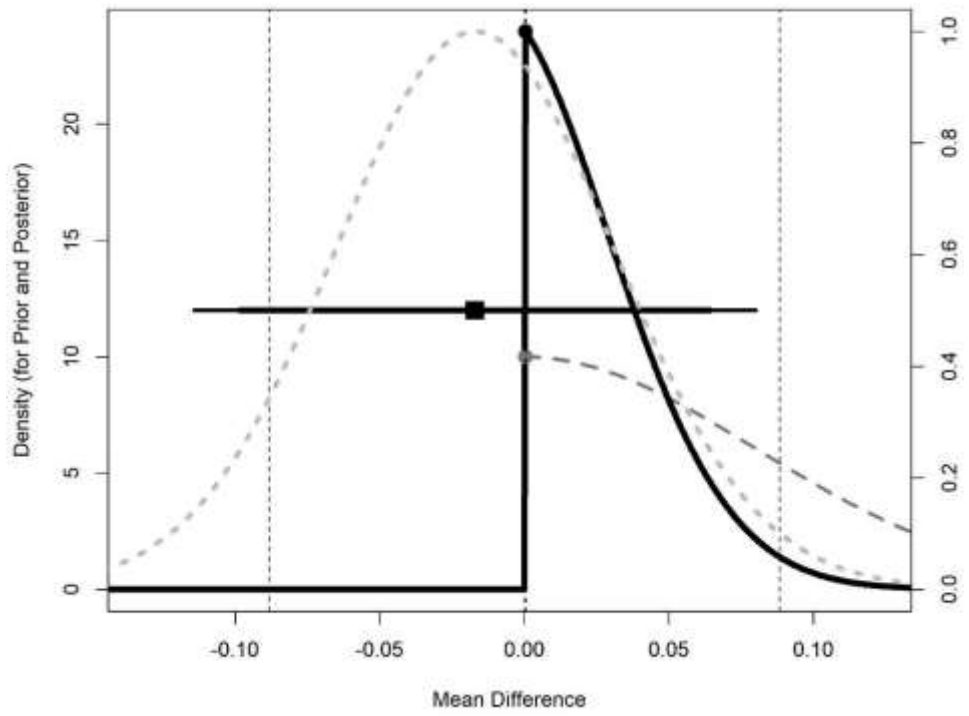

Figure 3. The results of Martins et al. (2016), Example 1. The black square indicates the observed mean difference of 0.02 (on a raw scale). The data is represented by the likelihood distribution (dotted grey line) which is always centered on the observed mean (black square). The dark-grey dashed line indicates the half-normal model of the alternative, and the solid black line visualizes how that model would be updated in light of the data (the posterior distribution). The vertical dashed lines at -0.088 and 0.088 are the equivalence bounds (on a raw scale). The 90% confidence interval (the thick black horizontal line) indicates that the smallest effect size of interest cannot be rejected (it overlaps with the equivalence bound of -0.088). The 95% confidence interval (thin horizontal black line) overlaps with zero, which indicates the null-hypothesis test can not reject an effect of zero.

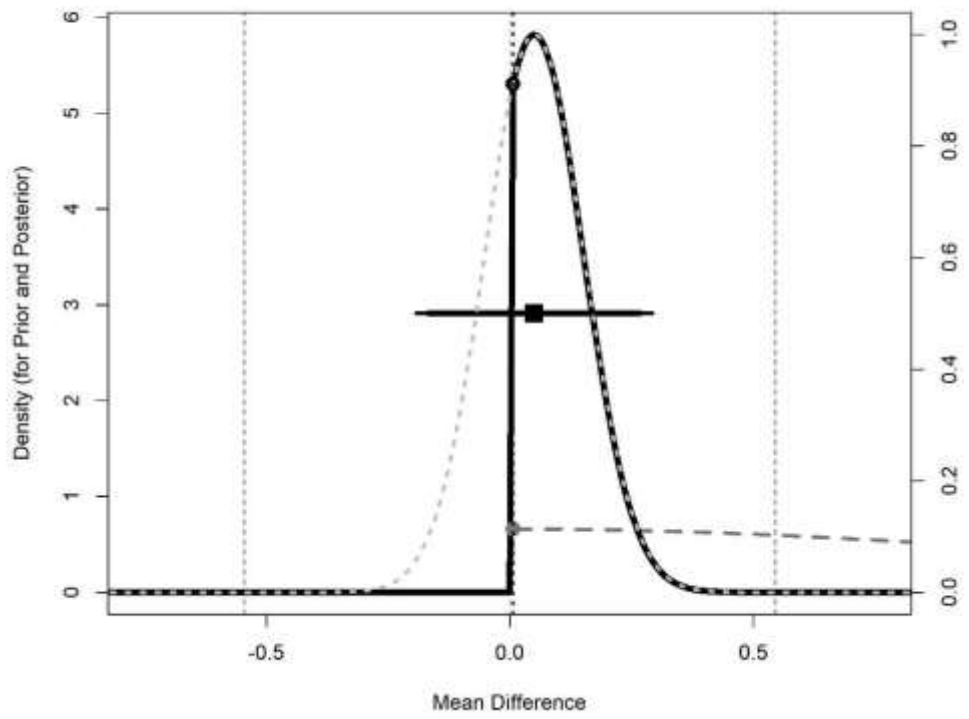

Figure 4. The results of Shega et al. (2014), Example 2. The 96.67% confidence interval (thick horizontal line, Bonferroni corrected for three comparisons) are within the equivalence bounds. The vague alternative model (dashed grey line) spreads the prior probability over the full range of the scale. Given the vague prior distribution, and the large amount of data, we see the posterior (solid black line) overlaps almost perfectly with the likelihood curve (dashed light-grey line) based only on the observed data. Note that in Bayesian estimation approaches the entire posterior distribution is used to draw inferences (for an introduction, see Kruschke, 2018).

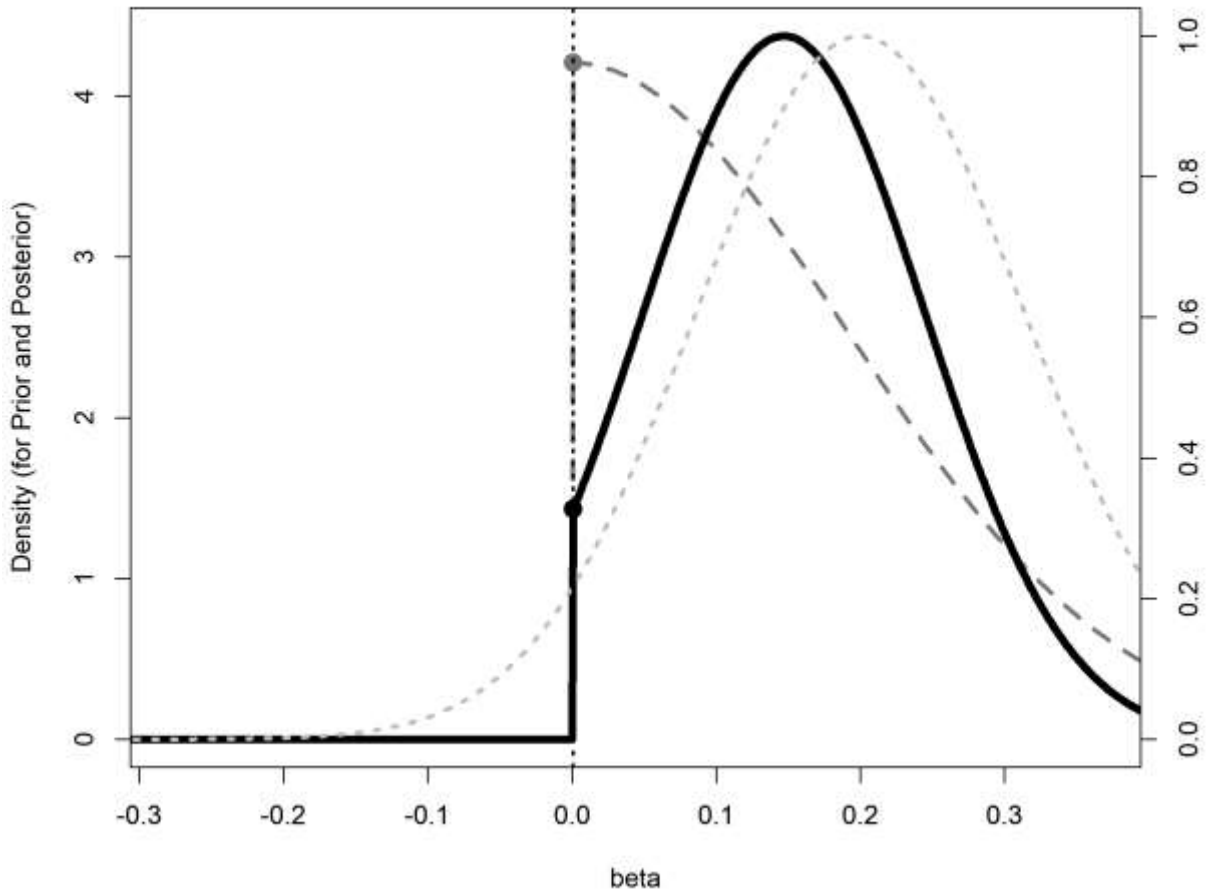

Figure 5. The results of Westerhof et al. (2015), Example 3. The Bayes factor of 2.98 equals the ratio of the density of the prior distribution at zero (dark grey dot) and the posterior distribution at zero (black dot), which is known as the Savage–Dickey density ratio method (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Because the Bayes factor is calculated in raw units (beta), and the equivalence test is performed on the correlation (*r*), the TOST results are not included in Figure 5.

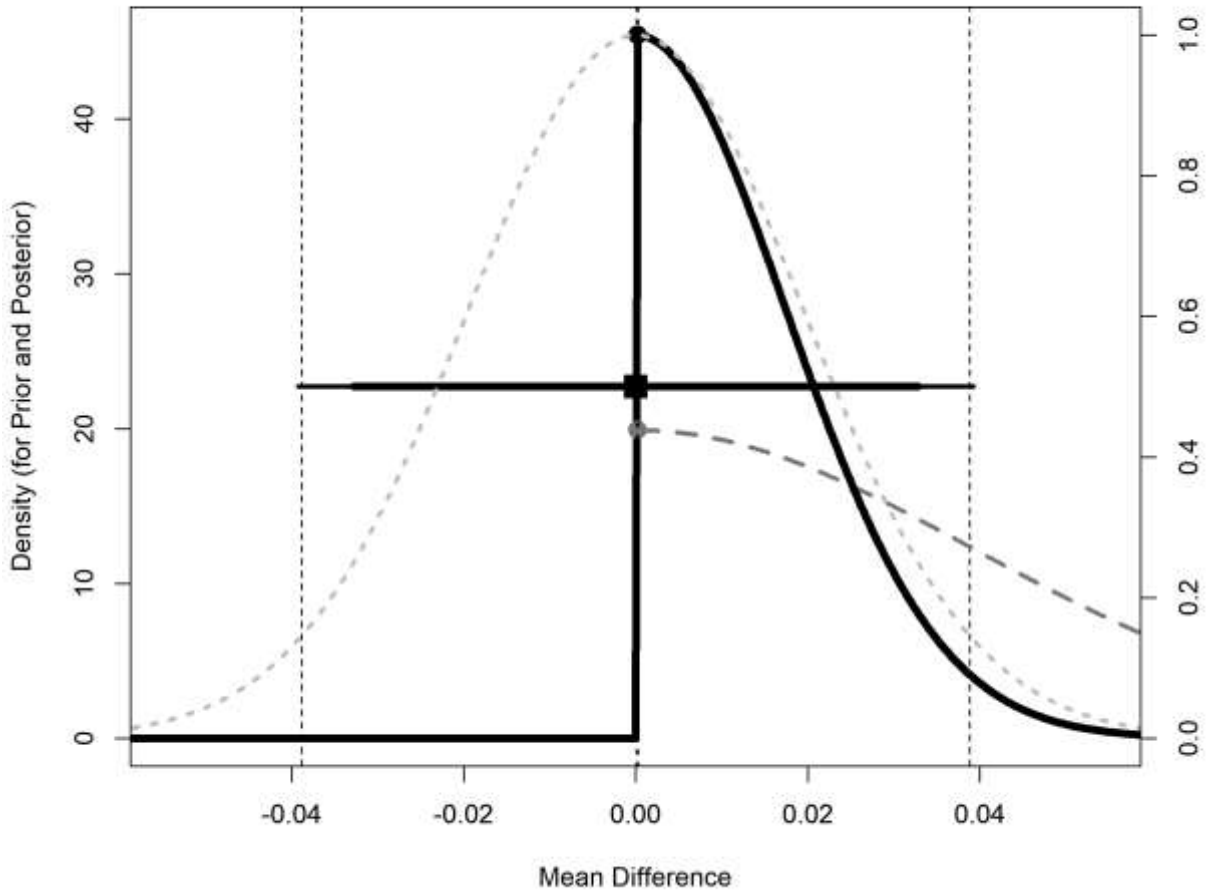

Figure 6. The results of Spaniol et al. (2013), Example 4. The 90% confidence interval (thick horizontal line) falls within the equivalence bounds (vertical dashed lines). The prior model (dark- grey dashed line) is specified with a half-normal distribution. The model of the data is represented by the likelihood distribution (dotted grey line). The posterior distribution (solid black line) is the updated estimate of the population effect size based on the prior and collected data.