


**Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the  
Bayes Factor**


Maximilian Linde<sup>1</sup>, Jorge N. Tendeiro<sup>1</sup>, Ravi Selker<sup>2</sup>, Eric-Jan Wagenmakers<sup>2</sup>, and Don  
van Ravenzwaaij<sup>1</sup>


<sup>1</sup>Unit of Psychometrics and Statistics, Department of Psychology, Faculty of Behavioural  
and Social Sciences, University of Groningen, Groningen, The Netherlands


<sup>2</sup>Unit of Psychological Methods, Department of Psychology, Faculty of Social and  
Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands

**Author Note**

Maximilian Linde  <https://orcid.org/0000-0001-8421-090X>

Jorge N. Tendeiro  <https://orcid.org/0000-0003-1660-3642>

Eric-Jan Wagenmakers  <https://orcid.org/0000-0003-1596-1034>

Don van Ravenzwaaij  <https://orcid.org/0000-0002-5030-4091>

This research was supported by a Dutch scientific organization VIDI fellowship grant awarded to Don van Ravenzwaaij (016.Vidi.188.001) and a Dutch scientific organization VICI fellowship grant awarded to Eric-Jan Wagenmakers (016.Vici.170.083). Correspondence concerning this article should be addressed to: Maximilian Linde, University of Groningen, Department of Psychology, Grote Kruisstraat 2/1, Heymans Building, room 217, 9712 TS Groningen, The Netherlands, Phone: (+31) 50 363 2702, E-mail: m.linde@rug.nl. Preliminary results of this research were presented at the Mathematical Psychology conference in 2019 in Montreal. In addition, this research is available as a pre-print (<https://psyarxiv.com/bh8vu>) and in a modified version as a blog post (<https://www.bayesianspectacles.org/preprint-decisions-about-equivalence-a-comparison-of-tost-hdi-rope-and-the-bayes-factor/>).

### Abstract

Some important research questions require the ability to find evidence for two conditions being practically equivalent. This is impossible to accomplish within the traditional frequentist null hypothesis significance testing framework; hence, other methodologies must be utilized. We explain and illustrate three approaches for finding evidence for equivalence: The frequentist two one-sided tests procedure, the Bayesian highest density interval region of practical equivalence procedure, and the Bayes factor interval null procedure. We compare the classification performances of these three approaches for various plausible scenarios. The results indicate that the Bayes factor interval null approach compares favorably to the other two approaches in terms of statistical power. Critically, compared to the Bayes factor interval null procedure, the two one-sided tests and the highest density interval region of practical equivalence procedures have limited discrimination capabilities when the sample size is relatively small: specifically, in order to be practically useful, these two methods generally require over 250 cases within each condition when rather large equivalence margins of approximately 0.2 or 0.3 are used; for smaller equivalence margins even more cases are required. Because of these results, we recommend that researchers rely more on the Bayes factor interval null approach for quantifying evidence for equivalence, especially for studies that are constrained on sample size.

*Keywords:* equivalence testing, two one-sided tests, highest density interval, region of practical equivalence, interval Bayes factor

## **Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor**

### **Introduction**

Most research aims to demonstrate the presence of an effect. For instance, a study might investigate whether a certain drug is more effective than a placebo. At other times, however, the goal is to find evidence for the equivalence of two conditions. Quantifying evidence that there is no effect can be useful in various applied and theoretical domains (Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). For example, providing evidence that internet-delivered psychotherapies are equally effective as standard face-to-face psychotherapy would benefit patients and clinicians because internet-delivered psychotherapies are comparatively cheap and immediately accessible at any time (see, e.g., Cuijpers, Van Straten, & Andersson, 2008, for a review). Alternatively, having evidence for the absence of an effect might be useful to refute a certain aspect of a theory, thereby advancing and refining that theory. Borrowing an example from Rouder et al. (2009), if the Weber-Fechner law (Fechner, 1860/1966, i.e., people's ability to detect a briefly presented stimulus is a function of the ratio of the intensity of the stimulus and the background intensity) holds, people's ability to detect the stimulus should not change when the two quantities of the ratio are multiplied by the same value.

The statistical framework that is typically employed for statistical inference, null hypothesis significance testing (NHST), merely enables researchers to find evidence against but not in favor of the null hypothesis (e.g., Gallistel, 2009; Goodman, 2008; van Ravenzwaaij, Monden, Tendeiro, & Ioannidis, 2019; Wagenmakers, 2007; Wagenmakers et al., 2018). Fortunately, there are alternatives to traditional NHST that allow researchers to quantify evidence in favor of the absence of an effect, which is the topic of this manuscript.

In frequentist statistical hypothesis testing, a decision is made about whether to reject or not to reject the null hypothesis for a given set of null and alternative hypotheses

(i.e.,  $\mathcal{H}_0$  and  $\mathcal{H}_1$ , respectively) and a given decision threshold. Depending on the decision and the true state of the hypotheses, four scenarios are possible: (1) Not rejecting  $\mathcal{H}_0$  when  $\mathcal{H}_0$  is in fact true, which is a correct decision and is called a true negative; (2) rejecting  $\mathcal{H}_0$  when  $\mathcal{H}_0$  is in fact true, which is an incorrect decision and is called a Type I error or a false positive; (3) not rejecting  $\mathcal{H}_0$  when  $\mathcal{H}_1$  is in fact true, which is an incorrect decision and is called a Type II error or a false negative; (4) rejecting  $\mathcal{H}_0$  when  $\mathcal{H}_1$  is in fact true, which is a correct decision and is called power or a true positive. Thus, two types of errors are possible. Under traditional NHST, the Type I error rate can be controlled by specifying the significance level ( $\alpha$ ) and the type II error rate ( $\beta$ ) can be minimized by increasing power because they are complements of each other (power =  $1 - \beta$ ).

Traditionally, there has been a strong focus on controlling the Type I error rate. This is relatively easy because one can set the desired  $\alpha$ . In contrast, pre-determining power is not straightforward because many factors must be considered. For example, when comparing two conditions, power depends on the sample size,  $\alpha$ , and the unknown population effect size (Cohen, 1988). Power considerations have been mostly neglected. Many studies in the behavioral, social, and medical sciences are under-powered (see, e.g., Button et al., 2013; Halpern, Karlawish, & Berlin, 2002; Ioannidis, 2005; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989; Vankov, Bowers, & Munafò, 2014). In general, it is important to evaluate statistical testing approaches based on both types of errors.

In this article, we first describe and subsequently compare three approaches for finding evidence for equivalence between two groups or, in other words, for finding evidence towards the absence of an effect: the frequentist two one-sided tests procedure (Hodges & Lehmann, 1954; Schuirmann, 1987; Westlake, 1976; see also Lakens, 2017; Lakens, Scheel, & Isager, 2018), the Bayesian highest density interval region of practical equivalence procedure (Kruschke, 2010, 2011, 2013, 2015, 2018; Kruschke, Aguinis, & Joo, 2012; Kruschke & Liddell, 2018a, 2018b), and the Bayes factor interval null procedure (Morey & Rouder, 2011; van Ravenzwaaij et al., 2019; see also Linde & van Ravenzwaaij, 2019a).

Henceforth, we refer to these three procedures as the TOST, HDI-ROPE, and BF procedures, respectively. In the remainder of this article, we will first describe and illustrate each of these procedures with the aid of summary statistics from an existing study by Steiner et al. (2015), an example borrowed from van Ravenzwaaij et al. (2019). Subsequently, we compare each of the three metrics in terms of their probability of concluding equivalence under conditions where the population means of two groups are practically equivalent (power) and under conditions where the population means of two groups are non-equivalent (Type I error rate). We end with a discussion of the implications of our findings and with practical recommendations for researchers.

### Running Example

In order to illustrate the three approaches for finding evidence towards equivalence, we utilize summary statistics from a study by Steiner et al. (2015); this example is borrowed from van Ravenzwaaij et al. (2019). The goal of this study was to examine whether or not storing red-cells for a long duration, in comparison to a short duration, is harmful for patients who receive a transfusion. A comparison was made between red-cells that are stored for either  $< 10$  days or  $> 21$  days. The dependent variable was the Multiple Organ Dysfunction Score (MODS), which was measured in patients 7 days after transfusion. The relevant summary statistics are provided in Table 2 of Steiner et al. (2015); more precise values are presented in van Ravenzwaaij et al. (2019). The  $< 10$ -day group (control,  $c$ ) consisted of  $n_c = 538$  cases, had a mean of  $\bar{x}_c = 8.516$ , and a standard deviation of  $s_c = 3.6$ . The  $> 21$ -day group (experimental,  $e$ ) consisted of  $n_e = 560$  cases, had a mean of  $\bar{x}_e = 8.683$ , and a standard deviation of  $s_e = 3.6$ .

### Frequentist Equivalence Testing

The TOST approach (Hodges & Lehmann, 1954; Schuirman, 1987; Westlake, 1976; see also Lakens, 2017; Lakens, Scheel, & Isager, 2018) tests whether or not the control and experimental conditions are practically equivalent on some measure (see, e.g., Christensen, 2007; Lesaffre, 2008; Meyners, 2012; Walker & Nowacki, 2011, for an overview). An

equivalence interval must be defined independent of the data (Meyners, 2012), which should include all values that are deemed small enough to be practically equivalent to no effect and, thus, constitute no meaningful effect (e.g., Meyners, 2012; Walker & Nowacki, 2011). The equivalence interval can be symmetric around 0 (e.g.,  $(-0.1, 0.1)$ ) or asymmetric around 0 (e.g.,  $(-0.2, 0.1)$ ). Furthermore, the equivalence interval can be expressed in standardized or unstandardized units. Henceforth and for all three approaches for finding evidence for equivalence, we will assume that the equivalence interval is symmetric around 0 with a standardized equivalence margin denoted by  $m$ , which is defined as the distance from the null value to the lower or upper boundary of the equivalence interval.

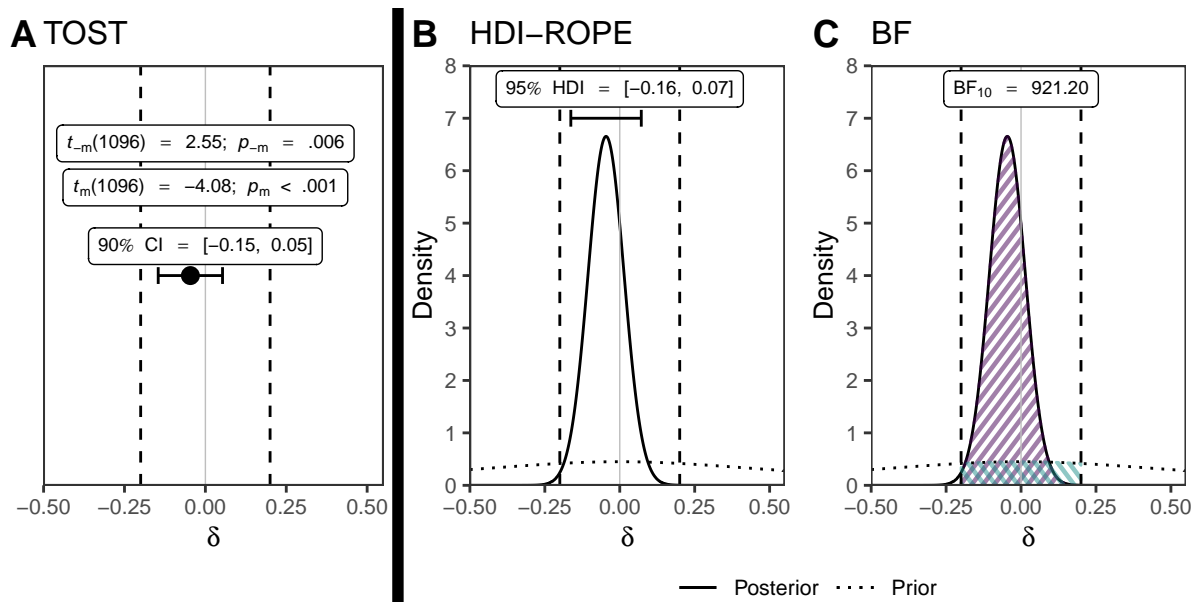
TOST entails conducting two one-sided tests (we refer the interested reader to Appendix A for a more detailed and mathematical description of TOST).  $\mathcal{H}_0$  and  $\mathcal{H}_1$  of the two one-sided tests combined are:

$$\mathcal{H}_0: \delta \leq -m \text{ OR } \delta \geq m \qquad \mathcal{H}_1: \delta > -m \text{ AND } \delta < m, \qquad (1)$$

where  $\delta$  is the population effect size between both groups. In words, we test whether  $\delta$  is significantly larger than  $-m$  and we test whether  $\delta$  is significantly smaller than  $m$ . If and only if both tests are statistically significant, we reject  $\mathcal{H}_0$  that the two conditions are non-equivalent (see, e.g., Meyners, 2012; Walker & Nowacki, 2011).

A perhaps more intuitive way to think of the TOST procedure is by means of a confidence interval (CI; cf. Meyners, 2012; Schuirmann, 1987). We reject  $\mathcal{H}_0$  at significance level  $\alpha$  if the  $(1 - 2\alpha)$  100% CI of  $\delta$  fully lies within the equivalence interval. So, for the typical  $\alpha = .05$ , the 90% CI needs to be fully contained within the equivalence interval. Note that it is not required for the CI to overlap with 0 for purposes of deciding in favor of equivalence (see van Ravenzwaaij et al., 2019, for an illustration).

We can use the TOST procedure to analyze the data from our running example (Steiner et al., 2015; van Ravenzwaaij et al., 2019). We use a significance level of  $\alpha = .05$  for this analysis. Panel A in Figure 1 shows the 90% CI of  $\delta$ . An equivalence margin of  $m = 0.2$  is chosen. It is clearly visible that the 90% CI is fully contained within the



**Figure 1**

*Illustration of (A) the frequentist TOST procedure; (B) the Bayesian HDI-ROPE procedure; and (C) the Bayesian BF procedure, using summary statistics provided in Table 2 of Steiner et al. (2015). See text for details.*

equivalence interval. The lower boundary of the 90% CI is higher than the lower boundary of the equivalence interval and the upper boundary of the 90% CI is lower than the upper boundary of the equivalence interval. Calculating the two one-sided  $t$ -tests results in  $p$ -values of  $p_{-m} = .006$  and  $p_m < .001$ . Because both  $p_{-m}$  and  $p_m$  are smaller than  $\alpha = .05$ , we reject  $\mathcal{H}_0$  of non-equivalence. The conclusion depends on the chosen equivalence margin. For example, with  $m = 0.1$  we could not reject  $\mathcal{H}_0$  of non-equivalence.

## Bayesian Techniques for Quantifying Evidence Towards Equivalence

### *Bayesian Inference*

Bayesian inference is an alternative to the frequentist school. It provides a framework for coherently updating beliefs about parameters and hypotheses based on new data (Kruschke, 2015). Our exposition of Bayesian inference will be kept to a minimum; we refer the reader to other sources for more detailed and technical introductions to Bayesian



inference (see, e.g., Dienes & Mclatchie, 2018; Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2018; Etz & Vandekerckhove, 2018; Gelman et al., 2013; Kass & Raftery, 1995; Kruschke, 2015; Kruschke & Liddell, 2018a, 2018b; O’Hagan & Forster, 2004; Rouder, Haaf, & Vandekerckhove, 2018; Rouder et al., 2009; Wagenmakers et al., 2018) and Bayesian approaches for quantifying evidence for equivalence specifically (see, e.g., Kruschke, 2011, 2018; Morey & Rouder, 2011; van Ravenzwaaij et al., 2019; see also Linde & van Ravenzwaaij, 2019a).

We can distinguish between two approaches to Bayesian inference: parameter estimation (e.g., reporting the posterior distributions for regression coefficients) and model comparison (e.g., comparing the relative likelihoods of the data under  $\mathcal{H}_0$  and  $\mathcal{H}_1$ ). At the heart of both of these approaches to Bayesian inference is Bayes’ rule, which stipulates how relevant data  $y$  refine beliefs about a parameter  $\delta$  under hypothesis  $\mathcal{H}_i$ . Bayes’ rule is:

$$\underbrace{p(\delta | y, \mathcal{H}_i)}_{\text{Posterior}} = \frac{\overbrace{p(y | \delta, \mathcal{H}_i)}^{\text{Likelihood}}}{\underbrace{p(y | \mathcal{H}_i)}_{\text{Evidence}}} \times \underbrace{p(\delta | \mathcal{H}_i)}_{\text{Prior}}, \quad (2)$$

meaning that the posterior is equal to the product of the prior and an updating factor, which is the ratio of the likelihood and the evidence. We will now explain the individual factors in Bayes’ rule (Equation 2) and how they combine to update probabilities in the context of equivalence designs.

Before we see any data  $y$ , we have beliefs about the plausibility of certain parameter values for  $\delta$ .<sup>1</sup> These beliefs can be expressed by placing a prior on  $\delta$ ,  $p(\delta | \mathcal{H}_i)$ . For example, if  $\delta$  is an effect size parameter and we expect a very small effect size, we can restrict the prior to very small values for  $\delta$ ; alternatively, if only relatively large effect sizes matter and are therefore of interest, we can place a higher emphasis on larger values for  $\delta$ . The likelihood is a conditional probability of the data  $y$ , given a specific value of  $\delta$ ,

<sup>1</sup> The full model of equivalence designs for  $t$ -tests actually includes two parameters,  $\delta$  and  $\sigma^2$  (common population variance), but for purposes of this demonstration a consideration of  $\delta$  alone suffices.

$p(y \mid \delta, \mathcal{H}_i)$ . The posterior,  $p(\delta \mid y, \mathcal{H}_i)$ , represents a compromise between the plausibility of parameter realizations before taking into account the observed data, the prior, and what the data tell us about the plausibility of parameter realizations, the likelihood. The evidence,  $p(y \mid \mathcal{H}_i)$ , is the probability of the data. It serves as a normalization constant, ensuring that the posterior is a proper probability density function. However, removing the evidence from Equation 2 would not alter the shape of the posterior. Therefore, in that case the posterior would be proportional to the product of the prior and the likelihood.

**Choosing a Prior.** The posterior is more influenced by the prior when data are sparse or when the prior is peaked and local; it is more influenced by the likelihood when data are abundant or when the prior is spread out and global (see, e.g., Kruschke, 2015). As such, the shape of the posterior heavily depends on our prior beliefs. There is an ongoing debate about whether data analysts should use ‘subjective’ or ‘objective’ priors. Proponents of the former advocate that subjective beliefs should be reflected in the prior, as this is exactly what Bayesianism is all about (Goldstein, 2006; Morey, Romeijn, & Rouder, 2016; Vanpaemel, 2010). In contrast, proponents of the latter suggest that Bayesian analyses should be as objective as possible by using objective priors (also called default priors) and that their use is sometimes inevitable when prior information is lacking (Bayarri, Berger, Forte, & García-Donato, 2012; Berger, 2006; Consonni, Fouskakis, Liseo, & Ntzoufras, 2018; Jeffreys, 1961; Rouder et al., 2009). We refer the readers to other sources for discussions on this topic (e.g., Berger, 2006; Berger & Berry, 1988; Consonni et al., 2018; Gelman & Hennig, 2017; Goldstein, 2006; Tendeiro & Kiers, 2019; Vanpaemel, 2010).

For all our Bayesian analyses described in this manuscript, we employ objective priors as described in Rouder et al. (2009), Gronau, Ly, and Wagenmakers (2020), and Morey and Rouder (2011). Specifically, we use a Cauchy prior centered on 0 and with a variable scale parameter  $r$  for  $\delta$ .<sup>2</sup> A Cauchy distribution is a  $t$  distribution with 1 degree of

---

<sup>2</sup> We place a Jeffreys’ prior (also called right Haar prior) on  $\sigma^2$ ,  $p(\sigma^2) \propto (\sigma^2)^{-1}$ .

freedom; it resembles a Normal distribution but it has less mass at the center and more mass at the tails (see Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder et al., 2009). The scale parameter  $r$  defines the width of the Cauchy distribution; that is, half of the mass lies between  $-r$  and  $r$ . Note that we use the same prior for the HDI-ROPE and the BF approaches to allow for a direct and unbiased comparison. We decided to use a Cauchy prior because it has several desirable mathematical properties in the context of calculating Bayes factors, which are described elsewhere (see, e.g., Bayarri et al., 2012; Consonni et al., 2018), and because it allows for a closed-form solution to calculate the HDI within the HDI-ROPE approach and the Bayes factor within the BF approach (see Gronau et al., 2020; Rouder et al., 2009).

### ***The HDI-ROPE Approach***

The HDI-ROPE approach (e.g., Kruschke, 2011, 2018) is one approach for making decisions about equivalence within the Bayesian paradigm that falls under the rubric of parameter estimation. As a first step, a region of practical equivalence (ROPE) is defined, whose values are deemed practically equivalent to the null value (Kruschke, 2015, 2018). We use Bayes' theorem (see Equation 2 and the previous section) to obtain the posterior distribution for  $\delta$ . Next, we estimate the 95% (or any other desired percentage) highest density interval (HDI) of this posterior. The probability mass of that interval is .95, given our prior and model, and the density at each point within the interval is higher than the density at any point outside the interval.

Having defined the ROPE and having calculated the 95% HDI for  $\delta$ , a decision can be made about whether the two conditions are equivalent. The HDI-ROPE decision rule results in one of three decisions (see Figure 1 in Kruschke, 2018, for an illustration): (1) If the HDI lies completely outside of the ROPE, we reject the null value and conclude there is no equivalence; (2) if the HDI lies completely within the ROPE, we accept the null value and conclude there is equivalence; (3) if the HDI overlaps with at least one of the boundaries of the ROPE, we can neither accept nor reject the null value and conclude

nothing. Note that the decision does not concern all the parameter values within the ROPE, but just the null value, which is the center of the ROPE (Kruschke, 2018). The ROPE defines an interval of parameter values that are considered practically equivalent to the null value. The analyst should reject the null value when the most credible values (i.e., HDI) are *sufficiently far away* (outside the ROPE) from the null value; the analyst should accept the null value when the most credible values (i.e., HDI) are *sufficiently close* (inside the ROPE) to the null value (Kruschke, 2018). The second scenario described above has a configuration which seems counter-intuitive at first glance. If the HDI fully lies within the ROPE but does not overlap with the null value, we still accept the null value for practical purposes (Kruschke, 2018).

The HDI-ROPE procedure can be used to analyze the data from our running example. Panel B in Figure 1 shows the prior (dotted line) and the posterior (solid line) for  $\delta$ . The two vertical dashed lines represent the standardized ROPE (i.e., the standardized equivalence interval,  $(-0.2, 0.2)$ ). The 95% HDI of the posterior of  $\delta$  is  $(-0.16, 0.07)$ . This can be interpreted as a probability of .95 that the population effect size  $\delta$  is somewhere between  $-0.16$  and  $0.07$ , conditional on the model and prior used in the estimation process. The HDI is fully contained inside the ROPE. Therefore, we decide that the two conditions are practically equivalent. Here again, the decision depends on the equivalence margin. Had we used  $m = 0.1$ , we could not make any decision because the HDI would be neither fully inside nor outside the equivalence interval. In practice, the obtained boundaries of the HDI will often be very similar to those of the frequentist confidence interval (see, e.g., Albers, Kiers, & van Ravenzwaaij, 2018). Hence, the HDI-ROPE approach with objective priors could be considered the Bayesian equivalent of the TOST approach.

### ***The BF Approach***

The Bayes factor (e.g., Jeffreys, 1939, 1948, 1961; Kass & Raftery, 1995) is used to compare two competing models or hypotheses. Hence, it falls under the category of model comparison approaches. Like the HDI in the HDI-ROPE approach, a Bayes factor is based

on Bayes' rule (see Equation 2). However, for the calculation of Bayes factors involving interval hypotheses, Bayes' rule is utilized in a hierarchical structure and, thus, needs to be adapted. If we replace parameter  $\delta$  in Equation 2 with model  $\mathcal{H}_1$  or model  $\mathcal{H}_0$  and divide the expressions for both models by one another, we obtain:

$$\underbrace{\frac{p(\mathcal{H}_1 | y)}{p(\mathcal{H}_0 | y)}}_{\text{Posterior Odds}} = \underbrace{\frac{p(y | \mathcal{H}_1)}{p(y | \mathcal{H}_0)}}_{\text{Bayes Factor, BF}_{10}} \times \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\text{Prior Odds}}. \quad (3)$$

After taking into account the data  $y$ , the Bayes factor is the factor by which we update the prior odds to obtain the posterior odds. For example, if  $\mathcal{H}_1$  was 3 times more probable than  $\mathcal{H}_0$  a priori and the Bayes factor is  $\text{BF}_{10} = 7$  (the subscript indicates that  $\mathcal{H}_1$  is in the numerator and  $\mathcal{H}_0$  in the denominator), the posterior odds would be  $3 \times 7 = 21$ . In this example, we can conclude that the data are seven times more likely under  $\mathcal{H}_1$  compared to  $\mathcal{H}_0$ . If the prior odds is 1, the Bayes factor is equal to the posterior odds. Given a specific Bayes factor (e.g., reported in an article), researchers can specify their own prior odds and use the Bayes factor to arrive at their own posterior odds.

The Bayes factor that we use here for equivalence designs corresponds to the non-overlapping hypotheses (NOH) Bayes factor approach, as described in Morey and Rouder (2011). Importantly, the Cauchy prior for  $\delta$  is split into two parts: The first part consists of all the values of the prior as defined by  $\mathcal{H}_1$ , where  $\delta$  lies between  $-m$  and  $m$ ; the second part consists of all the values of the prior as defined by  $\mathcal{H}_0$ , where  $\delta$  does not lie between  $-m$  and  $m$  (see Panel C of Figure 1). When we know the prior and posterior odds, we can rearrange Equation 3 to obtain the Bayes factor:

$$\text{BF}_{10} = \frac{p(\mathcal{H}_1 | y) / p(\mathcal{H}_1)}{p(\mathcal{H}_0 | y) / p(\mathcal{H}_0)}. \quad (4)$$

In words, we obtain the posterior odds as the area of the posterior inside the equivalence interval over the area of the posterior outside the equivalence interval. Similarly, we obtain the prior odds as the area of the prior inside the equivalence interval over the area of the prior outside the equivalence interval. Finally, we divide the posterior odds by the prior odds to obtain the Bayes factor.

The BF approach can be used to analyze the data from our running example. Panel C of Figure 1 shows the prior (dotted line) and the posterior (solid line) for  $\delta$ . The two vertical dashed lines represent the standardized equivalence interval  $(-0.2, 0.2)$ . Panel C of Figure 1 can be interpreted as a visualization of Equation 4 because all four factors of this equation are present in the figure. The posterior probability of  $\mathcal{H}_1$ ,  $p(\mathcal{H}_1 | y) \approx 0.99$ , is colored in purple; the posterior probability of  $\mathcal{H}_0$ ,  $p(\mathcal{H}_0 | y) \approx 0.01$ , is the remaining area of the posterior. The prior probability of  $\mathcal{H}_1$ ,  $p(\mathcal{H}_1) \approx 0.18$ , is shown in blue, whereas the prior probability of  $\mathcal{H}_0$ ,  $p(\mathcal{H}_0) \approx 0.82$ , is the remaining area of the prior. Using Equation 4 to combine the just described areas, we obtain a Bayes factor of  $\text{BF}_{10} = 921.20$ . According to heuristic evidence thresholds suggested by Jeffreys (1961), with updated labels provided by Lee and Wagenmakers (2013), we obtained extreme evidence in favor of  $\mathcal{H}_1$ , stating that the control and experimental conditions are practically equivalent. If we would have used an equivalence margin of  $m = 0.1$ , we would obtain a Bayes factor of  $\text{BF}_{10} = 43.32$ ; thus, the conclusion would remain unchanged.

In the next section, we compare the TOST, the HDI-ROPE, and the BF procedures in terms of their classification rates of equivalence, both in situations where the two population group means are practically equivalent and in situations where they are not. In order to be able to compare the performances of the three approaches, it is necessary to adopt a binary decision-making system. The reason is that the TOST approach can only make one of two decisions (“equivalence” or “not enough evidence for equivalence”). Since the HDI-ROPE approach has a tertiary decision-making system by nature (“equivalence”, “non-equivalence”, or “undecided”) and is thus incompatible with the decision-making system of the TOST approach, we adopt a binary decision-making system for the HDI-ROPE approach (“equivalence” or “non-equivalence/undecided”). The Bayes factor quantifies evidence in a continuous fashion. By imposing two thresholds against which the obtained evidence is evaluated, a tertiary decision-making system could be established (“equivalence”, “non-equivalence”, or “undecided”). Just as for HDI-ROPE, we have to

adopt a binary decision-making system to have results comparable to those of the TOST approach; hence, for purposes of this study the decision categories “non-equivalence” and “undecided” are combined (i.e., “equivalence” or “non-equivalence/undecided”).

### Methods

We estimated the proportions of equivalence decisions for the TOST, the HDI-ROPE, and the BF approaches for different plausible scenarios (the R code can be accessed at <https://osf.io/xur3t/>). These decisions could be either correct or incorrect, depending on the specific scenario at hand. A decision in favor of equivalence for truly equivalent conditions is a correct decision, whereas a decision in favor of equivalence for truly non-equivalent conditions is a false decision. More specifically, for the scenario where the population effect size lies within the equivalence interval, the proportion of equivalence decisions is the same as the true positive rate (power); in contrast, for the scenario where the population effect size lies outside of the equivalence interval, the proportion of equivalence decisions corresponds to the false positive rate (Type I error rate). Our scenarios were intended to mimic balanced independent-conditions two-sample experiments with a continuous outcome measure, a common variance in both conditions, and Normal distributed residuals within conditions. As will be explained shortly, the proportion of equivalence decisions can be obtained numerically in the case of the TOST procedure (see Appendix A) but must be estimated through simulations for the HDI-ROPE and the BF procedures. The global parameters, which determine the different scenarios, apply to both the analytical and simulated solutions.

### Global Parameters

We varied three parameters which were fully crossed, thus, establishing different scenarios. Those parameters were the population effect size  $\delta$ , the sample size within each condition  $n$ , and the margin  $m$  (i.e., the distance from the null value to the lower or upper bound of the equivalence interval) of a symmetric equivalence interval on the same standardized scale as  $\delta$ :

- $\delta = \{0, 0.01, 0.02, \dots, 0.5\}$
- $n = \{50, 100, 250, 500\}$
- $m = \{0.1, 0.2, 0.3\}$ .

We decided to use an almost continuous representation of  $\delta$  in order to have a high resolution of the proportion of equivalence predictions as a function of  $\delta$ . Note that we only used values for  $\delta$  in the interval  $[0, \infty)$ ; results should be the same (within random fluctuations) for the negative counterparts of  $\delta$  (i.e., within  $(-\infty, 0]$ ). Moreover, the choice of our standardized equivalence margins rests on rules of thumb for judging the magnitudes of effect sizes, developed by Cohen (1988). According to his rules of thumb, an effect size of  $\delta = 0.2$  is considered small. We also examined equivalence margins of 0.1 lower and higher (i.e., 0.1 and 0.3). Equivalence margins larger than this come close to what Cohen labeled medium-sized effects and are in most contexts unreasonably large to demarcate equivalence.

### **Numerical Approach for TOST**

We calculated the power for TOST for each combination of the global parameters  $\delta$ ,  $n$ , and  $m$ . For this we made use of code, written in R (R Core Team, 2019), provided by Shieh (2016), where an exact solution is offered based on numerical integration. We adapted and improved upon this code in two ways: (1) We applied more accurate numerical integration; (2) We made the code more readable and eliminated defined but unused variables.<sup>3</sup> All power calculations for TOST were conducted with  $\alpha = .05$ .

### **Simulation Approach for HDI-ROPE and BF**

A numerical or even analytical approach for the calculation of power for the HDI-ROPE and the BF procedures is not straightforward, so we conducted analyses on simulated data instead. We generated 1,000 data sets for each individual combination of

---

<sup>3</sup> The interested reader is referred to Appendix A for a detailed treatment and derivation of power for TOST and for the corresponding R code used here; this and other code is also available at <https://osf.io/f9dw2/>.

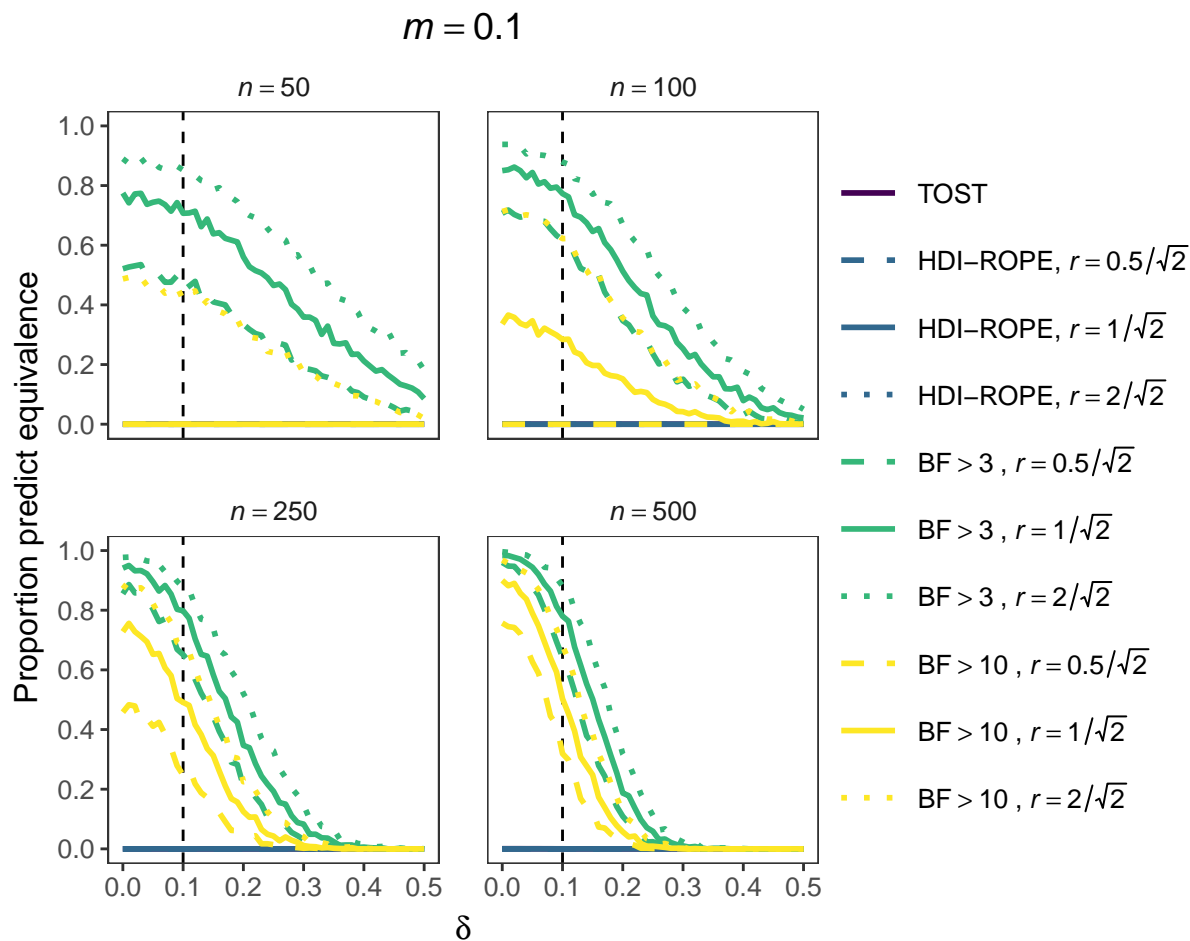


the global parameters  $\delta$ ,  $n$ , and  $m$ . Each data set was analyzed using the HDI-ROPE and BF approaches. Both approaches were applied three times for each data set because we employed three different priors. We used a Cauchy prior centered at 0 and with scale parameters of  $r = \{0.5/\sqrt{2} \approx 0.354, 1/\sqrt{2} \approx 0.707, 2/\sqrt{2} \approx 1.414\}$ . The use of different priors allowed us to examine the variation of outcomes and serves here as a coarse sensitivity analysis (Berger et al., 1994; Du, Edwards, & Zhang, 2019; Kass & Raftery, 1995; Lee & Wagenmakers, 2013). For each set of 1,000 data sets, we calculated the proportions of equivalence predictions for both the HDI-ROPE and the BF procedures.

We used a 95% HDI for the HDI-ROPE procedure. The calculation of the boundaries of the 95% HDI of the posterior was achieved using functions developed by Gronau et al. (2020, available at <https://osf.io/bsp6z/>), which are available for R (R Core Team, 2019). Bayes factors were obtained using the baymedr software (Linde & van Ravenzwaaij, 2019b) written in R (R Core Team, 2019). These Bayes factors were compared to two thresholds,  $\text{BF}_{thr} = \{3, 10\}$ , which follow approximate thresholds for judging the magnitude of the evidence:  $1 < \text{BF} < 3$  for ‘anecdotal evidence’;  $3 < \text{BF} < 10$  for ‘moderate evidence’;  $10 < \text{BF} < 30$  for ‘strong evidence’;  $30 < \text{BF} < 100$  for ‘very strong evidence’;  $\text{BF} > 100$  for ‘extreme evidence’ (cf. Jeffreys, 1961; see also Lee & Wagenmakers, 2013). Importantly, the Bayes factor that was obtained quantifies evidence in favor of  $\mathcal{H}_1$  of equivalence. Thus,  $\text{BF}_{10} > \text{BF}_{thr}$  constitutes evidence in favor of equivalence.

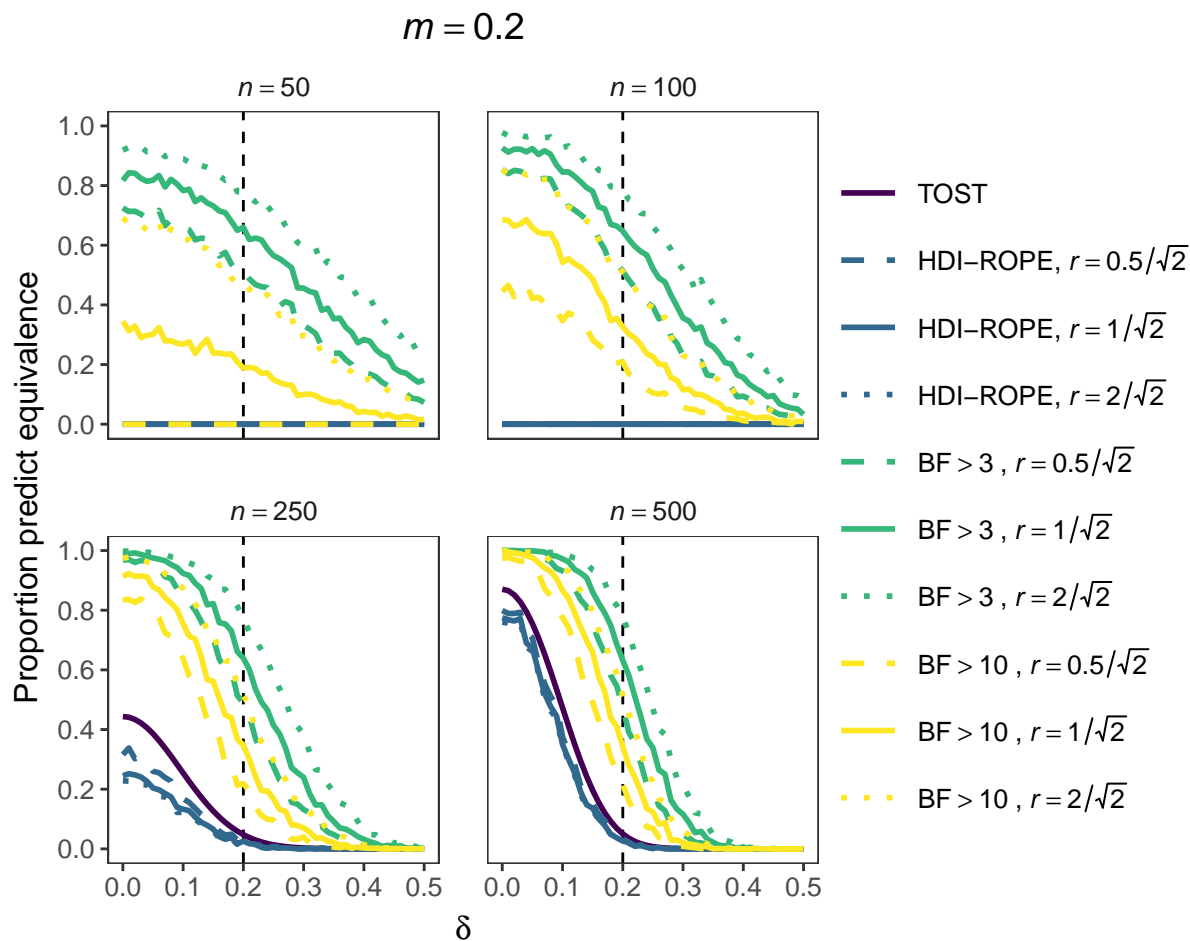
## Results

The results of the comparison between the TOST, HDI-ROPE, and BF approaches for equivalence margins of  $m = 0.1$ ,  $m = 0.2$ , and  $m = 0.3$  are shown in Figures 2, 3, and 4, respectively. Within each figure, the different panels show the results for the various sample sizes ( $n$ ) within conditions. The different colors represent the approaches for finding evidence towards equivalence, and for the BF approach the two different evidence thresholds. Different line types represent the three Cauchy prior scales ( $r$ ) for  $\delta$  within the two Bayesian approaches.

**Figure 2**

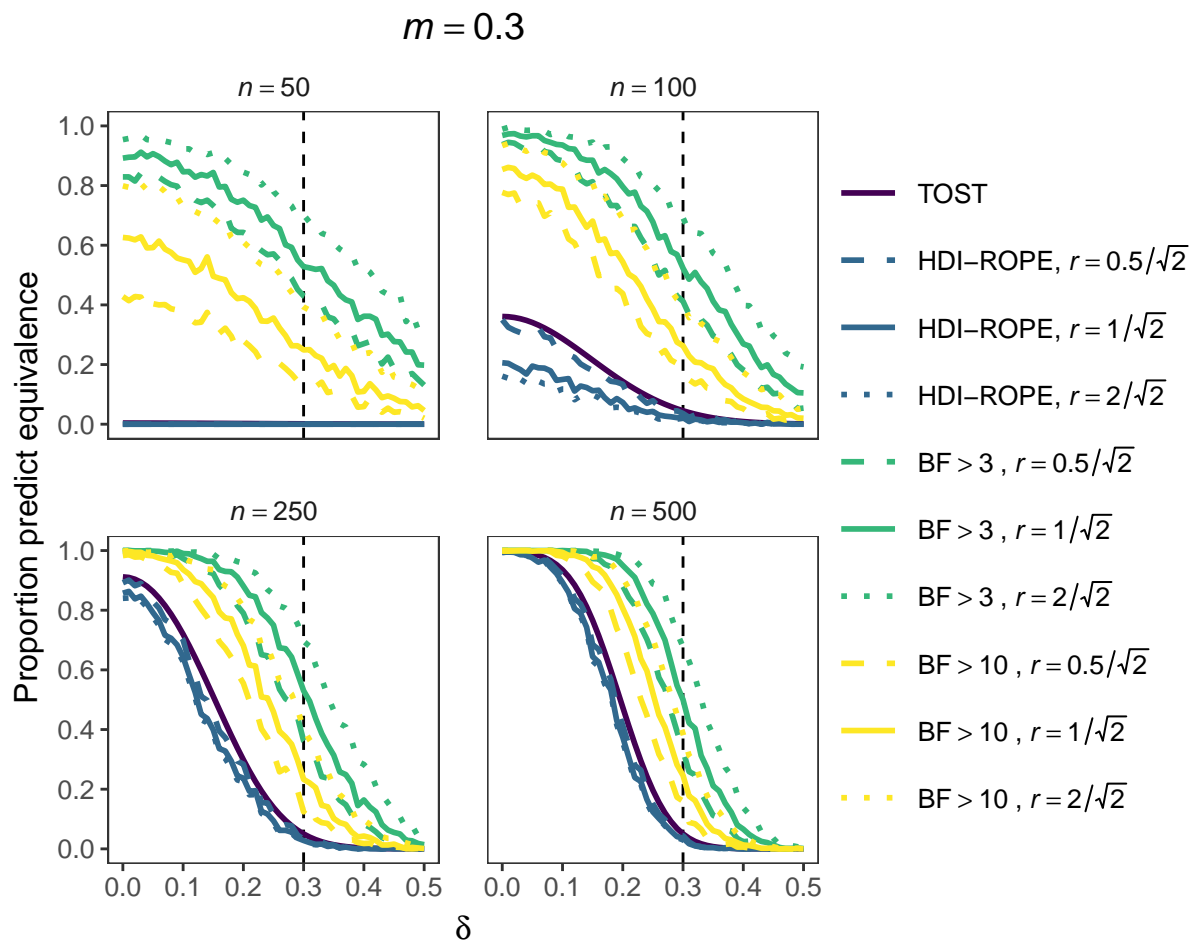
*Proportion of equivalence predictions with a standardized equivalence margin of  $m = 0.1$  (vertical dashed line). Panels contain results for different sample sizes. Colors denote different inferential approaches (and different decision thresholds within the BF approach). Line types denote different priors (for Bayesian metrics). Predictions of equivalence are correct if the population effect size ( $\delta$ ) lies within the equivalence interval (power), whereas predictions of equivalence are incorrect if  $\delta$  lies outside the equivalence interval (Type I error rate). Note that TOST and HDI-ROPE uniformly produce a proportion predicted equivalence of 0 and are thus not independently visible.*

We defined three criteria by which we judged the performance of the classifiers; these criteria are based on signal detection theory (see, e.g., Green & Swets, 1966;

**Figure 3**

*Proportion of equivalence predictions with a standardized equivalence margin of  $m = 0.2$  (vertical dashed line). Panels contain results for different sample sizes. Colors denote different inferential approaches (and different decision thresholds within the BF approach). Line types denote different priors (for Bayesian metrics). Predictions of equivalence are correct if the population effect size ( $\delta$ ) lies within the equivalence interval (power), whereas predictions of equivalence are incorrect if  $\delta$  lies outside the equivalence interval (Type I error rate). Note that TOST and HDI-ROPE for  $n = 50$  and  $n = 100$  uniformly produce a proportion predicted equivalence of 0 and are thus not independently visible.*

Macmillan & Creelman, 1990, 2005; Stanislaw & Todorov, 1999, for an overview): First, a higher proportion of equivalence decisions when conditions are truly equivalent (i.e.,

**Figure 4**

*Proportion of equivalence predictions with a standardized equivalence margin of  $m = 0.3$  (vertical dashed line). Panels contain results for different sample sizes. Colors denote different inferential approaches (and different decision thresholds within the BF approach). Line types denote different priors (for Bayesian metrics). Predictions of equivalence are correct if the population effect size ( $\delta$ ) lies within the equivalence interval (power), whereas predictions of equivalence are incorrect if  $\delta$  lies outside the equivalence interval (Type I error rate). Note that TOST and HDI-ROPE for  $n = 50$  uniformly produce a proportion predicted equivalence of 0 and are thus not independently visible.*

$|\delta| < m$ ) and a lower proportion of equivalence decisions when conditions are truly non-equivalent (i.e.,  $|\delta| > m$ ), with a maximum of equivalence decisions when  $\delta$  is zero (i.e.,

exactly in the middle of the equivalence margins) and a minimum of equivalence decisions when  $\delta$  is large (i.e.,  $\delta = 0.5$ ). In other words, the farther away  $\delta$  is from  $m$ , the clearer should be the decision.

Second, a steeper slope of the curves. A steep slope is desirable, because it indicates that there is only a small region of  $\delta$  for which the decision is ambiguous.

Third, a proportion of equivalence decisions close to 0.5 when  $\delta = m$  because this indicates that the classifier is unbiased. In more detail, if the proportion of equivalence decisions would be smaller than 0.5 at  $\delta = m$ , the curve would be shifted to the left and would thus decide for non-equivalence more often than equivalence (a conservative classifier); if the proportion of equivalence decisions would be higher than 0.5 at  $\delta = m$ , the curve would be shifted to the right and would thus decide for equivalence more often than non-equivalence (a liberal classifier); if the proportion of equivalence decisions would be 0.5 at  $\delta = m$ , the curve would decide for equivalence and non-equivalence equally often (an unbiased classifier). Ignoring potential diverging costs for deciding for equivalence or non-equivalence, an unbiased classifier maximizes accuracy (see Hahn & Harris, 2014; and chapter 2 in Macmillan & Creelman, 1990) for thorough treatments.

In light of these criteria, it is useful to consider optimal performance in terms of statistical inference, which helps interpreting the results. Ideally, the proportion of equivalence decisions should be 1 if conditions are truly equivalent (i.e.,  $|\delta| < m$ ) and 0 if conditions are truly non-equivalent (i.e.,  $|\delta| > m$ ). The proportion of equivalence decisions should switch from 1 to 0 at  $m$ . A classifier with this behavior could perfectly discriminate between truly non-equivalent and equivalent conditions. In contrast, a classifier that would always decide for non-equivalence does not have discriminatory power and is said to be maximally conservative. Further, a classifier that would always decide for equivalence is also lacking any discriminatory power and is said to be maximally liberal.

In general, it is expected that the average decision approaches the ideal decision as the distance of  $\delta$  to  $-m$  and  $m$  grows (first criterion). That is,  $\delta = 0$  should yield the

highest proportions of equivalence decisions because it is at the center of our two selected symmetric equivalence intervals. On the other hand, a population effect size of  $\delta \rightarrow \infty$  should result in the lowest proportions of equivalence decisions. These general predictions are reflected in the results: For all approaches and scenarios, the proportion of equivalence decisions is highest when  $\delta = 0$  and lowest when  $\delta = 0.5$ . Moreover, the results indicate that across all approaches, decision performances, as defined by the first and second criteria, improve as  $n$  increases. Exceptions are the TOST and HDI-ROPE approaches when  $m = 0.1$ ; with our choices of sample sizes, decisions remain maximally conservative.

Although the classification performances of all three approaches were quite similar for very large  $n$  and for  $m = 0.2$  or  $m = 0.3$  according to the first and second criteria, the third criterion of equal proportions of equivalence and non-equivalence decisions around the equivalence margins is not fulfilled by the TOST and the HDI-ROPE procedures. That is, these two approaches are biased towards making non-equivalence decisions. In the most extreme cases, when  $m = 0.1$ , when  $m = 0.2$  and  $n = 50$  or  $n = 100$ , and when  $m = 0.3$  and  $n = 50$ , these two approaches always decide for non-equivalence. Hence, for these scenarios, the two approaches are maximally conservative and do not possess any discriminatory capabilities.

TOST and HDI-ROPE have no discriminatory power for  $m = 0.1$  for our chosen sample sizes. When  $m = 0.2$  or  $m = 0.3$  the TOST and HDI-ROPE approaches remain conservative, but their performance improves as  $n$  increases. That is, the proportion of equivalence decisions is close to 0 when  $\delta$  is large or close to  $m$ ; at the same time, high proportions of equivalence decisions are only obtained for very large  $n$  and when  $\delta$  is close to 0. In addition, given a specific value for  $\delta$ , the proportion of equivalence decisions increases as  $m$  increases. This shows that the TOST and HDI-ROPE procedures require a large sample size or a wide equivalence margin in order to have any discriminatory power between equivalence and non-equivalence. Even in the scenario with highest  $m$  and  $n$ , TOST and HDI-ROPE are biased towards non-equivalence.

In comparison, the BF approach is more liberal and fulfills the third criterion more than the other two approaches. Although there is considerable fluctuation for the proportion of equivalence decisions around  $\delta = m$ , depending on the Cauchy prior scale  $r$  and the Bayes factor threshold, the proportion of equivalence decisions is closer to 0.5 compared to the other two approaches. This can be seen in Figures 2, 3, and 4 by the fact that the curves for the BF approach are shifted towards the right. Specifically, the BF approach features higher true positive rates when conditions are truly equivalent but also higher false positive rates when conditions are truly non-equivalent. Within the BF approach, more conservative results are obtained for higher Bayes factor thresholds and smaller Cauchy prior scales  $r$ . Therefore, in theory the BF approach can be made arbitrarily conservative when higher Bayes factor thresholds and smaller Cauchy prior scales are chosen (see also Lindley, 1957) but in practice the latter approach is not advised because an extremely narrow or wide prior compromises objectivity, does not necessarily represent the knowledge at hand, and lowers the diagnosticity of the test. Contrary to the other two approaches, the BF approach displays at least some limited discriminatory capabilities for situations with small  $n$ , as evident from the presence of a slope.

### Discussion

Quantifying evidence that two conditions are practically equivalent on some outcome measure of interest is important for various areas of research. The traditional  $t$ -test cannot be used to answer this question because it only allows quantification of evidence against but not in favor of equivalence (e.g., Bakan, 1966; van Ravenzwaaij et al., 2019). However, there are alternative procedures that can be utilized to quantify evidence in favor of equivalence, most prominently the frequentist TOST (Hodges & Lehmann, 1954; Schuirmann, 1987; Westlake, 1976; see also Lakens, 2017; Lakens, Scheel, & Isager, 2018) and the Bayesian HDI-ROPE (Kruschke, 2010, 2011, 2013, 2015, 2018; Kruschke et al., 2012; Kruschke & Liddell, 2018a, 2018b) and BF (Morey & Rouder, 2011; van Ravenzwaaij et al., 2019; see also Linde & van Ravenzwaaij, 2019a) procedures.

We compared true and false positive rates for each of these three approaches in the context of finding evidence towards equivalence for different scenarios that varied the sample size, the equivalence margin, and the population effect size. The TOST and the HDI-ROPE approaches behaved quite similarly. Both approaches did not discriminate between truly equivalent and truly non-equivalent conditions when the sample size or the equivalence margin were relatively small. In these situations, maximally conservative behavior was observed, which was evident by the absence of equivalence decisions. As the sample size or the equivalence margin increased, classifications improved but both approaches remained comparatively conservative in terms of deciding for equivalence. Although far from perfect, the BF approach exhibited some discriminatory capabilities even for small sample sizes and equivalence margins. Decisions improved as the sample size or the equivalence margin increased. In contrast to the TOST and HDI-ROPE approaches, the BF approach was comparatively liberal in terms of deciding for equivalence. This implies that the BF approach elicits more true positives but also more false positives compared to the other two approaches.

Although our results suggest that the TOST and HDI-ROPE approaches are conservative and the BF approach comparatively liberal, it should be noted that theoretically all three approaches can be made arbitrarily conservative (or liberal). The BF approach can be made more conservative by setting higher evidence thresholds and by using smaller Cauchy prior scales (or local priors around 0 in general). For the HDI-ROPE approach, we can also change the prior and adjust the probability of the HDI to control the degree of conservativeness. Lastly, for the TOST procedure the significance level  $\alpha$  could be increased or decreased to make it more liberal or conservative, respectively. Practically, however, it is undesirable to decrease the probability of the HDI and to increase  $\alpha$  for TOST. Decreasing the probability of the HDI implies lower certainty, which is undesirable. Similar concerns exist for increasing  $\alpha$  for TOST: Many have argued that the default significance level of  $\alpha = .05$  is already too liberal and should even be lowered in order to



avoid a high rate of false positive findings (e.g., Benjamin et al., 2018; Greenwald, Gonzalez, Harris, & Guthrie, 1996; but see, e.g., De Ruiter, 2019; Lakens, Adolphi, et al., 2018). Given these practical considerations, we conclude that the TOST and the HDI-ROPE approaches are not suitable for research with small sample sizes.

The TOST and HDI-ROPE approaches might be suitable options when the research problem allows sampling of a large number of cases. We argue that for relatively small sample sizes and narrow equivalence intervals, these two approaches should be avoided because they do not possess any or only very limited discriminatory capabilities (see Figures 2, 3, and 4). At the most extreme, using the TOST or HDI-ROPE approaches on samples with a few hundred cases and a reasonable equivalence interval results in a foregone decision for non-equivalence, making them poor choices for practical settings. In general, making decisions based on studies with small samples should be avoided. However, if larger sample sizes are impossible to achieve (e.g., neuropsychological studies on very rare disorders; or studies on infant speech development where the participants lose attention after a few trials), we prefer the use of Bayes factors to test for equivalence, since the BF approach could at least discriminate between truly equivalent and truly non-equivalent conditions to some extent in situations with sparse data. Furthermore, only the BF approach collapses into a regular point hypothesis test as the equivalence margin approaches 0; the TOST and the HDI-ROPE approaches fail to work in this situation (see Meyners, 2012, for an explanation).

As a result, we recommend the use of Bayes factors for quantifying evidence towards equivalence. Bayes factors for equivalence designs can easily be calculated with the `baymedr` (Linde & van Ravenzwaaij, 2019b) or the `BayesFactor` (Morey & Rouder, 2018) packages written in R (R Core Team, 2019) or with JASP (JASP Team, 2020).

## Conclusions

Quantifying evidence that two groups are equivalent on some outcome measure is an important tool for various areas of science. Unfortunately, traditional NHST cannot be

utilized for this because a non-significant  $p$ -value could indicate that  $\mathcal{H}_0$  of no effect is true or that there was not enough power to detect an effect; it is impossible to disentangle the two possibilities (e.g., Bakan, 1966; van Ravenzwaaij et al., 2019). The frequentist TOST, the Bayesian HDI-ROPE, and the interval BF approaches allow for quantifying evidence for equivalence. We compared the classification performances of these approaches for various scenarios and found that the BF approach is particularly useful for research with small sample sizes, although obtaining more data should always be preferred. With small samples, the TOST and HDI-ROPE approaches have no discriminatory power and result in a foregone decision for non-equivalence. For large sample sizes the three approaches all have near-perfect classification when the population effect size is close to zero or very large, but the BF approach outperforms both other approaches for population effect sizes near the edge of the equivalence interval. Taking all of this together, we recommend the use of interval Bayes factors for gathering evidence for equivalence.

### **Acknowledgements**

This research was supported by a Dutch scientific organization VIDI fellowship grant awarded to Don van Ravenzwaaij (016.Vidi.188.001) and a Dutch scientific organization VICI fellowship grant awarded to Eric-Jan Wagenmakers (016.Vici.170.083).

### References

- Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 31. doi:10.1525/collabra.149
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. doi:10.1037/h0020412
- Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*(3), 1550–1577. doi:10.1214/12-AOS1013
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. doi:10.1038/s41562-017-0189-z
- Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3), 385–402. doi:10.1214/06-BA115
- Berger, J. O. & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*(2), 159–165. doi:10.1016/0278-2316(88)90057-6
- Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., . . . Sivaganesan, S. (1994). An overview of robust Bayesian analysis. *Test*, *3*(1), 5–124. doi:10.1007/BF02562676
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi:10.1038/nrn3475
- Chow, S.-C., Shao, J., & Wang, H. (2002). A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics*, *12*(4), 441–456. doi:10.1081/BIP-120016229

- Christensen, E. (2007). Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology*, *46*(5), 947–954. doi:10.1016/j.jhep.2007.02.015
- Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, *30*(2), 178–191. doi:10.1017/S0305004100016595
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Consonni, G., Fouskakis, D., Liseo, B., & Ntzoufras, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679. doi:10.1214/18-BA1103
- Cuijpers, P., Van Straten, A., & Andersson, G. (2008). Internet-administered cognitive behavior therapy for health problems: A systematic review. *Journal of Behavioral Medicine*, *31*(2), 169–177. doi:10.1007/s10865-007-9144-1
- De Ruiter, J. (2019). Redefine or justify? comments on the alpha debate. *Psychonomic Bulletin & Review*, *26*(2), 430–433. doi:10.3758/s13423-018-1523-9
- Dienes, Z. & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218. doi:10.3758/s13423-017-1266-z
- Du, H., Edwards, M. C., & Zhang, Z. (2019). Bayes factor in one-sample tests of means with a sensitivity analysis: A discussion of separate prior distributions. *Behavior Research Methods*, *51*(5), 1998–2021. doi:10.3758/s13428-019-01262-w
- Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. doi:10.3758/s13423-017-1317-5
- Etz, A. & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin and Review*, *25*(1), 5–34. doi:10.3758/s13423-017-1262-3

- Fechner, G. T. (1966). *Elements of psychophysics*. New York, NY: Holt, Rinehart & Winston. (Original work published 1860)
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. doi:10.1037/a0015251
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). London, UK: Chapman & Hall/CRC.
- Gelman, A. & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *180*(4), 967–1033. doi:10.1111/rssa.12276
- Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Analysis*, *1*(3), 403–420. doi:10.1214/06-BA116
- Goodman, S. N. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. doi:10.1053/j.seminhematol.2008.04.003
- Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: John Wiley & Sons.
- Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, *33*(2), 175–183. doi:10.1111/j.1469-8986.1996.tb02121.x
- Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 137–143. doi:10.1080/00031305.2018.1562983
- Hahn, U. & Harris, A. J. L. (2014). What does it mean to be biased: Motivated reasoning and rationality. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Chap. 2, Vol. 61, pp. 41–101). San Diego, CA: Academic Press.
- Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association*, *288*(3), 358–362. doi:10.1001/jama.288.3.358

- Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2), 261–268.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. doi:10.1371/journal.pmed.0020124
- JASP Team. (2020). JASP (Version 0.12.2)[Computer software]. Retrieved from <https://jasp-stats.org/>
- Jeffreys, H. (1939). *Theory of probability*. Oxford, UK: The Clarendon Press.
- Jeffreys, H. (1948). *Theory of probability* (2nd). Oxford, UK: The Clarendon Press.
- Jeffreys, H. (1961). *Theory of probability* (3rd). Oxford, UK: Oxford University Press.
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. doi:10.2307/2291091
- Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5), 658–676. doi:10.1002/wcs.72
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, 6(3), 299–312. doi:10.1177/1745691611406925
- 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. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd). Boston, MA: Academic Press.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280. doi:10.1177/2515245918771304
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, 15(4), 722–752. doi:10.1177/1094428112457829

- Kruschke, J. K. & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, *25*(1), 155–177. doi:10.3758/s13423-017-1272-1
- Kruschke, J. K. & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. doi:10.3758/s13423-016-1221-4
- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177
- Lakens, D., Adolfs, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. doi:10.1038/s41562-018-0311-x
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963
- Lee, M. D. & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.
- Lesaffre, E. (2008). Superiority, equivalence, and non-inferiority trials. *Bulletin of the NYU Hospital for Joint Diseases*, *66*(2), 150–154.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. doi:10.1198/016214507000001337
- Linde, M. & van Ravenzwaaij, D. (2019a). baymedr: An R package for the calculation of Bayes factors for equivalence, non-inferiority, and superiority designs. arXiv: 1910.11616 [stat.OT]
- Linde, M. & van Ravenzwaaij, D. (2019b). *baymedr: Computation of Bayes factors for common biomedical designs*. R package version 0.1.0. Retrieved from <https://CRAN.R-project.org/package=baymedr>

- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.  
doi:10.2307/2333251
- Macmillan, N. A. & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and “nonparametric” indexes. *Psychological Bulletin*, *107*(3), 401–413. doi:10.1037/0033-2909.107.3.401
- Macmillan, N. A. & Creelman, C. D. (2005). *Detection theory: A user’s guide* (2nd). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163.  
doi:10.1037/1082-989X.9.4.425
- 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
- Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. doi:10.1037/a0024377
- Morey, R. D. & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.2. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>
- O’Hagan, A. & Forster, J. (2004). *Kendall’s advanced theory of statistics: Vol. 2B. Bayesian inference* (2nd). London, UK: Arnold.
- Phillips, K. F. (1990). Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *18*(2), 137–144.  
doi:10.1007/BF01063556



- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*(1), 102–113. doi:10.3758/s13423-017-1420-7
- 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
- Schuirman, 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
- Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316.
- Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE*, *11*(9), e0162093. doi:10.1371/journal.pone.0162093
- Siqueira, A. L., Whitehead, A., Todd, S., & Lucini, M. M. (2005). Comparison of sample size formulae for  $2 \times 2$  cross-over designs applied to bioequivalence studies. *Pharmaceutical Studies*, *4*(4), 233–243. doi:10.1002/pst.183
- Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi:10.3758/BF03207704
- Steiner, M. E., Ness, P. M., Assmann, S. F., Triulzi, D. J., Sloan, S. R., Delaney, M., . . . Stowell, C. P. (2015). Effects of red-cell storage duration on patients undergoing cardiac surgery. *The New England Journal of Medicine*, *372*(15), 1419–1429. doi:10.1056/NEJMoa1414219

- Tendeiro, J. N. & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. doi:10.1037/met0000221
- Tsai, C.-A., Huang, C.-Y., & Liu, J.-p. (2014). An approximate approach to sample size determination in bioequivalence testing with multiple pharmacokinetic responses. *Statistics in Medicine*, *33*(19), 3300–3317. doi:10.1002/sim.6182
- van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1), 71. doi:10.1186/s12874-019-0699-7
- Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040. doi:10.1080/17470218.2014.885986
- Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. doi:10.1016/j.jmp.2010.07.003
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi:10.3758/s13423-017-1343-3
- Walker, E. & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192–196. doi:10.1007/s11606-010-1513-8
- Wang, H. & Chow, S.-C. (2002). On statistical power for average bioequivalence testing under replicated crossover designs. *Journal of Biopharmaceutical Statistics*, *12*(3), 295–309. doi:10.1081/BIP-120014560
- Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*(4), 741–744. doi:10.2307/2529259

## Appendix

### TOST and Its Statistical Power

For the TOST procedure (Hodges & Lehmann, 1954; Schuirmann, 1987; Westlake, 1976; see also Lakens, 2017; Lakens, Scheel, & Isager, 2018) we assume here that we have two independent samples that are drawn from Normal distributions:  $x_{ij} \sim \text{Normal}(\mu_i, \sigma)$  for  $i \in \{c, e\}$  and  $j \in \{1, \dots, n_i\}$ . Sample means and standard deviations are denoted  $\bar{x}_i$  and  $s_i$ , respectively, and the pooled standard deviation is  $s = \sqrt{((n_c - 1)s_c^2 + (n_e - 1)s_e^2) / (n_c + n_e - 2)}$ . We use a symmetric unstandardized equivalence interval  $(-\lambda, \lambda)$ . The difference in population means of the two groups is  $\theta = \mu_e - \mu_c$ .

TOST consists of conducting two one-sided tests. The first one is concerned with the lower boundary  $-\lambda$  of the equivalence interval

$$\mathcal{H}_{0-\lambda}: \theta \leq -\lambda \qquad \mathcal{H}_{1-\lambda}: \theta > -\lambda \qquad (\text{A.1})$$

and the second one with the upper boundary  $\lambda$  of the equivalence interval

$$\mathcal{H}_{0\lambda}: \theta \geq \lambda \qquad \mathcal{H}_{1\lambda}: \theta < \lambda. \qquad (\text{A.2})$$

The null and alternative hypotheses of the two one-sided tests in Equations A.1 and A.2 can be compressed as

$$\mathcal{H}_0: \theta \leq -\lambda \text{ OR } \theta \geq \lambda \qquad \mathcal{H}_1: \theta > -\lambda \text{ AND } \theta < \lambda. \qquad (\text{A.3})$$

We conduct two one-sided  $t$ -tests. Assuming equal variances of the two conditions in the population, the standard error is  $se = s\sqrt{1/n_c + 1/n_e}$  with  $df = n_c + n_e - 2$  degrees of freedom.  $\mathcal{H}_0$  of non-equivalence is rejected with a Type I error rate of  $\alpha$  if

$$t_{-\lambda} = \frac{\bar{x}_e - \bar{x}_c - (-\lambda)}{se} > t_{1-\alpha, df} \qquad \text{AND} \qquad t_{\lambda} = \frac{\bar{x}_e - \bar{x}_c - \lambda}{se} < -t_{1-\alpha, df}, \qquad (\text{A.4})$$

where  $t_{1-\alpha, df}$  is the critical  $t$ -value at the  $1 - \alpha$  quantile and with  $df$  degrees of freedom.

Equation A.4 is equivalent to

$$-\lambda < (\bar{x}_e - \bar{x}_c) - t_{1-\alpha, df} \times se \quad \text{AND} \quad (\bar{x}_e - \bar{x}_c) + t_{1-\alpha, df} \times se < \lambda. \quad (\text{A.5})$$

Observe that

$$((\bar{x}_e - \bar{x}_c) - t_{1-\alpha, df} \times se, (\bar{x}_e - \bar{x}_c) + t_{1-\alpha, df} \times se) \quad (\text{A.6})$$

is the  $(1 - 2\alpha)$  100% CI for  $\bar{x}_e - \bar{x}_c$ . So, rejecting  $\mathcal{H}_0$  of TOST at significance level  $\alpha$  is equivalent to having the  $(1 - 2\alpha)$  100% CI for  $(\bar{x}_e - \bar{x}_c)$  completely inside the equivalence region  $(-\lambda, \lambda)$ .

This observation lets us define the rejection region ( $RR$ ) of TOST (e.g., Tsai, Huang, & Liu, 2014); from Equation A.6 it follows that

$$RR = \{-\lambda + t_{1-\alpha, df} \times se < \bar{x}_e - \bar{x}_c < \lambda - t_{1-\alpha, df} \times se\}. \quad (\text{A.7})$$

Values to the left of the  $RR$  pertain to non-significance of  $t_{-\lambda}$  while values to the right of the  $RR$  pertain to non-significance of  $t_{\lambda}$ . The  $RR$  can be visualized by plotting  $\bar{x}_e - \bar{x}_c$  on the  $x$ -axis and  $se$  on the  $y$ -axis, which yields an isosceles triangle with two vertices at  $\bar{x}_e - \bar{x}_c = -\lambda$  and  $\bar{x}_e - \bar{x}_c = \lambda$ , with  $se = 0$ , and one vertex at  $\bar{x}_e - \bar{x}_c = 0$ , with  $se$  reaching a maximum (cf. Meyners, 2012; Schuirmann, 1987). Importantly, the  $RR$  is only defined if  $se < \lambda/t_{1-\alpha, df}$  (Shieh, 2016).

The power function is the probability associated with the  $RR$  (see Equation A.7):

$$1 - \beta = p(-\lambda + t_{1-\alpha, df} \times se < \bar{x}_e - \bar{x}_c < \lambda - t_{1-\alpha, df} \times se). \quad (\text{A.8})$$

Suppose that the true difference in means,  $\theta$ , lies under  $\mathcal{H}_1$  (i.e.,  $-\lambda < \theta < \lambda$ ). In this case, we have that  $\bar{x}_e - \bar{x}_c \sim \text{Normal}(\theta, \sigma\sqrt{1/n_c + 1/n_e})$  and can therefore rewrite

Equation A.8:

$$\begin{aligned}
1 - \beta &= p \left( \frac{-\lambda - \theta + t_{1-\alpha, df} \times se}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} < \underbrace{\frac{(\bar{x}_e - \bar{x}_c) - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}}_{\sim \text{Normal}(0, 1)} < \frac{\lambda - \theta - t_{1-\alpha, df} \times se}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} \right) \\
&= \Phi \left( \frac{\lambda - \theta - t_{1-\alpha, df} \times se}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} \right) - \Phi \left( \frac{-\lambda - \theta + t_{1-\alpha, df} \times se}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} \right) \\
&= \Phi \left( \frac{\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} - t_{1-\alpha, df} \frac{s}{\sigma} \right) - \Phi \left( \frac{-\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} + t_{1-\alpha, df} \frac{s}{\sigma} \right),
\end{aligned} \tag{A.9}$$

where  $\Phi(\cdot)$  denotes the cumulative distribution function of the standard Normal distribution. As a consequence of Cochran's theorem (Cochran, 1934), we have  $V = df (s^2/\sigma^2) \sim \chi^2(df)$ . Rewriting  $s/\sigma$  as  $\sqrt{V/df}$ , Equation A.9 becomes

$$\begin{aligned}
1 - \beta &= \Phi \left( \frac{\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} - t_{1-\alpha, df} \sqrt{\frac{V}{df}} \right) \\
&\quad - \Phi \left( \frac{-\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} + t_{1-\alpha, df} \sqrt{\frac{V}{df}} \right).
\end{aligned} \tag{A.10}$$

The power function is thus rewritten as a function of the  $\chi^2$ -distributed random variable  $V$ . As a result,  $s$  was removed and instead its uncertainty was incorporated through the  $\chi^2$  distribution.

We can express the power through an integral:

$$1 - \beta = \int_0^\infty \left[ \Phi \left( \frac{\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} - t_{1-\alpha, df} \sqrt{\frac{v}{df}} \right) - \Phi \left( \frac{-\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} + t_{1-\alpha, df} \sqrt{\frac{v}{df}} \right) \right] \varphi_{df}(v) dv, \quad (\text{A.11})$$

where  $\varphi_{df}(\cdot)$  is the density of the  $\chi^2$  distribution with  $df$  degrees of freedom. The limits of the integral reflect the support of a typical  $\chi^2$  distribution, which is in the interval  $(0, \infty)$ . However, in our case these limits of the integral cannot apply because the  $RR$  is only defined if  $se < \lambda/t_{1-\alpha, df}$  (Shieh, 2016). This implies that

$$V = df \frac{s^2}{\sigma^2} < \frac{df \lambda^2}{\sigma^2 t_{1-\alpha, df}^2 \left( \frac{1}{n_c} + \frac{1}{n_e} \right)} := V^*. \quad (\text{A.12})$$

We must therefore constrain the integration interval to  $(0, V^*)$ :

$$1 - \beta = \int_0^{V^*} \left[ \Phi \left( \frac{\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} - t_{1-\alpha, df} \sqrt{\frac{v}{df}} \right) - \Phi \left( \frac{-\lambda - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} + t_{1-\alpha, df} \sqrt{\frac{v}{df}} \right) \right] \varphi_{df}(v) dv. \quad (\text{A.13})$$

### Function to Compute Power for TOST

Shieh (2016) provided a function, written in R (R Core Team, 2019), that numerically solves the integral in Equation A.13. We adapted and improved this function by applying more accurate numerical integration and by improving readability of the code and eliminating defined but unused variables (this and other code can be accessed at <https://osf.io/f9dw2/>):

```
tost <- function(
```

```

n1,          #sample size 1st group
n2,          #sample size 2nd group
alpha,       #type 1 error rate
margin,      #equivalence margin (unstandardized)
es,          #effect size
sd           #population standard deviation
) {
  sd_sq      <- sd^2
  num_int    <- 100000
  df         <- n1 + n2 - 2
  t_crit     <- qt(1 - alpha, df)
  n_fac      <- 1 / n1 + 1 / n2
  var        <- sd_sq * n_fac
  std        <- sqrt(var)
  cu         <- (df * margin^2) / (var * t_crit^2)
  cvec       <- seq(0, cu, length.out = num_int)
  st         <- sqrt(cvec / df) * t_crit
  int_arg    <- (pnorm((margin - es) / std - st) -
                pnorm((-margin - es) / std + st)) *
                dchisq(cvec, df)
  width      <- cvec[2] - cvec[1]
  epower     <- sum(width * int_arg)
  return(epower)
}

```

The code above relates to our formulas in the following way:

- $sd\_sq = \sigma^2$

- `num_int` is the number of integration points
- `df = df`
- `t_crit = t_{1-\alpha, df}`
- `n_fac = 1/n_1 + 1/n_2`
- `var = \sigma^2 (1/n_1 + 1/n_2)`
- `std = \sigma \sqrt{1/n_1 + 1/n_2}`
- `cu = V^*`
- `cvec` is a vector of equally spaced integration points of length `num_int` in the interval  $[0, V^*]$
- `epower` solves Equation A.13 through numerical integration.

### Power for TOST with Non-central $t$ Distributions

Power for TOST can also be approached with the non-central  $t$  distribution. First, we can rewrite the  $t_{-\lambda}$  and  $t_{\lambda}$  test statistics as follows (see, e.g., Phillips, 1990, Equation 5):

$$\begin{aligned}
 t_{-\lambda} &= \frac{\bar{x}_e - \bar{x}_c - (-\lambda)}{s \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} = \frac{(\bar{x}_e - \bar{x}_c - \theta) + (\theta + \lambda)}{s \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} = \frac{\frac{\bar{x}_e - \bar{x}_c - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} + \frac{\theta + \lambda}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}}{\frac{s \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}} \\
 &= \frac{\overbrace{\frac{\bar{x}_e - \bar{x}_c - \theta}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}}^{\mathcal{N}(0, 1)} + \overbrace{\frac{\theta + \lambda}{\sigma \sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}}^{\text{NCP}_{-\lambda}}}{\sqrt{\underbrace{\left(\frac{s^2}{\sigma^2 / df}\right)}_{\chi_{df}^2}} / df} \sim t_{df}(\text{NCP}_{-\lambda}),
 \end{aligned} \tag{A.14}$$



where  $t_{df}$  is the probability density function of the non-central  $t$  distribution with  $df$  degrees of freedom and the given non-centrality parameter NCP. The result is similar for  $t_\lambda$ :

$$t_\lambda = \frac{\bar{x}_e - \bar{x}_c - \lambda}{s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} \sim t_{df}(\text{NCP}_\lambda), \quad \text{NCP}_\lambda = \frac{\theta - \lambda}{\sigma\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}. \quad (\text{A.15})$$

Then, we rewrite Equation A.8:

$$\begin{aligned} 1 - \beta &= 1 - p\left(\bar{x}_e - \bar{x}_c < -\lambda + t_{1-\alpha, df} \times s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}\right) \\ &\quad - p\left(\lambda - t_{1-\alpha, df} \times s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}} < \bar{x}_e - \bar{x}_c\right) \\ &= 1 - p\left(\frac{\bar{x}_e - \bar{x}_c + \lambda}{s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} < t_{1-\alpha, df}\right) - p\left(\frac{\bar{x}_e - \bar{x}_c - \lambda}{s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}} > -t_{1-\alpha, df}\right) \end{aligned} \quad (\text{A.16})$$

$$\begin{aligned} &= 1 - \mathcal{T}_{df}\left(t_{1-\alpha, df} \left| \frac{\theta + \lambda}{\sigma\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right.\right) - 1 + \mathcal{T}_{df}\left(-t_{1-\alpha, df} \left| \frac{\theta - \lambda}{\sigma\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right.\right) \\ &= \mathcal{T}_{df}\left(-t_{1-\alpha, df} \left| \frac{\theta - \lambda}{\sigma\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right.\right) - \mathcal{T}_{df}\left(t_{1-\alpha, df} \left| \frac{\theta + \lambda}{\sigma\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right.\right), \end{aligned} \quad (\text{A.17})$$

where  $\mathcal{T}_{df}(\cdot | \text{NCP})$  is the cumulative distribution function for the non-central  $t$  distribution with  $df$  degrees of freedom and non-centrality parameter NCP. In the last equality of Equations A.16 and A.17, a property of function  $\mathcal{T}_{df}(\cdot | \text{NCP})$  is used, namely that  $\mathcal{T}_{df}(-x | \text{NCP}) = 1 - \mathcal{T}_{df}(x | -\text{NCP})$ . Equation A.17 is the power formula in Chow, Shao, and Wang (2002, p. 451, formula 2) and also Shieh (2016, formula A4 in supplement S1).

Equation A.17 gives another formula to compute power. Importantly, however, Equation A.17 is an approximation and is only exact under the constraint that  $s\sqrt{1/n_c + 1/n_e} < \lambda/t_{1-\alpha, df}$  (cf. Shieh, 2016, p. 2 in supplement S1). Shieh (2016) argues that this constraint is inconsequential in most situations (see also Siqueira, Whitehead, Todd, & Lucini, 2005; Wang & Chow, 2002). The probability that the constraint is not

met is:

$$\begin{aligned}
p\left(s\sqrt{\frac{1}{n_c} + \frac{1}{n_e}} > \frac{\lambda}{t_{1-\alpha, df}}\right) &= 1 - p\left(s < \frac{\lambda}{t_{1-\alpha, df}\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right) \\
&= 1 - p\left(\sqrt{df}\frac{s}{\sigma} < \frac{\sqrt{df}\lambda}{\sigma t_{1-\alpha, df}\sqrt{\frac{1}{n_c} + \frac{1}{n_e}}}\right) \\
&= 1 - \mathcal{F}_{\chi_{df}^2}\left(\frac{df\lambda^2}{\sigma^2 t_{1-\alpha, df}^2\left(\frac{1}{n_c} + \frac{1}{n_e}\right)}\right) \\
&= 1 - \mathcal{F}_{\chi_{df}^2}(V^*).
\end{aligned} \tag{A.18}$$

Here  $\mathcal{F}_{\chi_{df}^2}(\cdot)$  is the distribution function of the  $\chi^2$  distribution with  $df$  degrees of freedom.

So, the approximation given by Equation A.17 (see also Shieh, 2016, Equation A4 in supplement 1) is not good once  $V^*$  becomes small. For instance, when the equivalence interval is tight around 0 (i.e.,  $\lambda$  is small). Expressed differently, we get into problems when  $\chi^2(df)$  has a lot of mass above  $V^*$  (see Equation A.13).