2019 ■ Vol. 15 ■ no. 1

Traditional and bayesian approaches for testing mean equivalence and a lack of association

Joseph J. Hoyda [a], Alyssa Counsell [b] & Robert A. Cribbie [a, ✉]

[a]Department of Psychology, York University, Toronto
[b]Department of Psychology, Ryerson University, Toronto

Abstract ■ Researchers are often interested in demonstrating that variables are unrelated. However, declar-ing a lack of relationship (e.g., no mean difference or no correlation) through nonrejection of the traditional null hypothesis (e.g., H_0: $\mu_1 - \mu_2 = 0$; H_0: $\rho = 0$) is inappropriate. The two one-sided tests (TOST) method for testing mean equivalence is a popular approach to resolving this issue, and has been adapted for testing equivalence with various test statistics. In this study, two Bayesian alternatives to the TOST method for assessing equivalence of means or a lack of correlation were examined and compared to their equivalence testing analogs. The first is the Bayes factor method, which compares the relative evidence that the data were more likely under one hypothesis than another. The second method is Bayesian parameter estimation, using highest density intervals, which estimates a posterior distribution and seeks to demonstrate that the interval falls within bounds for establishing equivalence. The power rates of these procedures were first compared in a simulation study. Next, empirical examples of each of the approaches are shown using an openly available dataset on personality traits. Results identify the benefits and limitations of these competing alternatives under various testing conditions, and highlight the importance of using equivalence interval based methods in the behavioral sciences. .

Keywords ■ equivalence tests, Bayes factor, Bayesian estimation, correlation, t-test.

✉ cribbie@yorku.ca

JJH: 0000-0001-5962-9821; *AC*: 0000-0001-9449-6630; *RAC*: 0000-0002-9247-497X

10.20982/tqmp.15.1.p012

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ Daniel Lakens (Eindhoven University of Technology)

Introduction

Researchers in the behavioral sciences are frequently interested in determining if groups of subjects are equivalent on an outcome variable. For example, Thompson (2015) hypothesized that tattooed women would be as psychosocially healthy as non-tattooed women. The primary difficulty in demonstrating equivalence is that traditional null hypothesis significance testing (NHST) methods are designed only to reject, rather than accept, the null hypothesis. In order to demonstrate equivalence using this method, the researcher would have to retain the null hypothesis. It is well documented, however, that this strategy is inappropriate because absence of evidence is not evidence of absence (Altman & Bland, 1995; Rogers, Howard, & Vessey, 1993). Thus, it is important that researchers be aware of, and have access to, methods that can properly evaluate equivalence.

Over the past few decades, the field of equivalence testing has provided a variety of alternative hypothesis testing methods to assess whether groups of subjects do not differ significantly on an outcome variable (e.g., Anderson & Hauck, 1983; Wellek, 2010; Westlake, 1976). The two-one sided tests (TOST) method (Schuirmann, 1987) for mean equivalence is one of the most popular approaches to resolving this issue, and has been adapted for other test statistics like correlation (e.g., Goertzen & Cribbie, 2010). In this paper, we highlight two additional methods relevant to equivalence testing from the field of Bayesian statistics, namely, Bayes factors (BF) and estimating highest density intervals (HDIs). We will focus on two popular instances in which researchers are interested in demonstrating a

lack of relationship: 1) equivalence of independent group means; and 2) a lack of correlation among continuous variables.

This paper aims to serve as a practical guide for researchers who wish to explore the benefits of Bayesian alternatives to traditional equivalence testing techniques. First, the TOST and two Bayesian approaches (BF and HDI) for mean equivalence and lack of correlation are outlined. Next, the power rates of these procedures are compared in a simulation study. The results of this study serve to identify the benefits and limitations of these competing alternatives under various testing conditions. Lastly, empirical examples using each of the approaches are provided based on data on personality traits. The examples demonstrate the ease with which researchers can employ these techniques, and further serve to highlight their theoretical differences.

Two One-Sided Testing Approach to Demonstrating Equivalence

The TOST procedure addresses the problem of retaining the null with traditional methods by reversing the null and alternative hypotheses. More specifically, when testing for mean equivalence, the traditional null hypothesis that there is no mean difference (e.g., H_0: $\mu_1 = \mu_2$) is replaced by a null hypothesis (actually two null hypotheses, see below) that state that there is a mean difference. Therefore, in order to determine equivalence, the researcher must reject the null hypothesis that there is a difference, in favor of an alternative hypothesis that there is no mean difference. Further, instead of using a strict nil hypothesis, which assumes that the difference is exactly zero (e.g., H_0: $\mu_1 - \mu_2 = 0$), the TOST procedure uses an equivalence interval (e.g., $-\varepsilon, \varepsilon$), where ε represents the smallest effect that would still be considered meaningful within the nature of the research (Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers et al., 1993). This concept of equivalence takes into account the fact that mean differences between groups are rarely, if ever, exactly zero, which is what is typically assumed by traditional NHST methods. Additionally, equivalence intervals give researchers the freedom to define the minimum effect size that would be considered meaningful for their research.

When testing for population mean equivalence, the TOST procedure uses two simultaneous one-sided t-tests to evaluate whether the difference in the population means falls within an equivalence interval. That is, they evaluate whether the difference in the means is small relative to the lower and upper bounds of the equivalence interval. The null hypothesis of non-equivalence then logically has two components:

$$H_{01} : \mu_1 - \mu_2 \geq \varepsilon$$
$$H_{02} : \mu_1 - \mu_2 \leq -\varepsilon$$

In order to demonstrate that the group means are equivalent, both null hypotheses must be rejected, which implies that $\mu_1 - \mu_2$ falls within the bounds of $(-\varepsilon, \varepsilon)$ (Schuirmann, 1987). Note that we are focusing on a symmetric equivalence interval (i.e., $|-\varepsilon| = \varepsilon$), however this is not necessary. H_{01} is rejected when $t_1 \leq t_{\alpha,df}$ and H_{02} is rejected when $t_2 \geq t_{1-\alpha,df}$, where t_1 and t_2 are calculated by:

$$t_1 = \frac{(M_1 - M_2) - \varepsilon}{s_{M_1 - M_2}}$$
$$t_2 = \frac{(M_1 - M_2) - (-\varepsilon)}{s_{M_1 - M_2}}$$

where

$$s_{M_1 - M_2} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 + n_2 - 2)} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

in which n_1 and n_2 are sample sizes for both groups.

Note that $s_{M_1 - M_2}$ is the standard error of the difference, M_i represents the ith group's sample mean, ni represents the ith group's sample size and si represents the ith group's sample standard deviation. The degrees of freedom are the same as that for an independent samples t-test (i.e., $df = n_1 + n_2 - 2$). The null hypothesis of non-equivalence is rejected when both H_{01} and H_{02} are rejected. An equivalent approach to the TOST method is to demonstrate that the $100(1 - 2\alpha)$% confidence interval of the difference between M_1 and M_2 falls completely within the equivalence interval (Westlake, 1976). Numerous extensions to the TOST method are also available, including a heteroscedastic two independent-samples procedure often referred to as the Schuirmann-Welch test (TOST-SW; Gruman, Cribbie, & Arpin-Cribbie, 2007). The TOST-SW is identical to the TOST except that the pooled standard error ($s_{M_1 - M_2}$) is replaced by a nonpooled standard error, i.e., $\sqrt{s_1^2/n_1 + s_2^2/n_2}$, and the degrees of freedom are that of the heteroscedastic Welch statistic (see Gruman et al., 2007). Further extensions include trimmed means and Winsorized variances which are appropriate with skewed or heavy-tailed distributions (see van Wieringen & Cribbie, 2014).

The TOST method has also been applied to the problem of detecting a lack of correlation by Goertzen and Cribbie (2010). In this case, the null hypotheses, $H_{01} : \rho \geq \rho^*$ and $H_{02} : \rho \leq -\rho^*$, are rejected if $t_1 \leq t_{\alpha,N-2}$ and $t_2 \geq t_{1-\alpha,N-2}$, where

$$t_1 = \frac{r - \rho^*}{\sqrt{\frac{1-r^2}{N-2}}} \qquad t_2 = \frac{r - (-\rho^*)}{\sqrt{\frac{1-r^2}{N-2}}}$$

in which ρ^* represents the half-width of the (symmetrical) lack of association interval $(-\rho*, \rho^*)$, N represents the sample size, $t_{\alpha,N-2}$ represents the α level critical value from the t distribution with $N-2$ degrees of freedom, and r represents the sample correlation value.

Bayesian Alternatives to the TOST Approach

While the TOST resolves some of the initial criticisms in regards to retaining the null hypothesis, it still uses traditional hypothesis testing. One of the primary advantages of Bayesian approaches, insofar as it is related to equivalence testing, is the ability to quantify evidence for, as well as against, the traditional null hypothesis. This can be done by quantifying the probability of the null hypothesis, given the data, or by providing relative evidence in favor of the null hypothesis relative to some competing hypothesis. These approaches are described in more detail below.

Bayesian statistics differ from frequentist approaches in (at least) two important ways. First, in frequentist statistics parameters are considered fixed, whereas in Bayesian statistics parameters are considered unknown and described probabilistically. Second, prior information regarding the model parameters is incorporated into the probability model for the parameters. Bayes theorem can be written as:

$$p\left(A|B\right) = \frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)}$$

where $p(A)$ is the prior credibility of parameter values, $p(B)$ is the overall probability of the data given the model, $p(A|B)$ is the conditional probability of A given B, and $p(B|A)$ is the conditional probability of B given A. There are three fundamental aspects of the Bayes calculations that are important for determining an outcome. The likelihood distribution, represented by $p(B|A)$, represents the plausibility of the data, given the hypothesized model. The prior distribution, represented by $p(A)$, represents any information that is known about the parameters before the study. Finally, the posterior distribution, represented by $p(A|B)$, is the final probability distribution that represents a compromise between the prior and the likelihood. In other words, the posterior is equal to the likelihood multiplied by the prior and divided by the overall evidence being examined. Here, a researcher obtains information about the probability of their hypotheses given prior information and data, rather than the probability of the data assuming that the null hypothesis is true.

The concept of the prior distribution is one of the most important contributions of Bayesian estimation. The use of priors allows the researcher to incorporate previous knowledge that may have been accumulated through other studies or research. For instance, if it is known that a specific parameter value is particularly unlikely, the model can be modified in order to take this into account. This can be especially useful if the researcher has a large body of research from which to draw. The other primary advantage is that Bayesian estimation allows researchers to accumulate knowledge by continuing to update their data; the resulting posterior distribution of one experiment can be used as an informative prior in future research. It is also important to note that although priors can be informative in cases where a lot is known about the possible parameter values, they can also be noninformative in cases where the researcher wants to minimize assumptions regarding parameters.

Bayesian HDI Approach to Equivalence Testing

The Bayesian highest density interval approach advocated by Kruschke (2013) estimates the probability of potential parameter values and allows researchers to calculate a Bayesian equivalent to a confidence interval (HDI) around those parameters. Applying this method to a two sample t-test, a researcher would estimate the distributions of five parameters: each of the population means (μ_1, μ_2), each of the population standard deviations (σ_1, σ_2) and a distribution shape parameter common across groups (ν). From these posterior distributions, information is garnered pertaining to the difference between the population means (i.e., $\mu_1 - \mu_2$); more specifically, the HDI for $\mu_1 - \mu_2$ is computed in order to determine if the HDI falls completely within the predefined equivalence interval (Kruschke labels the equivalence interval the 'region of practical equivalence' or ROPE). The posterior distribution represents the probability of those parameters, given the data, and, following Bayes theorem above, can be expressed as:

$$p\left(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu|D\right) = p\left(D|\mu_1, \mu_2, \sigma_1, \sigma_2, \nu\right) \times \frac{p\left(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu\right)}{p\left(D\right)}$$

in which $p\left(D|\mu_1, \mu_2, \sigma_1, \sigma_2, \nu\right)$ represents the likelihood of the data given the hypothesized parameters (i.e., the product of t distribution probability densities associated with each of the data values), $p\left(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu\right)$ represents the prior probability of the parameters (derived from the distributions of the five parameters) and $p\left(D\right)$ represents the marginal likelihood, or the overall probability of the data, and its purpose is to ensure that the posterior is a valid probability by making its area sum to 1.

In testing for a lack of association, a researcher estimates six parameters: the correlation (ρ), each of the population means (μ_1, μ_2), each of the population standard deviations (σ_1, σ_2) and a distribution shape parameter common across groups (ν). This posterior distribution represents the probability of those parameters, given the data,

and is calculated as:

$$p\left(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho | D\right) = p\left(D | \mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho\right) \times \frac{p\left(\mu_1, \mu_2, \sigma_1, \sigma_2, \nu, \rho\right)}{p\left(D\right)}$$

As with the t-test, these posterior distributions are used to compute the HDI (in this case the parameter of interest is ρ) in order to determine if it falls entirely within the predefined equivalence interval. In both the t-test and test for correlation, researchers can conclude equivalence if the $100(1-2\alpha)$% HDI falls within the equivalence region.

The posterior distributions are estimated using sampling techniques, rather than direct computation. The sampling is conducted using Markov Chain Monte Carlo (MCMC) methods. MCMC uses many samples of potential parameter values that fit the sample data and prior distributions simultaneously. From these samples for each parameter, the HDI for $\mu_1 - \mu_2$ or ρ can be explicitly determined by using the credible values from each sample.

Bayes Factor Approach to Equivalence Testing

The BF method developed by Jeffreys (1961) emphasizes the selection of one model over another by estimating the ratio of the likelihoods for the two models being compared. Thus, it compares the degree to which the data is probable under each of these models (e.g., the models associated with the null and alternative hypotheses). In its original formulation for demonstrating equivalence, the model associated with H_0 uses a spiked prior based on the point nil hypothesis ($\mu_1 - \mu_2 = 0$) and is then compared to an alternative model H_1. The BF can be expressed as:

$$BF = \frac{p(D|H_0)}{p(D|H_1)}$$

where $p(D|H_0)$ is the probability of the data, given H_0, and $p(D|H_1)$ is the probability of the data, given H_1. The BF therefore expresses the relative plausibility of these two models, resulting in a number that quantifies how much more likely one model is over another. For example, a BF of 10 indicates that the data are 10 times more likely to have occurred under the null hypothesis than under the alternate hypothesis; further, if it can be assumed that the two hypotheses were equally likely a priori, then the BF indicates that the null hypothesis is 10 times more likely than the alternate. Jeffreys (1961) qualifies the size of BF values by indicating that values less than 3 are not worth mentioning, between 3 and 10 are substantial, between 10 and 30 are strong, between 30 and 100 are very strong, and greater than 100 are decisive (also see Kass & Raftery, 1995).

The original BF approach by Jeffreys (1961) can be used for equivalence testing (i.e., quantifying relative evidence in favor of the null hypothesis), but it was not originally proposed to be used with an equivalence interval and thus the null hypothesis, as described above, was typically the nil hypothesis (i.e., $\delta = 0$). An important consequence of using a model based on the nil hypothesis is that the nil hypothesis will sometimes fail for trivial reasons. For instance, it has been argued by Meehl (1978) that nil hypotheses never hold to an arbitrary level of precision. Morey and Rouder (2011) address this issue by permitting the null hypothesis to be an interval within a specified distribution, with the alternate hypothesis representing the complement of the null. The goal of this method is to avoid rejecting the null hypothesis if the failure is due to trivial or uninteresting effects, making it a good fit for equivalence testing.

When using the BF approach, it is important to recognize that the concept of a prior is very different in the BF setting than in the Bayesian parameter estimation setting. In a Bayesian estimation setting, the prior explicitly specifies prior information about the parameter, whereas in a BF setting the prior specifies the comparison distribution (in our setting the alternative hypothesis). Take, for example, a situation in which a researcher would like to use a precise prior. Adopting a precise prior (e.g., a normal distribution with a standard deviation of .1 instead of a uniform $[-.1, 1]$ distribution for ρ) leads to a narrower HDI (and a greater chance of demonstrating equivalence) in an estimation setting. However, in a BF setting where relative evidence is being quantified, specifying a more precise prior actually specifies an alternative distribution that is more similar to the null and thus leads to less evidence in favor of the null (relative to a setting in which a less precise prior is adopted). In other words, the more similar the competing priors, the harder it will be to differentiate between them.

Simulation Study

A simulation study was conducted to provide researchers with power comparisons of the discussed TOST and Bayesian procedures under conditions thought to be common in behavioral science research. The goal of these simulations is to assist researchers in understanding how to quantify and compare the power of these theoretically diverse procedures. Three methods were compared: 1) TOST; 2) BF; and 3) Bayesian HDI. Each of these methods was applied to the problem of detecting the equivalence of two population means and to the problem of detecting a negligible correlation.

For the BF approach, both a nil hypothesis (BF-N) based approach, as proposed by Jeffreys (1961) and an equivalence interval based approach (BF-EQ), as developed by Morey and Rouder (2011), were included. For the BF

method, BF cutoffs of 10, 30 and 100 were used in accordance with Jeffreys's (1961) recommendations to evaluate when equivalence was successfully detected. These cutoffs refer to the required BF size to detect equivalence, since we are exploring the relative likelihood of the null hypothesis relative to the alternate hypothesis (e.g., a BF of 45 will be counted as equivalent by the BF10 and BF30 cutoffs, but not the BF100). This allows researchers to see how the power differs across the different cutoffs. For the HDI method, a 90% HDI was used to be consistent with both the $100(1-2\alpha)$% confidence interval approach (Westlake, 1976) and the TOST method when $\alpha = 0.05$. For both the correlation and t-test methods, 1000 replications were used for each condition. We acknowledge here that the HDI and BF methods were not necessarily designed for dichotomous types of decisions, e.g., equivalent means vs. nonequivalent means, however these methods are often used in this manner and for this study, this is necessary in order to compare the results to the TOST method.

For the mean difference problem, standardized equivalence intervals, based on population Cohen's d effect sizes, were set at $(-\varepsilon, \varepsilon) = (-.1, .1), (-.2, .2), (-.3, .3)$ and $(-.4, .4)$, except for the BF-N method, which uses a nil hypothesis. Although a standardized effect size of Cohen's $d = .4$ might seem large for an equivalence testing study, Rusticus and Eva (2016) found that participants did not find a mean difference to be meaningful until around $d = .5$. Total sample sizes used were 50, 100, 200 and 1000 with equal group sizes (e.g., with $N = 50$, $n_1 = 25$ and $n_2 = 25$). Standard deviations were set at 1 in each group. The posterior distribution for the t-test simulations was generated with the `BEST` package (Kruschke & Meredith, 2015) in R, using the default 100,000 iterations of MCMC. For the BF method, the `BayesFactor` R package was used (Morey & Rouder, 2015).

To explore how sample size and equivalence interval size affect power in tests for a lack of correlation, standardized equivalence intervals for ρ were set to $(-.05, .05), (-.1, .1), (-.15, .15)$ and $(-.2, .2)$, except for the BF-N method, which uses a nil hypothesis. Sample sizes were set at $N = 50, 100, 200$ and 1000. These sample sizes were chosen to be comparable to the sample sizes of the t-test simulations and to the sample sizes common in psychological research. The posterior distribution for the correlation simulations was generated with the `BayesianFirstAid` package in R (Bååth, Kruschke, & Meredith, 2014), using the default 15,000 iterations of MCMC. The BayesFactor package was used (Morey & Rouder, 2015) to calculate every version of the BF method, including the nil hypothesis.

For both the t-test and correlation simulations, noninformative priors were used to ensure minimal influence on the estimates of the parameter values. For the HDI approach, the priors on μ_1 and μ_2 were normal distributions with large standard deviations (i.e., 1000 times the pooled standard deviation of the groups), the priors on the standard deviations were uniform distributions ranging from $\sigma/1000$ to 1000σ, and lastly the prior distribution on the shape parameter was a shifted exponential ($\lambda = 1/29$, shift = 1; which allows for similar prior likelihood of a normal or heavy-tailed distribution) and represents the degrees of freedom for a t-distribution (recall that as the degrees of freedom increase the distribution approximates a normal distribution). In testing for a lack of correlation, the same priors as above were used in addition to a uniform $(-1, 1)$ prior for ρ to reflect the full range of possible values for a correlation coefficient. For the BF approach for the mean comparison, noninformative priors were also used for the population means and variances. More specifically, a Jeffreys prior was placed on σ_2, while a Cauchy prior is placed on the standardized effect size (d). For the BF approach for correlation, a uniform distribution was used as a prior for ρ (see Ly, Verhagen, & Wagenmakers, 2016). When using the BF-EQ method, the null hypothesis was an interval within the specified distribution [e.g., $\delta \sim \text{Cauchy}(0, 1)$, $\delta \in (-\varepsilon, \varepsilon)$; $\delta \sim \text{Uniform}(-1, 1)$, $\delta \in (-\varepsilon, \varepsilon)$], while the alternate hypothesis is the complement of the null [e.g., $\delta \sim \text{Cauchy}(0, 1)$, $\delta \notin (-\varepsilon, \varepsilon)$ and $\delta \sim \text{Uniform}(-1, 1)$, $\delta \notin (-\varepsilon, \varepsilon)$]. For the BF method, we also varied the scale of the alternative hypothesis distribution (often referred to as the 'rscale'), using values of .5, .75 and 1. Increasing the scale of the alternative hypothesis increases the variability of the distribution and thus makes it more dissimilar to the null distribution. For the mean comparison simulation, a scaling factor of 1 is equivalent to a standard Cauchy distribution, for correlation a scaling factor of 1 is equivalent to a uniform (-1,1) distribution. See (Morey & Rouder, 2011) for more details regarding the priors.

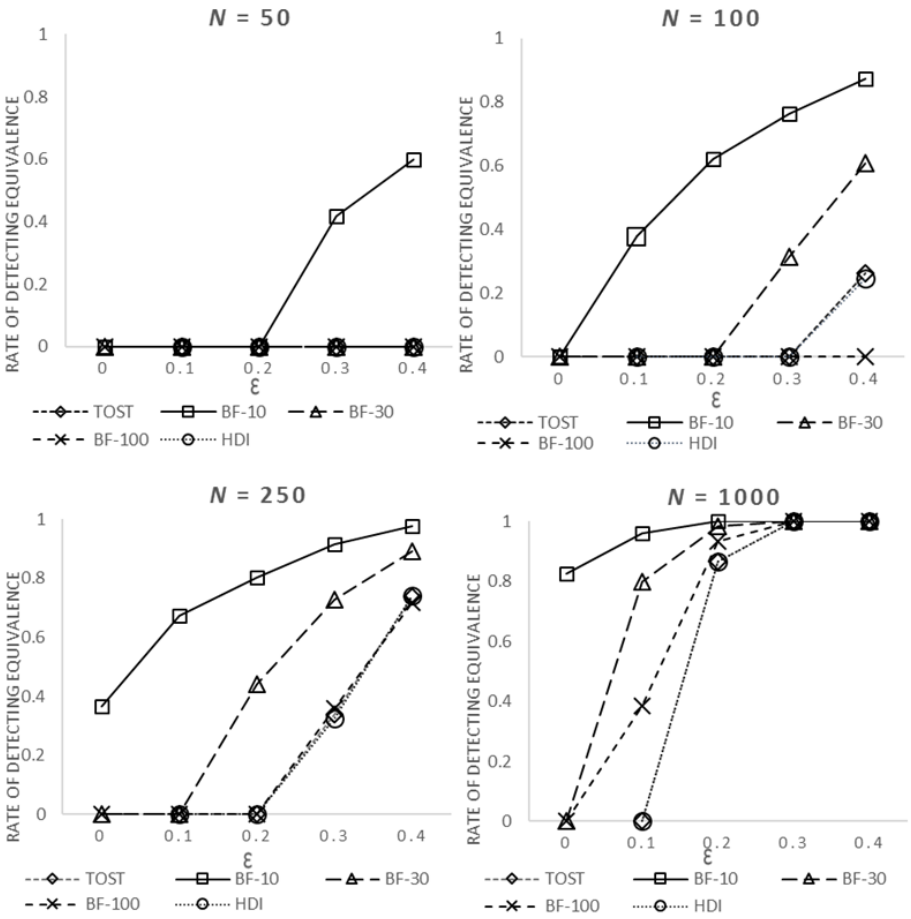
Results

The results for the different scalings of the alternative distribution produced expected effects; namely, by decreasing the scale of the alternative hypothesis the distribution becomes more like the null distribution and thus power for detecting equivalence decreases. Thus, due to space limitations only results for a scaling factor of 1 are presented here, however full tables of results are available at https://osf.io/tfdxq/. This issue is also discussed further in the discussion section.

Mean Equivalence

Power results for the TOST, BF approaches and HDI method are displayed in Figure 1, where the y-axis represents the

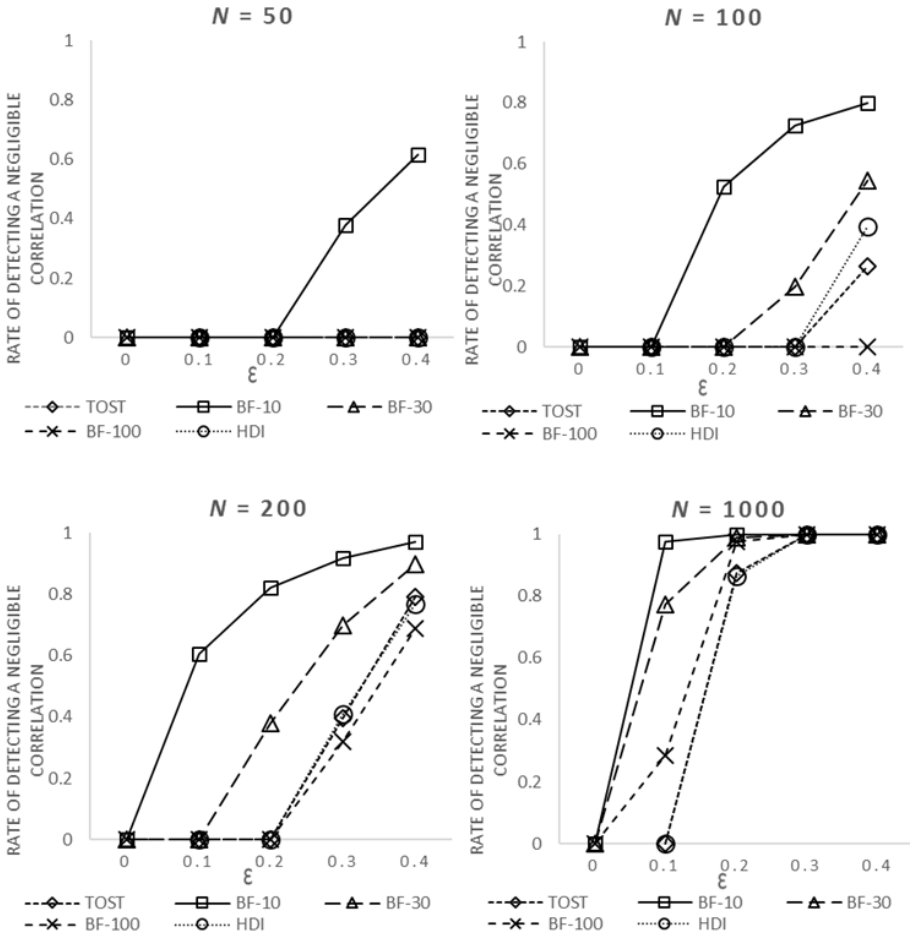
Figure 1 ■ Probability of concluding equivalence of means with each of the equivalence testing and Bayesian methods. The rate of detecting equivalence for the BF-N method is represented when $\varepsilon = 0$. The TOST and HDI methods are not represented when $\varepsilon = 0$ because these methods require an equivalence interval that is greater than 0.



probability of declaring the populations equivalent. The results for the 90% HDI and TOST procedures were practically identical in every sample size (displayed in separate panels) and equivalence interval (displayed along the x-axis of each panel). This is not surprising because, even though there are interpretational differences between the Bayesian and frequentist results, the procedures are numerically aligned. Along with the BF100 cutoff, they were also consistently the most conservative (i.e., the least likely to detect equivalence) of the tests used. However, the power of the BF100 cutoff increased faster relative to the HDI and TOST procedures as sample size increased; the BF100 cutoff was ultimately the lowest in power for detecting equivalence when sample size was low ($N \leq 100$), but the TOST and HDI methods were lower in power compared to the BF100 cutoff as sample size increased to $N = 1000$. Overall, the BF10 and BF30 cutoffs both detected equivalence more frequently than all other methods in every case.

In general, a large equivalence interval was needed to detect equivalence when the sample size was low. In fact, with $N \leq 100$, every method had power rates for detecting equivalence of 0 with an equivalence interval based on an effect size below $d = 0.2$. Even when sample size was increased to $N = 200$, a BF cutoff of 10 was the only measure to detect equivalence with an equivalence interval set at $(-0.2, 0.2)$. The BF with a nil hypothesis was unable to detect equivalence at any sample size, highlighting the potential limitations of using the overly strict nil hypothesis.

Figure 2 ■ Probability of concluding a lack of correlation with each of the equivalence testing and Bayesian methods. The rate of detecting equivalence for the BF-N method is represented when $\varepsilon = 0$. The TOST and HDI methods are not represented when $\varepsilon = 0$ because these methods require an equivalence interval that is greater than 0.

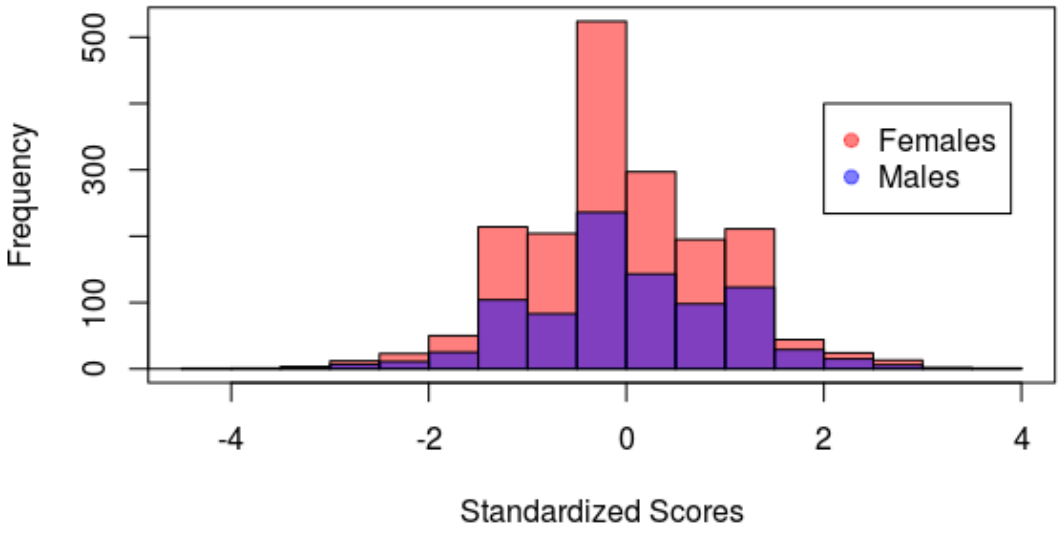


Negligible Correlation Simulation

Power results for the TOST, BF approaches and HDI method are displayed in Figure 2, where here the y-axis represents the probability of declaring the relationship between the two variables negligible. Results for the lack of correlation simulations were very similar to the lack of mean difference simulation results reported above. Namely, the HDI and TOST procedures displayed practically identical equivalence detection rates in all conditions (i.e., all sample size and equivalence interval conditions). Along with the BF100 cutoff, they were also again the most conservative methods. As seen in the *t*-test simulation results, the power to detect equivalence of the BF method increased more relative to the HDI and TOST procedures as sample size increased; the BF100 cutoff was lower in power compared to the HDI and TOST methods when sample size was low ($N \leq 100$), but was higher in power when sample size increased to $N = 1000$. Again, the BF10 and BF30 cutoffs both detected equivalence more frequently than all other methods.

The most significant departure from the mean difference results was that the BF with a nil hypothesis was able to detect a lack of correlation when $N \geq 200$. However, this was only the case when using a BF cutoff of 10. Even when the sample size was $N = 1000$, the BF30 cutoff was unable to detect a lack of correlation using a nil hypothesis. The TOST, HDI, and BF100 methods were all unable to detect negligible correlation with a tighter equivalence interval than $(-.15, .15)$ when the sample size was $N \leq 200$. The BF30 method had reasonable power to detect negligible correlation with an equivalence interval of $(-.15, .15)$

Figure 3 ■ Overlapping histogram of males and females in scores of conscientiousness.



or higher when $N \geq 100$, and an equivalence interval of $(-.1, .1)$ or higher when $N \geq 200$.

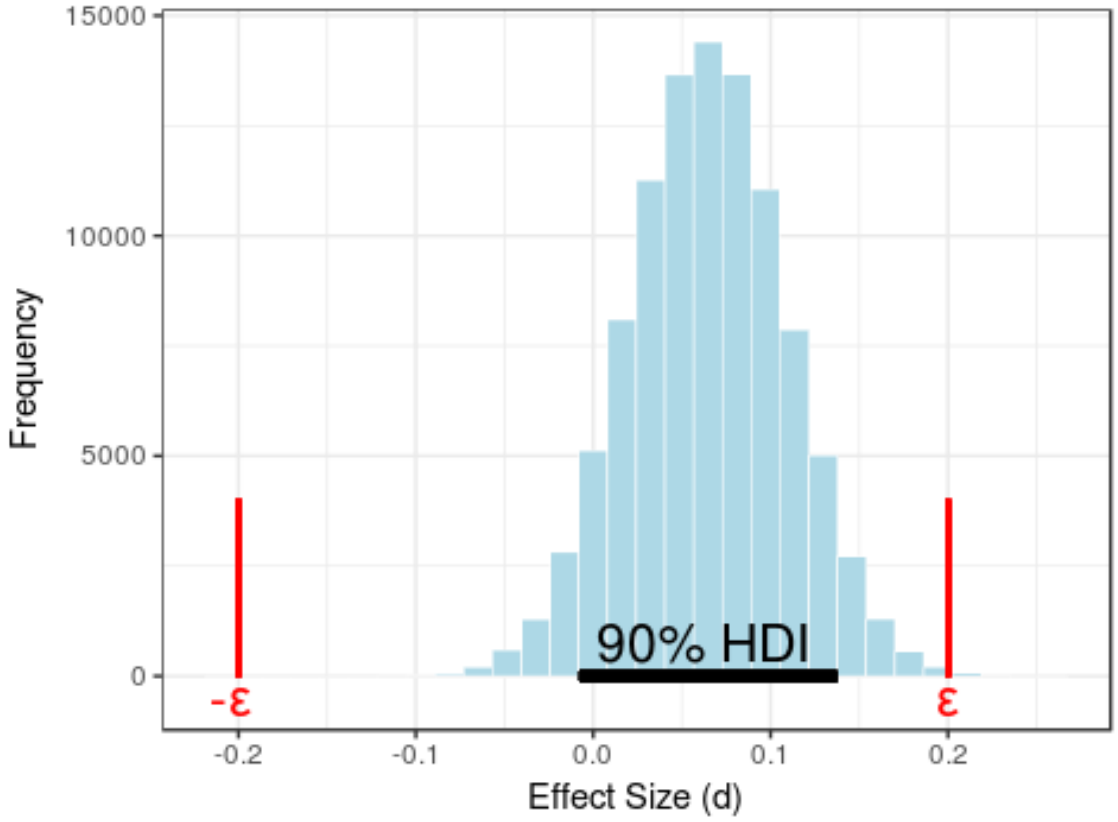
Application with Empirical Data

To demonstrate the TOST and the Bayesian HDI and BF approaches for detecting the equivalence of two population means and negligible correlation, we performed an analysis on the Big Five Inventory (BFI) dataset available in the R `psych` package (Revelle, 2016). The BFI dataset is based on a collection of 25 personality self-report items taken from the International Personality Item Pool that measure levels of the Big Five personality traits: Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion (Goldberg, 1999). The survey items are based on a 6-point response scale, from 1 (*Very Inaccurate*) to 6 (*Very Accurate*). The dataset includes the responses of 2800 subjects taken from the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project (Revelle, Wilt, & Rosenthal, 2010).

To demonstrate testing for mean equivalence, the scores of males and females in levels of conscientiousness were compared to test the research hypothesis that they are equivalent. This hypothesis follows up on a recent paper that found no significant difference between genders in levels of conscientiousness (Lehmann, Denissen, Allemand, & Penke, 2013). After excluding all incomplete cases, the final sample size was 2707 subjects, with 888 male participants ($M_{age} = 28.0$, $SD_{age} = 11.0$) and 1819 female participants ($M_{age} = 29.1$, $SD_{age} = 11.1$). All procedures use an equivalence interval based on a small value of Cohen's d $(-.2, .2)$, except for the BF-N method, which uses a nil hypothesis. The posterior distribution for the HDI was generated with the `BEST` package (Kruschke & Meredith, 2015), using the same noninformative priors as used in the simulation study described above.

To demonstrate testing for a lack of correlation, the association between age and agreeableness was analyzed to test the hypothesis that personality traits remain stable in fully developed adults. The choice to measure levels of agreeableness was based on a recent longitudinal study, which found that agreeableness remains stable among middle-aged cohorts (Wortman, Lucas, & Donnellan, 2012). To this end, only participants between the ages of 30 and 60 were included. After excluding participants outside of this age range or with incomplete answers, the final sample size was $N = 978$. All procedures use an equivalence interval based on a Pearson's r set at $(-.1, .1)$, except for the BF method that uses a nil hypothesis (BF-N). The posterior distribution was generated with the `BayesianFirstAid` package (Bååth et al., 2014), us-

Figure 4 ■ Posterior distribution for the standardized effect size with a 90% HDI. The red vertical lines indicate the bounds of the equivalence interval $(-0.2, 0.2)$.



ing the same approach as in the simulation study (e.g., noninformative priors, default iterations). The R (Revelle, 2016) code for conducting these analyses can be found at https://osf.io/tfdxq/.

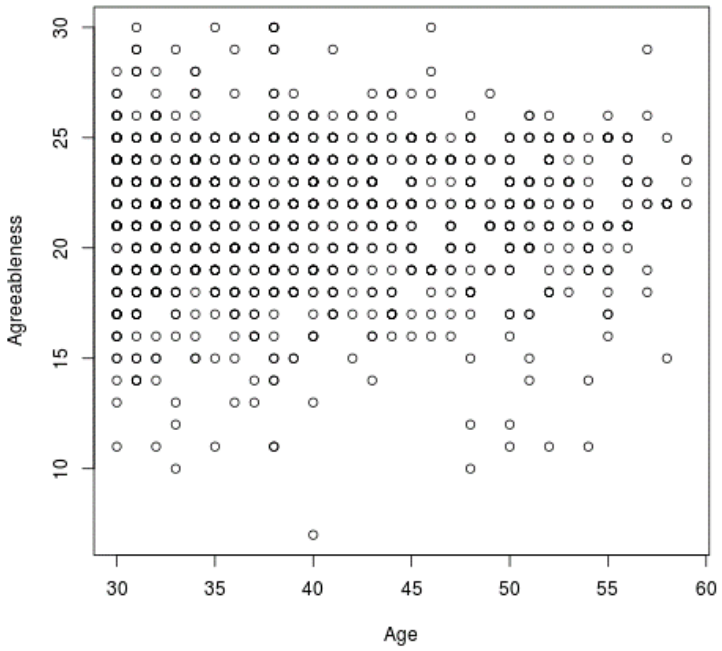
Equivalence of Conscientiousness by Gender

Conscientiousness scores were standardized so that an equivalence interval based on Cohen's d could be used with all tests. Figure 3 shows an overlapping histogram of the standardized scores. The TOST-SW variant of the TOST method was used, which adjusts for unequal population variances. Using the TOST-SW method, the null hypothesis that the difference between the scores of male ($M_M = 0.03$, $SD_M = 1.05$) and female ($M_F = -0.02$, $SD_F = 0.97$) participants in levels of conscientiousness is greater or equal to $d = .2$ (or less than or equal to $d = -.2$) can be rejected, $t_1(1648.6) = -1.262$, $p_1 = 0.01$; $t_2(1648.6) = 3.501$, $p_2 < 0.01$.

Using a BF approach with a nil hypothesis resulted in a BF of 16.13. This means that the null model, which predicts that there is no difference, is 16.13 times more likely than the alternative model, which predicts that there is a difference. In terms of the BF cutoffs used in the simulation study, this indicates that only the BF10 cutoff would conclude that the scores of male and females are equivalent. However, when an equivalence interval is incorporated into the test based on bounds of $(-.2, .2)$, the calculated BF rises to 77546. Thus, every level of the BF cutoffs used in the simulation study (10, 30 and 100) would detect equivalence. This drastic jump in the BF highlights once again the theoretical importance of including equivalence intervals in the BF approach when it is warranted.

Using Bayesian estimation, the mean of credible parameter values for μ_1 is 19.2, with a 90% HDI from 19.1 to 19.3, and the mean for μ_2 is 19.0, with a 90% HDI from 18.9 to 19.1. As seen in Figure 4, the 90% HDI for $\mu_1 - \mu_2$ $(-0.0077, 0.137)$ is fully within the equivalence interval (-.2, .2), and thus there is support for the hypothesis that there is no difference between males and females in scores of conscientiousness.

Figure 5 ■ Scatterplot of the variables age and agreeableness.



Lack of Correlation between Age and Agreeableness

Figure 5 shows that there does not appear to be any non-linear relationship between age and agreeableness. In addition, it suggests that the assumption of homoscedasticity is valid and that there are no multivariate outliers. The correlation coefficient was $r = .043$, 95% CI $[-0.02, 0.11]$.

Using the correlation variant of the TOST to check for a lack of association, the null hypothesis that the correlation between age ($M = 40.08$, $SD = 7.75$) and agreeableness ($M = 21.45$, $SD = 3.34$) falls outside the equivalence interval can be rejected, $t_1(976) = -1.775$, $p_1 = .038$; $t_2(976) = 4.479$, $p_2 < 0.001$.

Using a BF approach with a nil hypothesis results in a calculated BF of 10.04, indicating that the null model, which assumes that there is no correlation, is 10.04 times more likely than the alternative model, which assumes that there is a correlation. In terms of the BF cutoffs used in the simulation study, this indicates that only the BF10 cutoff would conclude a lack of correlation. However, when an equivalence interval is incorporated $(-1, .1)$, the calculated BF rises to 235. Thus, every level of the BF cutoffs used in the simulation study (10, 30 and 100) would conclude equivalence. As with the mean difference example, this highlights the substantial increase in power that comes from using an equivalence interval in BF calculations.

Using the Bayesian estimation approach, the mean of credible parameter values for ρ is 0.044, with a 90% HDI from $(-0.008$ to $0.099)$. As seen in Figure 6, the 90% HDI is within the equivalence interval $(-0.1, 0.1)$. Thus, there is support for the hypothesis that there is no association between age and levels of agreeableness.

Discussion

This paper aims to serve as a guide for researchers who wish to explore and compare the benefits and differences between the TOST procedure and two Bayesian alternatives for assessing a lack of association (e.g., lack of mean difference, negligible correlation). To this end, a simulation study was performed to assist researchers in understanding how to quantify and compare the power of these procedures. Specifically, the results serve to illustrate how these competing alternatives differ under various testing conditions. Next, these methods were applied to a dataset to serve as a demonstration of how Bayesian analysis can be used to test for a lack of mean difference and negligible

correlation.

The simulation study revealed important practical differences between the three approaches examined in this paper. Power differences between the tests remained relatively consistent across both the lack of mean difference and negligible correlation situations. However, it is interesting to note that in most cases either a large sample size or large equivalence interval is required to detect a lack of association. This highlights the inherent difficulty in successfully demonstrating a lack of association with limited data, regardless of the method being used. On the other hand, recent research (Rusticus & Eva, 2016) has highlighted that the smallest meaningful association among variables might actually be larger than what has traditionally been accepted as such (e.g., $d = .2$, $r = .1$); this research has important implications for setting appropriate equivalence intervals.

In the simulation study, the BF100 cutoff had a power of 0 in all cases when sample size was low ($N \leq 100$). However, it is important to note that the BF100 cutoff was more powerful in both the mean difference and correlation simulations than the TOST and HDI methods when sample size was $N = 1000$. This suggests that sample size affects the power of these methods differently, although further study is required to make any definitive statements. A related note is that the power for detecting equivalence is also affected by setting the scale of the alternative hypothesis (BF prior). We only present results for the default scaling factor of 1, since we believe it is important when testing for equivalence to have sufficient separability between the null and alternative distributions. Setting the scaling factor below 1 results in an alternative distribution that contains a nontrivial probability that the parameter falls near 0 (i.e., contains substantial overlap with the null distribution). More research and discussion is necessary regarding the most appropriate alternative distribution to use when utilizing BFs to test for equivalence.

Since noninformative priors were used, it is not surprising that the HDI and TOST results were similar. While the HDI and TOST methods displayed similar power across conditions, keep in mind that the HDI method is capable of incorporating informative priors and more logical statements can be made regarding the conclusions. With a large body of research informing a researcher's study, it may be wise to use informative priors based on previous research. If incorporating informative priors, the HDI method may be a better choice. It is also important to note that although the adopted TOST, HDI and BF approaches are all heteroscedastic in nature, the BF-EQ is not. Although this was not an issue with the simulation study or applications described in this paper, caution should be taken in interpreting the results of the BF-EQ approach when the variances are unequal.

Finally, both the simulation study and the demonstrations point to the importance of using equivalence intervals in psychological research. In the simulation study, the use of a nil hypothesis significantly reduced the power of the BF method for detecting equivalence. It was unable to detect equivalence in any of the mean difference simulation conditions, even when the sample size was over 1000, and was only able to detect a lack of correlation when sample sizes were large ($N \geq 200$) using a BF cutoff of 10. These power issues with the BF-N were further highlighted in the demonstrations. In both the mean difference and correlation demonstrations, the BF approach with an equivalence interval due to (Morey & Rouder, 2011) was drastically more powerful than the BF approach with a nil hypothesis developed by Jeffreys. This was particularly evident in the lack of mean difference example, where the BF jumps from 16.13 using a nil hypothesis to 77546 when using an equivalence interval.

The demonstrations using the BFI dataset also illustrate how easy it can be to apply Bayesian alternatives to the frequentist methods that psychologists often employ in their analyses. Although the TOST uses familiar NHST terminology that is familiar to most researchers, the results of the BF method are no more difficult to interpret. Likewise, results of the HDI method are easily understood by anyone with experience calculating confidence intervals. In the end, each method has unique properties that can make it a more pragmatic choice, depending on the situation. The BF is a good choice if researchers want to compare how much more likely the null is to the alternative (or vice versa), and the HDI method is a good choice if they want information about the full posterior distribution(s). The Bayesian methods are also recommended if the researchers have previous information that could be included in the priors. In fact, it is important for future research to explore the effect of informative priors on the procedures outlined in this paper. For instance, when testing for a lack of correlation, researchers might want to explore priors that give less weight to values that are closer to 1 or -1, as opposed to using a uniform distribution.

The demonstrations using the BFI dataset also illustrate how easy it can be to apply Bayesian alternatives to the frequentist methods that psychologists often employ in their analyses.

Overall, Bayesian methods provide useful alternatives

to traditional equivalence testing techniques like the TOST. With the power of modern computers, Bayesian techniques now serve as a practical alternative to frequentist approaches that take very little effort to employ, and which have unique benefits. Although we do not wish to argue that Bayesian methods are superior to frequentist approaches, or vice versa, for testing for mean equivalence or negligible correlation, we feel that knowledge of Bayesian methods should be more widespread in the realm of the behavioral sciences so that discussion on the relative merits of each method is encouraged.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, *311*, 485. doi:10.1136/bmj.311.7003.485

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics - Theory and Methods*, *12*(23), 2663–2692. doi:10.1080/03610928308828634

Bååth, R., Kruschke, J., & Meredith, M. (2014). Bayesian replacements for the most commonly used statistical tests in R (Version 0.1). Retrieved from https://github.com/rasmusab/bayesian_first_aid

Cribbie, R. A., Gruman, J., & Arpin-Cribbie, C. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*, 1–10. doi:10.1002/jclp.10217

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 527–537. doi:10.1348/000711009X475853

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in europe, vol. 7* (pp. 11–22). Tilburg, The Netherlands: Tilburg University Press.

Gruman, J., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods*, *6*, 133–140.

Jeffreys, H. (1961). *The theory of probability (3rd ed.)* Oxford, UK: Oxford University Press.

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

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

Kruschke, J., & Meredith, M. (2015). Best: Bayesian estimation supersedes the t-test (Version 0.4). Retrieved from https://CRAN.R-project.org/package=BEST

Lehmann, R., Denissen, J. J. A., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. *Developmental Psychology*, *49*(2), 365–383. doi:10.1037/a0028277

Ly, A., Verhagen, A. J., & Wagenmakers, E.-j. (2016). Harold jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. doi:http://dx.doi.org/10.1016/j.jmp.2015.06.004

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. doi:10.1037/0022-006X.46.4.806

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. (2015). Bayesfactor: Computation of Bayes factors for common designs. *R package version 0. 9*, 12–4. Retrieved from http://bayesfactorpcl.r-forge.r-project.org/

Revelle, W. (2016). Psych: Procedures for personality and psychological research (R package) (Version 1.6.12). Retrieved from https://CRAN.R-project.org/package=psych

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In M. A., G., & B. Szymura (Eds.), *Gruszka* (pp. 27–49). Handbook of individual differences in cognition . New York, NY: Springer.

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. doi:10.1037/0033-2909.113.3.553

Rusticus, S. A., & Eva, K. W. (2016). Defining equivalence in medical education evaluation and 20 research: Does a distribution-based approach work? *Advances in Health Sciences Education, 21*, 359–373. doi:10.1007/s10459-015-9633-x

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.

Thompson, K. (2015). Comparing the psychosocial health of tattooed and non-tattooed women. *Personality and Individual Differences*, *74*, 122–126. doi:10.1016/j.paid.2014.10.010

van Wieringen, K., & Cribbie, R. A. (2014). Robust normative comparison tests for evaluating clinical significance. *British Journal of Mathematical and Statistical Psychology*, *67*, 213–230. doi:10.1111/bmsp.12015

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority, second edition*. New York, NY: CRC Press.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741–744.

Wortman, J., Lucas, R. E., & Donnellan, M. B. (2012). Stability and change in the Big Five personality domains: Evidence from a longitudinal study of australians. *Psychology and Aging*, *27*(4), 867–874. doi:10 . 1037 / a0029322

Citation

Hoyda, J. J., Counsell, A., & Cribbie, R. A. (2019). Traditional and bayesian approaches for testing mean equivalence and a lack of association. *The Quantitative Methods for Psychology*, *15*(1), 12–24. doi:10.20982/tqmp.15.1.p012

Received: 28/10/2017 ∼ Accepted: 05/01/2019