ORIGINAL RESEARCH PAPER

Equivalence testing for standardized effect sizes in linear regression

Harlan Campbell

University of British Columbia, Department of Statistics Vancouver, British Columbia, Canada harlan.campbell@stat.ubc.ca

ARTICLE HISTORY

Compiled March 8, 2022

Abstract

We introduce equivalence testing procedures for standardized effect sizes in a linear regression analysis. Such tests are useful for confirming the lack of a meaningful association between a continuous outcome and continuous/binary covariates and may be particularly valuable if the covariates of interest are measured on different and somewhat arbitrary scales. We consider how to define valid hypotheses and how to calculate *p*-values for these equivalence tests. Via simulation study we examine type I error rates and statistical power; and we compare the proposed frequentist equivalence testing to an alternative Bayesian testing approach.

KEYWORDS

equivalence testing, non-inferiority testing, linear regression, standardized effect sizes

Thank you to Prof. Paul Gustafson for the helpful advice with preliminary drafts and thank you also to Prof. Daniël Lakens and Prof. Ken Kelley for feedback and insights.

1. Introduction

All too often researchers will conclude that the effect of an explanatory variable, X, on an outcome variable, Y, is absent when a null-hypothesis significance test (NHST) yields a non-significant p-value (e.g., when the p-value > 0.05). Unfortunately, such a procedure is logically flawed. As the saying goes, "absence of evidence is not evidence of absence" (Hartung et al., 1983; Altman & Bland, 1995). Indeed, a non-significant result can instead be due to insufficient statistical power, and while a NHST can provide evidence to reject the null hypothesis, it cannot provide evidence to accept the null.

To properly conclude that an association between X and Y is absent or at most negligible (i.e., to confirm the lack of an association), the recommended frequentist tool, the equivalence test (also known as the "non-inferiority test" for one-sided testing), is well-suited (Wellek, 2010). Let θ be the parameter of interest representing the association between X and Y. An equivalence test reverses the question that is asked in a NHST. Instead of asking whether we can reject the null hypothesis of no effect, i.e., reject $H_0: \theta = 0$, an equivalence test examines whether the magnitude of θ is at all meaningful by asking: Can we reject the possibility that θ is as large or larger than our smallest effect size of interest, Δ ? The null hypothesis for an equivalence test can therefore be defined as $H_0: \theta \notin (-\Delta, \Delta)$. In other words, equivalence implies that θ is small enough that any non-zero effect would be at most equal to Δ . To be clear, the interval $(-\Delta, \Delta)$, known as the "equivalence margin," represents the range of values for which θ can be considered negligible.

In psychology research and in the social sciences more broadly, the practice of equivalence testing is relatively new but "rapidly expanding" (Koh & Cribbie, 2013). Equivalence testing is now being used across all types of applied psychological research. Recent examples include Fruehauf et al. (2021) who use equivalence testing to study cognitive control in obsessive-compulsive disorder, and Leonidaki & Constantinou (2021) who use equivalence testing in a study of cognitive behavioural therapy. There are also many papers in the psychological methods literature which consider the merits of equivalence testing. For example, Goertzen & Cribbie (2010) highlight the

importance of using equivalence tests to establish the independence of two different variables and, more specifically, Ball et al. (2013) consider how equivalence tests may be useful for effectively testing the so-called "gender similarities hypothesis." Anderson & Maxwell (2016) recommend that psychology researchers use equivalence tests to adequately support claims of null effects in replication studies.

However, despite become increasingly common, equivalence testing remains challenging for many researchers. Specifically, defining and justifying the equivalence margin is cited as one of the "most difficult issues" (Hung et al., 2005). Lakens et al. (2018) provide some guidance for using equivalence tests in psychological research but note that defining the margin will be the "biggest challenge for researchers" because psychological theories are often "too vague." If the equivalence margin is too large, any claim of equivalence will be considered meaningless. On the other hand, if the margin is somehow too small, the probability of declaring equivalence will be substantially reduced (Wiens, 2002). While the margin is ideally based on some objective criteria, these can be difficult to justify, and there is generally no clear consensus among stakeholders (Keefe et al., 2013; Campbell & Gustafson, 2021b).

Scores from many psychological measures/scales are interpretable and meaningful, and researchers should, whenever possible, use validated and well-scaled measures where the units of measurement are well understood. However, in certain scenarios, the parameters of interest are measured on different and somewhat arbitrary scales. This makes the task of defining the equivalence margin more challenging. Without units of measurement that are easy to interpret, defining and justifying an appropriate equivalence margin can be all but impossible (Lakens et al., 2018).

When working with parameters measured on arbitrary scales, researchers will often prefer to work with standardized effect sizes to aid with interpretation (Baguley, 2009). It therefore stands to reason that, for equivalence testing in such a situation, it would also be preferable to define the equivalence margin in terms of a standardized effect size. Unfortunately, equivalence testing with standardized effects is not always straightforward. Contrary to certain recommendations, one cannot merely define the equivalence margin in terms of a standardized effect size and proceed as normal. For ex-

ample, as we will demonstrate later, Lakens (2017)'s suggestion that, for a two-sample test for the equivalence in means, one may simply define the equivalence margin in terms of the observed standard deviation is incorrect. The equivalence margin cannot be defined as a function of the observed data as this will invalidate the test. Instead, one must define the parameter of interest to be the standardized parameter, such that the randomness associated with standardization is properly taken into account.

For linear regression analyses, reporting standardized regression coefficients is quite common (West et al., 2007; Bring, 1994) and can be achieved by normalizing the outcome variable and all covariates before fitting the regression. However, there are no established equivalence tests for standardized regression coefficients. Therefore, our objective in this paper is to establish equivalence testing procedures for standardized effect sizes in a linear regression.

We note that equivalence tests for correlations (such as those proposed by Goertzen & Cribbie (2010)) can be used for testing standardized regression coefficients in a simple linear regression (i.e., when there is only 1 covariate) since the standardized regression coefficient is exactly equal to the correlation between the outcome variable and the single covariate. However, tests for correlations are inadequate for multivariable regression.

In the first section of this paper, we consider how to define valid hypotheses and calculate p-values for equivalence tests of standardized regression coefficients. We then conduct a small simulation study to better understand the operating characteristics of the tests and discuss different ways the tests could be used. We note that all the tests we propose in this paper are derived from inverting the noncentrality parameter confidence intervals put forward by Kelley (2007). Note that Kelley (2007) derived these confidence intervals by pivoting cumulative distribution functions.

Several Bayesian methods (e.g., the "default" Bayes factor (Rouder & Morey, 2012), and the Bayes factor interval null procedure (Morey & Rouder, 2011)) have also been proposed for establishing equivalence with standardized effect sizes. The pros and cons of frequentist versus Bayesian testing methods are a topic of great debate (e.g., Campbell & Gustafson (2021a)). While the focus of this paper is fre-

quentist equivalence testing, in the second half of the paper we briefly review one of the proposed Bayesian alternatives for establishing the equivalence of linear regression parameters. Following this, we demonstrate how all of the different testing methods can be applied in practice with several practical examples. We conclude with general recommendations on how to perform equivalence testing of standardized regression coefficients.

Equivalence testing for regression coefficients

An equivalence test for unstandardized regression coefficients

Consider a multivariable linear regression where Y is the outcome variable and X is the $N \times (K+1)$ fixed covariate matrix (with a column of 1s for the intercept); see Azen & Budescu (2009) for an accessible review. Going forward, we use the notation $X_{i\times}$ to refer to all K+1 values corresponding to the i-th observation; and X_k to refer to the k-th covariate.

Note that the regression may include both categorical and continuous covariates. For example, suppose a researcher is looking to investigate possible predictors of anxiety among high-school students. In this hypothetical study, Y might be a student's score on an anxiety assessment questionnaire; X_1 might be a binary variable indicating whether or not the student received counselling services (0 = "did not receive counselling; 1 = "did receive counselling"); X_2 might be a continuous covariate corresponding to the student's age in years; and X_3 might be a continuous covariate corresponding to the student's household income in dollars.

We operate under the standard linear regression assumption that the N observations in the data are independent and normally distributed such that:

$$Y_i \sim \text{Normal}(X_{i \times \beta}^T, \sigma^2), \quad \forall i = 1, ..., N;$$
 (1)

where $\beta = (\beta_0, \beta_1, \dots, \beta_K)^T$ is a parameter vector of K + 1 regression coefficients, and σ^2 is the population variance parameter (i.e., the variability of the random

errors). Least squares estimates for the linear regression model are denoted by $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_K})^T$, and $\widehat{\sigma}^2$; see equations (21) and (22) which are provided in the Appendix for completeness.

Recall that, for k in $1, \ldots, K$, the interpretation of the β_k coefficient is the average change in the response variable (Y) for every unit change in the explanatory variable (X_k) when holding all other explanatory variables constant. For example, in our hypothetical study about anxiety, the β_1 coefficient would be interpreted as the average number of additional points on the anxiety assessment score associated with a student receiving counselling services, the β_2 coefficient would be interpreted as the average number of additional points on the anxiety assessment score associated with every additional year in age, and the β_3 coefficient would be interpreted as the average number of additional points associated with every additional dollar in household income.

An equivalence test for an unstandardized regression coefficient asks the following question: Can we reject the possibility that β_k is as large or larger than our smallest effect size of interest? Formally, the null and alternative hypotheses for the equivalence test are stated as:

$$H_0: \beta_k \leq \Delta_{k,lower} \quad \text{or} \quad \beta_k \geq \Delta_{k,upper}, \quad \text{vs.}$$

$$H_1: \beta_k > \Delta_{k,lower} \quad \text{and} \quad \beta_k < \Delta_{k,upper},$$

$$(2)$$

where the equivalence margin, $(\Delta_{k,lower}, \Delta_{k,upper})$, defines the range of values considered negligible, for k in $0, \ldots, K$. Often, the equivalence margin will be symmetrical such that $\Delta_k = \Delta_{k,upper} = -\Delta_{k,lower}$, but this is not necessarily so. Also, in some scenarios, instead of a two-sided equivalence test, a one-sided equivalence test, known as a non-inferiority test, is required. A one-sided test can defined by simply setting the margin as a one sided-interval: $(-\infty, \Delta_{k,upper})$, or as $(\Delta_{k,lower}, \infty)$; see Wellek (2010).

Returning to our hypothetical example, suppose that in order for the impact of counselling services to be considered at all meaningful, the services would have to be associated with a minimum two point difference on the anxiety assessment questionnaire. In this case, the researcher would simply define $\Delta_{1,lower} = -2$ and $\Delta_{1,upper} = 2$. The equivalence margin for k = 1 would be (-2, 2). For the other covariates, k = 2 and k = 3, it may be more difficult to define an equivalence margin since β_2 and β_3 are measured in terms of "points per year" and "points per dollar". To define an appropriate margin, the researcher would have to ask: What are the minimum meaningful per year and per dollar numbers of points to consider?

Recall that there is a one-to-one correspondence between an equivalence test and a confidence interval (CI); see Dixon et al. (2018) for details. As such, an equivalence test can be constructed by simply inverting a confidence interval. For example, we will reject the above null hypothesis ($H_0: \beta_k \leq \Delta_{k,lower}$ or $\beta_k \geq \Delta_{k,upper}$), at a $\alpha = 0.05$ significance level, whenever a $(1 - 2\alpha) = 90\%$ CI for β_k fits entirely within $(\Delta_{k,lower}, \Delta_{k,upper})$. Inverting the CI for β_k leads to two one-sided t-tests (TOST) with the following p-values:

$$p_k^{lower} = 1 - F_t \left(\frac{\widehat{\beta_k} - \Delta_{k,lower}}{SE(\widehat{\beta_k})}, N - K - 1 \right), \quad \text{and} \quad p_k^{upper} = 1 - F_t \left(\frac{\Delta_{k,upper} - \widehat{\beta_k}}{SE(\widehat{\beta_k})}, N - K - 1 \right),$$
(3)

for k in 0,...,K; where $F_t(\cdot,\cdot;df)$ denotes the cumulative distribution function (cdf) of the t-distribution with df degrees of freedom, and where $SE(\widehat{\beta_k}) = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{kk}}$.

In order to reject the equivalence test null hypothesis $(H_0 : \beta_k \le \Delta_{k,lower})$ or $\beta_k \ge \Delta_{k,upper}$, both p-values, p_k^{lower} and p_k^{upper} , must be less than α . As such, for the k-th regression coefficient, β_k , a single overall p-value for the equivalence test can be calculated as: p-value $_k = \max(p_k^{lower}, p_k^{upper})$. In the Appendix, the "equivBeta" function is provided to conduct all the necessary calculations in R.

An equivalence test for standardized regression coefficients

It has been previously suggested that the bounds of an equivalence margin can be defined in terms of sample estimates; see Lakens (2017) and Lakens et al. (2018). This is incorrect. To illustrate why, suppose that, for a two-sample equivalence test for the difference in means, μ_d , one were to define a symmetric equivalence margin, $(-\Delta, \Delta)$, in terms of the observed standard deviation, $\hat{\sigma}$, such that $\Delta = 0.5 \times \hat{\sigma}$. Lakens et

al. (2018) consider this example and claim (incorrectly) that "when the equivalence bounds are based on standardized differences, the equivalence test depends on the standard deviation in the sample."

Recall that in order for a hypothesis test to be valid, the hypotheses must be statements about the unobserved parameters and not about the observed sample. Therefore, since the hypotheses for the test in the example, $H_0: |\mu_d| \geq 0.5 \times \hat{\sigma}$, vs. $H_1: |\mu_d| < 0.5 \times \hat{\sigma}$, are defined as functions of the observed data (i.e., in terms of $\hat{\sigma}$), the test is invalid.

Instead, the correct procedure is to define the parameter of interest, θ , to be the standardized effect size, e.g. define $\theta = \mu_d/\sigma$. Then, one can define the margin on the standardized scale without invalidating the hypotheses. To be clear, $H_0: |\theta| \geq 0.5$ vs. $H_1: |\theta| < 0.5$ is a completely valid test, while $H_0: |\mu_d| \geq 0.5 \times \hat{\sigma}$, vs. $H_1: |\mu_d| < 0.5 \times \hat{\sigma}$ is invalid. In this example, the valid equivalence test requires the use of a non-central t-distribution; see Appendix for details on how to conduct the valid test and Weber & Popova (2012) for a worked-through example. While in practice, the difference between setting $H_0: |\theta| \geq 0.5$ and $H_0: |\mu_d| \geq 0.5 \times \hat{\sigma}$ may be small, it should nevertheless be acknowledged since one should always (ideally) take into account the uncertainty involved in estimating the standard deviation. In the Appendix, we show results from a small simulation study (Simulation Study 2) which suggest that, in practice, using the invalid test can lead to a higher than advertised type 1 error when sample sizes are large, and a minor loss of efficiency when sample sizes are small.

Returning now to linear regression, in order to define the equivalence margin in terms of a standardized effect size, we will define the parameter of interest to be \mathcal{B}_k , the standardized regression coefficient. Standardizing a regression coefficient is done by multiplying the unstandardized regression coefficient, β_k , by the ratio of the standard deviation of X_k to the standard deviation of Y. The population standardized regression coefficient parameter, \mathcal{B}_k , for k in 1,...,K, is defined as:

$$\mathcal{B}_k = \beta_k \frac{s_k}{\sigma_Y},\tag{4}$$

where $\sigma_Y = (\beta^T \text{Cov}(X)\beta + \sigma^2)^{\frac{1}{2}}$ and s_k is the standard deviation of X_k . To be clear, the standardized regression coefficient, \mathcal{B}_k , is given in units of standard deviations of Y per standard deviation of X_k . The standardized regression coefficient can be estimated by:

$$\widehat{\mathcal{B}_k} = \widehat{\beta_k} \frac{s_k}{\widehat{\sigma_Y}},\tag{5}$$

where $\widehat{\sigma_Y}$ is the estimated standard deviation of the outcome variable. Note that, as is standard in the classical regression model, the covariates are assumed fixed leaving their standard deviations known and not needing to be estimated; see Azen & Budescu (2009). It is for this reason that s_k does not don a circumflex in the equation (5). Also note that in the psychometric literature, standardized regression coefficients are often known as Beta-coefficients, while the conventional unstandardized regression coefficients are often called B-coefficients.

How should researchers interpret the magnitude of standardized regression coefficients? There are no customarily cited references for guidance on this question. However, it is worth noting that when K = 1, the \mathcal{B}_1 parameter is exactly equal to the correlation between Y and X_1 , and Acock (2008) writes that standardized regression coefficients can be "interpreted similarly to how you interpret correlations in that $\mathcal{B} < 0.2$ is considered a weak effect, \mathcal{B} between 0.2 and 0.5 is considered a moderate effect, and $\mathcal{B} > 0.5$ is considered a strong effect." Also, when K = 1, and X_1 is a balanced binary variable, the standardized regression coefficient will be approximately equal in magnitude to half of the Cohen's d (note that the approximation will depend on the magnitude of the effect size, see Figure 1). As such, interpreting $\mathcal{B} = 0.10$ as small, $\mathcal{B} = 0.25$ as medium, and $\mathcal{B} = 0.40$ as large would be consistent with the habitually cited benchmarks for Cohen's d suggested by Cohen (1988).

Returning to the hypothetical anxiety study example, the interpretation of the standardized regression coefficients could be as follows. Suppose the scores obtained from students on the anxiety questionnaire range from 0 to 57 points with a standard deviation of 8 points. Furthermore, suppose the students in the study are from diverse

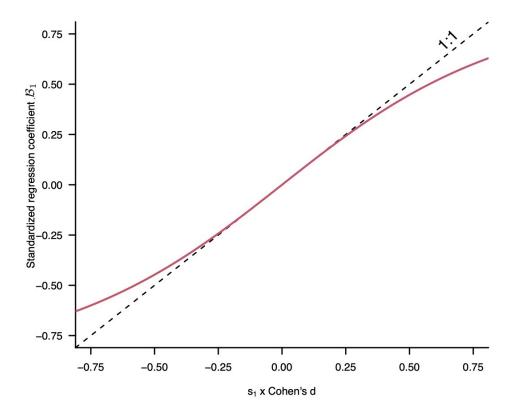


Figure 1. When K=1 and X_1 is a binary variable, the standardized regression coefficient, \mathcal{B}_1 , will be approximately equal in magnitude to the standard deviation of X_1 , s_1 , multiplied by Cohen's d. Indeed, $\mathcal{B}_1 = s_1(\beta_1/\sigma_Y)$, whereas Cohen's $d = (\beta_1/\sigma)$. The quality of the approximation will therefore depend on how well σ approximates $\sigma_Y = (\beta^T \operatorname{Cov}(X)\beta + \sigma^2)^{\frac{1}{2}}$. If X_1 is a balanced binary covariate, $s_1 = 0.5$ and as such, the standardized regression coefficient will be approximately equal in magnitude to half of the Cohen's d.

backgrounds such that there is a substantial variation their household incomes with a mean household income of \$47,100 and standard deviation of \$51,600 (as in Goodman et al. (2003)). Then, one could interpret $\widehat{\mathcal{B}}_3 = -0.10$ as follows. For every one additional standard deviation in household income (i.e., for every \$51,600 increase), the score on the anxiety assessment questionnaire would be, on average, one tenth of a standard deviation lower (i.e., would be about 0.8 points lower). This could be considered a weak effect. Now suppose that the students' ages ranged from 11 to 21 years old, with a standard deviation of 1.8 years (also as in Goodman et al. (2003)). Then one could interpret $\widehat{\mathcal{B}}_2 = 0.25$ as follows. For every one additional standard deviation in age, the score on the anxiety assessment questionnaire would be, on average, one quarter of a standard deviation higher (i.e., being 1.8 years older is associated with a

2 point increase). This could be considered a moderate effect. Finally, suppose half of the students have received counselling services $(X_1 = 1)$ and the other half have not $(X_1 = 0)$. As such, we have $s_1 = 0.5$ and one could interpret $\widehat{\mathcal{B}}_3 = -0.50$ as follows. Obtaining counselling services (coded as a difference in X_1 values equal to two standard deviations, $1-0=2\times s_1$) is associated with a decrease in the anxiety assessment score, on average, of one standard deviation. In other words, a student receiving counselling services has, on average, a score on the anxiety assessment questionnaire that is 8 points lower than a student who does not receive counselling services, given a constant household income and age. This could be considered a strong effect.

An equivalence test for \mathcal{B}_k can be defined by the following null and alternative hypotheses:

$$H_0: \mathcal{B}_k \leq \Delta_{k,lower}$$
 or: $\mathcal{B}_k \geq \Delta_{k,upper}$, vs. $H_1: \mathcal{B}_k > \Delta_{k,lower}$ and: $\mathcal{B}_k < \Delta_{k,upper}$,

where the equivalence margin is $(\Delta_{k,lower}, \Delta_{k,upper})$, for k in $1, \ldots, K$. We caution that, in certain scenarios, it may be challenging to define appropriate bounds on the standardized metric and that for well-understood scales, it may in fact be easier to define equivalence bounds on the original metric. However, when scales are not well-understood, defining equivalence bounds on the standardized metric may be preferable.

By inverting a confidence interval for \mathcal{B}_k (see Kelley (2007) for details), we obtain the following, for k in 1,...,K:

$$\mathbf{p}_{k}^{lower} = 1 - F_{t} \left(\frac{\widehat{\mathcal{B}_{k}}}{SE(\widehat{\mathcal{B}_{k}})}; df = N - K - 1, ncp = \Delta_{k,lower} \frac{\sqrt{N\left(1 - R_{X_{k}X_{-k}}^{2}\right)}}{\sqrt{1 - \left((1 - R_{X_{k}X_{-k}}^{2}\right)\Delta_{k,lower}^{2} + R_{YX_{-k}}^{2}\right)}} \right),$$

$$(6)$$

and:

$$p_k^{upper} = 1 - F_t \left(\frac{-\widehat{\mathcal{B}_k}}{SE(\widehat{\mathcal{B}_k})}; df = N - K - 1, ncp = -\Delta_{k,upper} \frac{\sqrt{N\left(1 - R_{X_k X_{-k}}^2\right)}}{\sqrt{1 - ((1 - R_{X_k X_{-k}}^2)\Delta_{k,upper}^2 + R_{Y X_{-k}}^2)}} \right),$$

with:

$$SE(\widehat{\mathcal{B}_k}) = \sqrt{\frac{(1 - R_{YX}^2)}{(1 - R_{X_k X_{-k}}^2)(N - K - 1)}},$$
 (7)

where R_{YX}^2 is the coefficient of determination from the linear regression of Y predicted from X; where $R_{X_kX_{-k}}^2$ is the coefficient of determination from the linear regression of X_k predicted from the remaining K-1 regressors; where $R_{YX_{-k}}^2$ is the coefficient of determination from the linear regression of Y predicted from all but the k-th covariate; and where $F_t(\ \cdot\ ; df, ncp)$ denotes the cdf of the non-central t-distribution with df degrees of freedom and noncentrality parameter ncp. (Note that when ncp = 0, the non-central t-distribution is equivalent to the central t-distribution. See Kelley (2007) for details on the use of non-centrality parameters.) For the k-th covariate, the null hypothesis, $H_0: \mathcal{B}_k \leq \Delta_{k,lower}$ or: $\mathcal{B}_k \geq \Delta_{k,upper}$, is rejected if and only if the p-value, p-value

In the Appendix, the "equivstandardBeta" function is provided to conduct all the necessary calculations in R. In addition, a Shiny app that allows one to conduct equivalence tests for the standardized regression coefficient in a simple linear regression model is available at https://hhappydog.shinyapps.io/std_beta/.

Under the assumption that $\mathcal{B}_k = 0$, for given values of N, K, and a symmetric equivalence margin of $(-\Delta_k, \Delta_k)$, a simple analytic formula can provide a reasonable approximation of the equivalence test's statistical power:

$$power = Pr(\text{reject } H_0 | \mathcal{B}_k = 0) \approx F_{half-t}(t^*; df = N - K - 1), \tag{8}$$

where $F_{half-t}(\ \cdot\ ;df)$ denotes the cdf of the half-t-distribution with df degrees of freedom; and where t^* is equal to the $(1-\alpha)\%$ critical value of a non-central t-distribution with df = N - K - 1 degrees of freedom, and noncentrality parameter $ncp = \Delta_k \frac{\sqrt{N(1-R_{X_kX_{-k}}^2)}}{\sqrt{1-((1-R_{X_kX_{-k}}^2)\Delta_k^2+R_{YX_{-k}}^2)}}$. Using this formula, certain simple sample size calculations for the equivalence test for \mathcal{B}_k can be carried out in a relatively straightforward manner. To be clear, the above formula may overestimate power in the case of

correlated covariates when sample sizes are relatively small (as we shall see in the simulation study). In the Appendix, the "powerestimate" function is provided to calculate the approximate power using equation (8) in R.

Simulation Study 1

We conducted a simple simulation study in order to better understand the operating characteristics of the proposed equivalence test for standardized regression coefficients. The equivalence test in the simulation study targeted the first standardized coefficient, \mathcal{B}_1 , and considered a symmetric equivalence margin, $(-\Delta_1, \Delta_1)$, such that the hypothesis test in question can be stated as:

$$H_0: |\mathcal{B}_1| \geq \Delta_1$$
, vs.

$$H_1: |\mathcal{B}_1| < \Delta_1.$$

The outcome variable was simulated from a Normal distribution such that $Y_i \sim \text{Normal}(X_{i\times}^T\beta,\sigma^2), \forall i=1,...,N$ with one of three values for the variance: $\sigma^2=0.05$, $\sigma^2=0.15$, or $\sigma^2=0.50$. Depending on the specific value of σ^2 , the true population standardized coefficient, \mathcal{B}_1 , for these simulated data is either: 0.070, 0.124, or 0.200. Parameters for the simulation study were chosen so that we obtain three unique values for \mathcal{B}_1 approximately evenly spaced between 0 and 0.20. These values correspond to relatively small effect sizes (Hemphill, 2003). In order to examine situations with $\mathcal{B}_1=0$, we also simulated data from additional scenarios where the regression coefficients were fixed such that $\mathcal{B}_1=\beta_1=0.00$. For these additional scenarios, σ^2 was set equal to 0.5.

Samples sizes were set between N=180 and N=3,500 in order to consider a wide range of values representative of sample sizes in large psychological studies (Kühberger et al., 2014; Fraley & Vazire, 2014)). Note that in many psychological studies, sample sizes can be much much smaller. Indeed, Marszalek et al. (2011) found a median sample size of N=40 in a representative survey of four top-tier psychology journals.

For each of the different configurations within the simulation study, we simulated 10,000 unique datasets and calculated an equivalence test p-value with each of 49 differ-

ent values of Δ_1 (ranging from 0.01 to 0.25). We then calculated the proportion of these p-values less than $\alpha=0.05$. We specifically chose to conduct 10,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a negligible amount (for looking at type 1 error with $\alpha=0.05$, Monte Carlo SE will be approximately $0.002\approx\sqrt{0.05(1-0.05)/10,000}$; see Morris et al. (2019)).

Simple linear regression

In the first part of the simulation study, we considered equivalence testing in a simple linear regression with K=1. The p-values were calculated with equation (6) (the "eq_B" test) and also, for comparison, with the equivalence test for correlations based on Fisher's Z transformation (the "eq_fz" test) proposed by Goertzen & Cribbie (2010) (see equation (27) and details in the Appendix). The single covariate was simulated from a standard Normal distribution and we set $\beta = (-0.2, 0.05)^T$. We simulated data for each of six scenarios (2 × 3), one for each combination of the following parameters:

- one of two sample sizes: N = 180, or N = 540; and
- one of three values for the variance: $\sigma^2 = 0.05$, $\sigma^2 = 0.15$, or $\sigma^2 = 0.50$,

and for an additional two scenarios with $\mathcal{B}_1 = 0$, $\sigma^2 = 0.50$, and either N = 180, or N = 540.

Multivariable linear regression

In the second part of the simulation study, we considered equivalence testing in a multivariable regression with either K=2 or K=4 covariates. The p-values were calculated using only equation (6). We simulated data for each of twenty-four scenarios $(4 \times 2 \times 3)$, one for each combination of the following parameters:

- one of four sample sizes: N = 180, N = 540, N = 1,000, or N = 3,500; and
- one of two orthogonal designs with K=2, or K=4 continuous covariates simulated from a standard Normal distribution; (with $\beta=(-0.20,0.10,0.20)^T$ for K=2, and $\beta=(0.20,0.10,0.14,-0.12,-0.08)^T$, for K=4); and

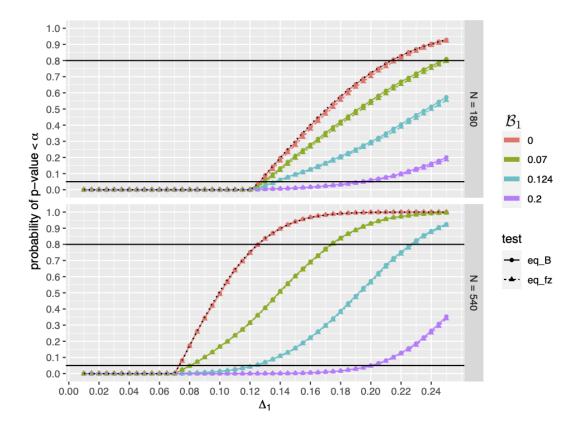


Figure 2. Simulation Study 1 - Results for simple linear regression with K=1. Upper panel shows results for N=180; Lower panel shows results for N=540. Two different equivalence tests were applied to each simulated dataset: (1) "eq.B" the proposed equivalence test for standardized regression coefficients (i.e., equation (6)); and (2) "eq.fz" the equivalence test for correlations based on Fisher's Z transformation proposed by Goertzen & Cribbie (2010) (see equation (27) and details in the Appendix). The lower solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$ and the upper solid horizontal black line indicates the reference power of 0.80. The curved dotted black lines correspond to numbers obtained using the proposed formula for approximate power calculation (equation (8)). Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

• one of three values for the variance: $\sigma^2=0.05$, $\sigma^2=0.15$, or $\sigma^2=0.50$, and an additional eight scenarios with $\mathcal{B}_1=0$ and $\sigma^2=0.50$.

Simulation study results

The simulation study was done using the R statistical software with default simulation routines (R Core Team, 2020). Figures 2 and 3 show results for the simple linear regression settings (i.e., for K=1). Figures 4 and 5 show results for the multivariable linear regression settings (i.e., for K=2 and K=4). Note that Figures 3 and 5 are "insets" (i.e., are "magnified portions") of Figures 2 and 4, respectively, to allow a focus on the type 1 error rate.

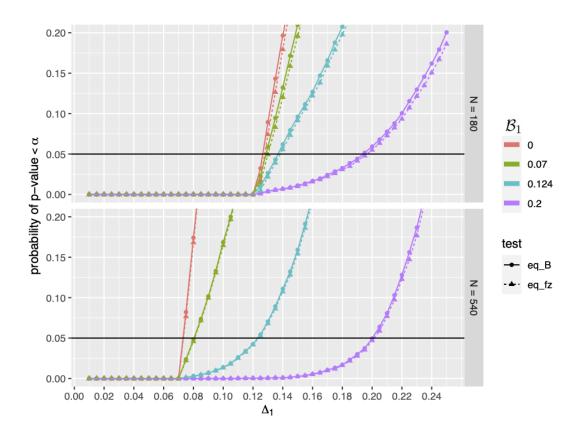


Figure 3. Simulation Study 1 - Note that this Figure is an "inset" (i.e., a "magnified portion") of Figure 2. Results for simple linear regression with K=1. Upper panel shows results for N=180; Lower panel shows results for N=540. Two different equivalence tests were applied to each simulated dataset: (1) "eq_B" the proposed equivalence test for standardized regression coefficients (i.e., equation (6)); and (2) "eq_fz" the equivalence test for correlations based on Fisher's Z transformation proposed by Goertzen & Cribbie (2010) (see equation (27) and details in the Appendix). Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$. Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

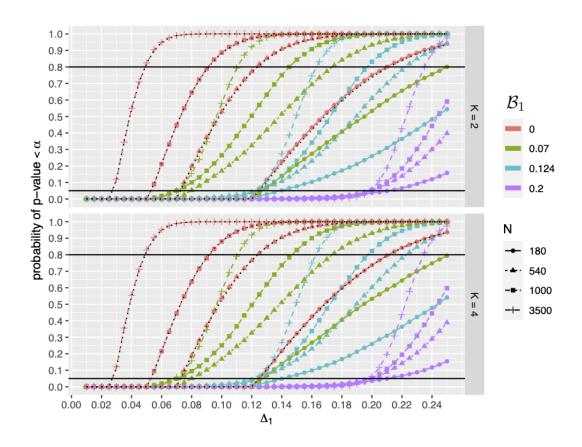


Figure 4. Simulation Study 1 - Upper panel shows results for K=2; Lower panel shows results for K=4. The proposed equivalence test for standardized regression coefficients (i.e., "eq_B", see equation (6)) was applied to each simulated dataset. The lower solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$ and the upper solid horizontal black line indicates the reference power of 0.80. The curved dotted black lines correspond to numbers obtained using the proposed formula for approximate power calculation (equation (8)). Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

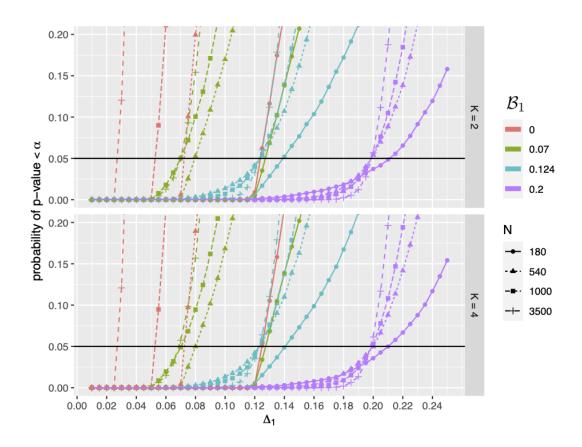


Figure 5. Simulation Study 1 - Note that this Figure is an "inset" (i.e., a "magnified portion") of Figure 4. Upper panel shows results for K=2; lower panel shows results for K=4. The proposed equivalence test for standardized regression coefficients (i.e., "eq_B", see equation (6)) was applied to each simulated dataset. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$. Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

First, let us briefly comment on how the proposed test standardized regression coefficients ("eq_B") and the test for correlations based on Fischer's Z transformation ("eq_fz") compare. For all scenarios, both tests are almost identical in terms of the probability of obtaining a p-value less than 0.05. However, the proposed test for standardized linear regression coefficients ("eq_B") appears to be marginally more efficient. In Figure 2, we see that the probability of obtaining a p-value less than 0.05 when $\Delta_1 > \mathcal{B}_1$ is slightly higher for the "eq_B" test than for the "eq_fz" test.

Now let us focus specifically on the results for the proposed test standardized regression coefficients ("eq.B"). Consider the results obtained when $\mathcal{B}_1 = 0.200$. We see that when the equivalence bound, Δ_1 , also equals 0.200 (i.e., when $\mathcal{B}_1 = \Delta_1 = 0.200$), the type 1 error rate is exactly 0.05, as it should be, when N = 540, 1,000, and 3,500. For N = 180, the test is somewhat conservative. This situation represents the boundary of the null hypothesis. As the equivalence bound increases beyond the true effect size (i.e., for $\Delta_1 > \mathcal{B}_1$), the alternative hypothesis is then true and it is more and more likely we will correctly conclude equivalence.

For smaller values of \mathcal{B}_1 (i.e., for $\mathcal{B}_1 = 0.070$ and $\mathcal{B}_1 = 0.124$), when the equivalence bound equals the true effect size (i.e., when $\mathcal{B}_1 = \Delta_1$), the test can be rather conservative, particularly for small N. Even when $\Delta_1 > \mathcal{B}_1$, the equivalence test may reject the null hypothesis for less than 5% of cases. This is due to the fact that with a small N, the sampling variance of $\widehat{\mathcal{B}}_1$ may be far too large to reject $H_0: |\mathcal{B}_1| \geq \Delta_1$. Consider that, when N is small and σ^2 is relatively large, the 90% CI for \mathcal{B}_1 may be far too wide to fit entirely within the equivalence margin.

In the Appendix (see Simulation Study 1 - alternative settings), we show results from an alternate version of the simulation study where we considered (correlated) binary covariates. The results obtained with the alternate settings are similar. In order for the test to have any substantial power, \mathcal{B}_1 must be substantially smaller than Δ_1 and limited sample sizes may prevent the proposed equivalence test from having sufficient power to rule out any truly "negligible" effect. As expected, the power of the test increases with larger values of N, smaller values of K, and larger values of Δ_1 .

In Figure 2 and Figure 4, we have overlaid a dotted black line corresponding to

numbers one obtains using the proposed formula for approximate power calculation (equation (8)). We note that the empirical results agree relatively well with the numbers obtained from equation (8). However, we observed that with certain alternative settings (correlated binary covariates), there is a discrepancy between the dotted black line and the solid red line for N = 180; see Figure 8 in the Appendix. This suggests that, when covariates are correlated and sample sizes are relatively small, the formula for approximating power (equation (8)) may overestimate the true power.

Finally, note that in some cases the equivalence test's power is strictly zero. For example, when $\mathcal{B}_1 = 0.00$, N = 180 and K = 2, the Δ must be greater or equal to 0.115 for there to be any possibility of rejecting H_0 . Otherwise, for $\Delta < 0.115$, the power is zero; see Figure 5. In summary, when using a $\alpha = 0.05$ maximum type I error rate, reasonable power for finding equivalence requires either large equivalence bounds, very large sample sizes, or both.

An non-inferiority test for the increase in \mathbb{R}^2

The increase in the squared multiple correlation coefficient associated with adding a variable in a linear regression model, $\operatorname{diff} R_k^2$, is a commonly used measure for establishing the importance of the added variable (Dudgeon, 2017). We define $\operatorname{diff} R_k^2 = R_{YX}^2 - R_{YX_{-k}}^2$.

In a linear regression model, the R_{YX}^2 is equal to the square of the Pearson correlation coefficient between the observed and predicted outcomes (Nagelkerke, 1991; Zou et al., 2003). Despite the R_{YX}^2 statistic's ubiquitous use, its corresponding population parameter, which we will denote as P_{YX}^2 , as in Cramer (1987), is rarely discussed. When considered, it is sometimes is known as the "parent multiple correlation coefficient" (Barten, 1962) or the "population proportion of variance accounted for" (Kelley, 2007).

When K=1 (i.e., for simple linear regression), note that: $\operatorname{diff} R_1^2 = R_{YX}^2 = \widehat{\mathcal{B}}_1^2$. To be clear, when K=1, the standardized regression coefficient, \mathcal{B}_1 , will be equal to ρ_{YX_1} , the correlation between the outcome variable, Y, and the single covariate, X_1 , and also equal to P_{YX} , such that:

$$\mathcal{B}_1 = \rho_{YX_1} = P_{YX}.\tag{9}$$

When K > 1, things are not as simple. In general, we have that the diff R_k^2 measure is a re-calibration of $\widehat{\mathcal{B}_k}$ (see Dudgeon (2017)), such that:

$$\operatorname{diff} R_k^2 = \widehat{\mathcal{B}}_k^2 (1 - R_{X_k X_{-k}}^2). \tag{10}$$

Similarly, we have that for the corresponding population parameter: $\operatorname{diff} P_k^2 = \mathcal{B}_k^2 (1 - P_{X_k X_{-k}}^2)$.

It may be preferable to consider an effect size (and what can be considered a "negligible difference") in terms of $diff P_k^2$ instead of in terms of \mathcal{B}_k . If this is the case, one can conduct a non-inferiority test (a one-sided equivalence test), for k in 1,...,K, with the following hypotheses:

$$H_0: 1 > \text{diff} P_k^2 \ge \Delta_k$$
, vs.

$$H_1: 0 \le \text{diff} P_k^2 < \Delta_k.$$

The *p*-value for the above non-inferiority test is obtained by simply replacing $\widehat{\mathcal{B}}_k$ with $\sqrt{\text{diff}R_k^2/(1-R_{X_kX_{-k}}^2)}$ and can be calculated, for fixed regressors as follows:

$$p - \text{value}_{k} = F_{t} \left(\frac{\sqrt{(N - K - 1) \text{diff} R_{k}^{2}}}{\sqrt{(1 - R_{YX}^{2})}}; \quad df = N - K - 1,$$

$$ncp = \frac{\sqrt{\Delta_{k} N \left(1 - R_{X_{k}X_{-k}}^{2} \right)}}{\sqrt{1 - \left((1 - R_{X_{k}X_{-k}}^{2}) \Delta_{k} + R_{YX_{-k}}^{2} \right)}} \right).$$
(11)

In related work, Campbell & Lakens (2021) introduce a non-inferiority test (a one-sided equivalence test) to test the null hypotheses:

$$H_0: 1 > P_{YX}^2 \ge \Delta$$
, vs.

$$H_1: 0 \le P_{YX}^2 < \Delta.$$

Note that when K=1, we have that: $\text{diff}P_1^2=P_{YX}^2=\mathcal{B}_1^2$, and assuming that the desired equivalence bounds used for the equivalence test of \mathcal{B}_1 are symmetric (i.e., assuming that $\Delta_{1,upper}=-\Delta_{1,lower}$), the three tests will be entirely equivalent.

Conditional equivalence testing

Ideally, a researcher uses an equivalence test to examine a preregistered hypothesis concerning the absence of a meaningful effect. However, in practice, one might first conduct a null hypothesis significance test (NHST) (i.e., calculate a p-value, p_{NHST} , using equation (23) or (24)) and only proceed to the equivalence test (i.e., calculate a second p-value, p_{EQUIV} , using equation (6)) if the NHST fails to reject the null. Such a two-stage sequential testing scheme has been discussed by Seaman & Serlin (1998) and more recently by Campbell & Gustafson (2018) under the name of "conditional equivalence testing" (CET).

Under the two-step CET scheme, if the first p-value, p_{NHST} , is less than the type 1 error α -threshold (e.g., if $p_{NHST} < 0.05$), one concludes with a "positive" finding: \mathcal{B}_k is significantly different than 0. On the other hand, if the first p-value, p_{NHST} , is greater than α and the second p-value, p_{EQUIV} , is smaller than α (e.g., if $p_{NHST} \geq 0.05$ and $p_{EQUIV} < 0.05$), one concludes with a "negative" finding: there is evidence of statistically significant equivalence, i.e., \mathcal{B}_k is at most negligible. If both p-values are larger than α , the result is inconclusive: there are insufficient data to support either finding.

Note that some researchers may wish to perform an equivalence test even if the initial NHST is "significant." For example, Burgoyne et al. (2020), in each of their analyses, conduct an NHST and then, regardless of the outcome, conduct an equivalence test. In this paper, we are not advocating for (or against) the CET approach, but simply use it to facilitate a comparison with Bayes factor testing which also categorizes outcomes as either positive, negative or inconclusive.

A Bayesian alternative for establishing equivalence in a linear regression

As discussed in the Introduction, there are a multitude of different Bayesian methods available for establishing equivalence. Rouder & Morey (2012)'s proposed "default" Bayes factors (based on the work of Liang et al. (2008)) is one approach that has proven to be particularly popular in psychology research for linear regression models (Etz, 2015; Morey et al., 2015). We briefly review the default Bayes factors approach for linear regression in order to consider how it might compare to the frequentist equivalence tests we proposed.

The Bayes Factor, BF_{10} , is defined as the probability of the data under the alternative model relative to the probability of the data under the null model:

$$BF_{10} = \frac{\Pr(Data \mid Model \ 1)}{\Pr(Data \mid Model \ 0)} = \frac{\Pr(Model \ 1 \mid Data) \times \Pr(Model \ 1)}{\Pr(Model \ 0 \mid Data) \times \Pr(Model \ 0)}, \tag{12}$$

with the "10" subscript indicating that the alternative model (i.e., "Model 1") is being compared to the null model (i.e., "Model 0"). Interpretation of the Bayes factor is straightforward. For example, a BF_{10} equal to 0.20 indicates that the null model is five times more likely than the alternative model.

Bayesian methods require one to define appropriate prior distributions for all model parameters (i.e., define $\Pr(Model\,0)$ and $\Pr(Model\,1)$) (Consonni & Veronese, 2008). Rouder & Morey (2012) suggest using a Jeffreys-Zellner-Siow (JZS) "objective prior" and provide and overview of its various advantages; see also Heck (2019). A scaled version of this prior, whereby the r scale parameter is set equal to a specific value allows one to specify prior beliefs about the expected standardized effect size in a rather straightforward manner. For instance, the BayesFactor R package uses the scaled-JZS prior with a default of $r=\sqrt{2}/4$, corresponding to a prior belief in a "medium effect size."

To test the k-th regression coefficient, in a multivariable linear regression model, one computes a Bayes factor for a model that includes the k-th covariate against a

model that does not, such that:

Model 0:
$$Y_i \sim \text{Normal}(X_{i,-k}^T \beta_{-k}, \sigma^2), \quad \forall i = 1, ..., N;$$
 (13)

Model 1:
$$Y_i \sim \text{Normal}(X_{i \times}^T \beta, \sigma^2), \quad \forall i = 1, ..., N;$$
 (14)

where β_{-k} ($X_{i,-k}$) is the vector (matrix) of regression coefficients (covariates), with the k-th coefficient (covariate) omitted. If this Bayes factor were to be above a certain threshold (e.g., if $BF_{10} > 6$), one would conclude that β_k is different than 0 (i.e., evidence in support of Model 1). On the other hand, if this Bayes factor were to be bellow a certain threshold (e.g., if $BF_{10} < 1/6$), one would conclude that there is evidence for $\beta_k = 0$ (i.e., evidence in support of Model 0).

A threshold of 3 (or 1/3) can be considered "moderate evidence," a threshold of 6 (or 1/6) can be considered "strong evidence," and a threshold of 10 (or 1/10) can be considered "very strong evidence" (Jeffreys, 1961). Simulation studies are recommended to calculate the statistical power of Bayes factor testing procedures; see Schönbrodt & Wagenmakers (2018).

Note that, just like frequentist testing of standardized regression coefficients, Bayes factor testing of regression coefficients is scale invariant so that the Bayes factor is entirely independent of the specific measurement units. To be clear, the Bayes factor will not change if the regression coefficients are measured in different units. Also note that some recent work (Tendeiro & Kiers, 2019) has argued that the posterior odds, rather than the Bayes factor, should be used for Bayesian testing. Rouder & Morey (2012) discuss this in a section of their paper entitled "Adding value through prior odds."

In the Appendix we conduct a small simulation study (see Simulation Study 3) to compare a CET frequentist testing scheme (based on NHST and equivalence testing of standardized regression coefficients) to the Bayesian testing approach based on Rouder & Morey (2012)'s default Bayes factors. The results of the simulation study suggest that, given the same data, both approaches will often arrive at the same overall conclusion (i.e., both approaches will obtain either a positive, negative or inconclusive

result). The level of agreement however is highly sensitive to the choice of equivalence margin and the choice of the Bayes factor evidence threshold. While we do not consider the impact of selecting different priors with the Bayes factors, it is reasonable to assume that the level of agreement between Bayes factor testing and frequentist testing will also be rather sensitive to the chosen priors, particularly when N is small; see Berger (1985).

In related work, Campbell & Gustafson (2021a) backwards engineer a Bayes factor testing scheme so that it closely matches the frequentist CET in terms of its operating characteristics. For a given sample size and a given prior, Campbell & Gustafson (2021a) show that the Bayesian testing approach can be made to be practically identical to the frequentist testing approach by simply selecting the appropriate BF threshold.

Practical Examples

Evidence for gender bias -or the lack thereof- in academic salaries

As a first example to illustrate the various testing methods, we turn to the "Salaries" dataset (from R CRAN package "car"; see Fox et al. (2012)). This dataset has been used as an example in other work: as an example for "anti-NHST" statistical inference in Briggs et al. (2019); and as an example for data visualization methods in Moon (2017) and Ghashim & Boily (2018).

The data consist of a sample of salaries of university professors collected during the 2008-2009 academic year. In addition to the posted salaries (a continuous variable, in \$US), the data includes 5 additional variables of interest: (1) sex (2 categories: (1) Female, (2) Male); (2) years since Ph.D. (continuous, in years); (3) years of service (continuous, in years); (4) discipline (2 categories: (1) theoretical, (2) applied). (5) academic rank (3 categories: (1) Asst. Prof., (2) Assoc. Prof., (3) Prof.).

The sample includes a total of N=397 observations with 358 observations from male professors and 39 observations from female professors. The minimum measured salary is \$57,800, the maximum is \$231,545, and the median salary is \$107,300. A

primary question of interest is whether there is a difference between the salary of a female professor and a male professor when accounting for possible observed confounders: rank, years since Ph.D., years of service, and discipline. The mean salary for male professors in the sample is \$115,090, while the mean salary for female professors in the sample is \$101,002. For illustration purposes, we consider both a simple linear regression (K = 1) (ignoring the confounders) and a multivariable linear regression (K = 6).

A simple linear regression

Consider a simple linear regression (i.e., $Y \sim \text{Normal}(\beta_0 + \beta_1 X_1, \sigma^2)$) for the association between salary (Y, measured in \$) and $\text{sex }(X_1, \text{ where "0" corresponds to "female," and "1" corresponds to "male."). Standard least squares estimation results in the following parameter estimates: <math>\widehat{\beta}_0 = 101002$, $\text{SE}(\widehat{\beta}_0) = 4809$, and $\widehat{\beta}_1 = 14088$, $\text{SE}(\widehat{\beta}_1) = 5065$ (see equation (21)); $\widehat{\sigma} = 30034.61$ (see equation (22)); $\widehat{\mathcal{B}}_1 = 0.14$ (see equation (5)), $\text{SE}(\widehat{\mathcal{B}}_1) = 0.05$ (see equation (7)); and $R_{YX}^2 = \text{diff} R_1^2 = 0.019$.

We can conduct an equivalence test to determine if the difference in salaries between male and female professors is at most no more than some negligible amount. Suppose that any difference of less than $\Delta = \$5,000$ is considered negligible. Then a p-value for the equivalence test, $H_0: |\beta_1| \geq 5000$ vs. $H_1: |\beta_1| < 5000$, can be calculated using equation (3). We obtain a p-value = 0.963 and therefore fail to reject the equivalence test null hypothesis.

If it were not possible to determine a specific number of dollars to be considered negligible, we could conduct an equivalence test for the standardized regression coefficient, \mathcal{B}_1 . Suppose we consider very small effect sizes to be negligible and define an equivalence margin as (-0.10, 0.10). Then we can calculate a p-value for $H_0: |\mathcal{B}_1| \geq 0.10$ vs. $H_1: |\mathcal{B}_1| < 0.10$ as per equation (6). We obtain

p-value = $max(\mathbf{p}_1^{lower}, \mathbf{p}_1^{upper}) = 0.780$, where:

$$p_1^{lower} = F_t \left(\frac{\widehat{\mathcal{B}}_k}{SE(\widehat{\mathcal{B}}_k)}; df = N - K - 1, ncp = \Delta_{1,lower} \frac{\sqrt{N\left(1 - R_{X_k X_{-k}}^2\right)}}{\sqrt{1 - \left(\left(1 - R_{X_k X_{-k}}^2\right)\Delta_{1,lower}^2 + R_{Y X_{-k}}^2\right)}} \right) \\
= F_t \left(\frac{0.14}{0.05}; df = 397 - 1 - 1, ncp = -0.10 \frac{\sqrt{397(1 - 0)}}{\sqrt{1 - \left(\left(1 - 0\right) \times 0.01 + 0\right)}} \right) \\
= F_t \left(2.78, df = 395, ncp = -2.00 \right) < 0.001 \\
< 0.001 \tag{15}$$

$$\mathfrak{p}_{1}^{upper} = F_{t} \left(\frac{-\widehat{\mathcal{B}_{k}}}{SE(\widehat{\mathcal{B}_{k}})}; df = N - K - 1, ncp = -\Delta_{1, upper} \frac{\sqrt{N\left(1 - R_{X_{k}X_{-k}}^{2}\right)}}{\sqrt{1 - \left((1 - R_{X_{k}X_{-k}}^{2}\right)\Delta_{1, upper}^{2} + R_{YX_{-k}}^{2}\right)}} \right)$$

$$= F_{t} \left(-2.78, df = 395, ncp = -2.00 \right) = 0.780.$$

$$= 0.780. \tag{16}$$

Alternatively, we could conduct an equivalence test for diff P_1^2 , the increase in the coefficient of determination attributable to including the sex variable in the model $(H_0: \text{diff} P_1^2 \geq \Delta \text{ vs. } H_1: \text{diff} P_1^2 < \Delta)$. Setting $\Delta = 0.10^2 = 0.01$, we obtain a p-value (as per equation (11)) identical to what we obtained with the previous test:

$$p - \text{value} = 1 - F_t \left(\frac{\sqrt{(N - K - 1) \text{diff} R_k^2}}{\sqrt{(1 - R_{YX}^2)}}; df = N - K - 1, ncp = \frac{\sqrt{N\Delta}}{\sqrt{\left(1 - \Delta + R_{X_k X_{-k}}^2\right)}} \right)$$

$$= 1 - F_t \left(\frac{\sqrt{(397 - 1 - 1)0.019}}{\sqrt{(1 - 0.019)}}; df = 397 - 1 - 1, ncp = \frac{\sqrt{397 \times 0.01}}{\sqrt{(1 - 0.01 + 0)}} \right)$$

$$= 0.780. \tag{17}$$

Bayes factors are easy to compute as well. With the BayesFactor package and the "regressionBF" function (with the default prior-scale $r = \sqrt{2}/4$), we obtain a $BF_{10} = \sqrt{2}/4$)

4.5 which suggests that the alternative model (i.e., the model with "sex" included) is about four and a half times more likely than the null model (i.e., the intercept only model). Note that we obtain the identical result using the "linearReg.R2stat" function. However, when using the "lmBF" function, we obtain a value of $BF_{10} = 6.2$ which suggests that the alternative model is about 6 times more likely than the null model. Both functions are comparing the two very same models so this result is somewhat surprising.¹

A multivariable linear regression

Now consider a multivariable linear regression model, with K=6:

$$Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \sigma^2), \tag{18}$$

where $X_1 = 0$ corresponds to "female," and $X_1 = 1$ corresponds to "male"; X_2 corresponds to years since Ph.D.; X_3 corresponds to years of service; $X_4 = 0$ corresponds to "theoretical," and $X_4 = 1$ corresponds to "applied"; and where $(X_5 = 0, X_6 = 0)$ corresponds to "Asst. Prof.", $(X_5 = 1, X_6 = 0)$ corresponds to "Assoc. Prof.", and $(X_5 = 0, X_6 = 1)$ corresponds to "Prof.".

Table 1 lists parameter estimates obtained by standard least squares estimation and Table 2 lists the p-values for each of the hypothesis tests we consider as well as Bayes factors. The Bayes factors are calculated using the "regressionBF" function from the BayesFactor package (with the default prior-scale $r = \sqrt{2}/4$).

We obtain a Bayes factor for k = 1 of $B_{10} = 1/3.9$, indicating only moderate evidence in favour of the null model. This corresponds to an "inconclusive" result with a Bayes factor threshold of 6, or 10 (or any threshold higher than 3.9 for that matter). The result for k = 1 from a CET would also be "inconclusive" (for $\alpha = 0.05$

¹The apparent contradiction can be explained by the fact that the two "default BF" functions are using different "default priors." The "regressionBF" function (as we are using it, see Appendix) assumes "sex" is a continuous variable, while the "lmBF" function assumes that "sex" is a categorical variable. The "default priors" are defined accordingly, in different ways. This may strike one as rather odd, since both models are numerically identical. However, others see logic in such practice: Rouder & Morey (2012) suggest, somewhat vaguely, that researchers "be mindful of some differences when considering categorical and continuous covariates" and "recommend that researchers choose priors based on whether the covariate is categorical or continuous"; see Section 13 of Rouder & Morey (2012) for details.

and $\Delta=0.10$), since both the NHST p-value (= 0.216) and the equivalence test p-value (= 0.076) are larger than $\alpha=0.05$. As such, we conclude that, when controlling for observed confounders, there are insufficient data to support either an association, or the lack of an association, between sex and salary. More data will be required to answer the question. This inconclusive result might motivate researchers to undertake another study on the question with a much larger sample size.

Note that the conclusions obtained with the frequentist and Bayesian approaches do not entirely agree for the other covariates, see Table 2. For both the "years since Ph.D" (k=2) and the "years of service" (k=3) covariates, the frequentist CET obtains a positive result whereas the Bayes factor obtains an inconclusive result.

\overline{k}	covariate	β_k	$SE(\hat{\beta}_k)$	\mathcal{B}_k	$SE(\widehat{\mathcal{B}_k})$
0	intercept	65955.23	4588.60	-	-
1	sex (male)	4783.49	3858.67	0.047	0.038
2	years since Ph.D.	535.06	240.99	0.228	0.103
3	years of service	-489.52	211.94	-0.210	0.091
4	discipline (applied)	14417.63	2342.88	0.237	0.039
5	rank (Asst. Prof.)	12907.59	4145.28	0.157	0.050
6	rank (Prof.)	45066.00	4237.52	0.700	0.066
			$\hat{\sigma} = 22538.65$		$R_{Y,X}^2 = 0.455$

Table 1. Parameter estimates obtained by standard least squares estimation for the full multivariable linear regression model.

k	\mathcal{B}_k	p_{NHST}	p_{EQUIV}	BF_{10}	CET conclusion	Bayesian conclusion
			$\Delta = 0.10$	$r = \sqrt{2}/4$	$\alpha = 0.05$	BF threshold $= 6$
1	0.05	0.216	0.076	1/3.9	Inconclusive	Inconclusive
2	0.23	0.027	0.892	1.4	Positive	Inconclusive
3	-0.21	0.021	0.885	1.7	Positive	Inconclusive
4	0.24	< 0.001	1.000	6.5×10^{6}	Positive	Positive
5	0.16	0.002	0.868	13.6	Positive	Positive
6	0.70	< 0.001	1.000	1.8×10^{20}	Positive	Positive

Table 2. Calculated values and conclusions for both frequentist and Bayesian testing for the salaries multivariable linear regression model.

Six key premises of mindset theory

As a second practical example, we consider Burgoyne et al. (2020) who obtained data from N=438 individuals and fit several simple linear regressions to these data in order to investigate six key premises of "mindset theory." For each of the six key premises, Burgoyne et al. (2020) regressed a different continuous variable against an individual's

"mindset score" and used a non-inferiority test (i.e., a one-sided equivalence test) to determine whether the standardized regression coefficient was significantly smaller, or larger, than a predetermined value.

Specifically, Burgoyne et al. (2020) defined the non-inferiority margin as either -0.2 or 0.2 (depending on the direction of the effect predicted by mindset theory) and used the test for correlations proposed by Goertzen & Cribbie (2010) based on Fisher's Z transformation (see details in Appendix). The test for correlations is valid for testing standardized regression coefficients since the standardized regression coefficient is exactly equal to the correlation between the outcome and exposure variables in simple linear regressions (i.e., when K = 1); see equation (9).

We calculated p-values for each of the six regressions based instead on our proposed test for standardized regression coefficients (equation (6)) in order to see how these p-values would compare to the p-values obtained by Burgoyne et al. (2020). In Table 3, the p-values calculated based on equation (6) are listed alongside the p-values obtained by Burgoyne et al. (2020). We note that for each of the six simple linear regressions, the two p-values are very similar with the p-values obtained by Burgoyne et al. (2020) only marginally larger. This is consistent with the findings in our simulation study which suggested that our proposed test for standardized regression coefficients may be slightly more efficient (see Figure 2).

Burgoyne et al. (2020) also wished to investigate whether the association between the "Raven failure score" and the mindset score is no more than negligible when controlling for cognitive ability. This requires a multivariable linear regression and Goertzen & Cribbie (2010)'s test for correlations is therefore not applicable. Our proposed test for standardized regression coefficients is, on the other hand, well-suited for the task. In row 7 of Table 3, we list the p-value obtained using equation (6) as $p_{\mathcal{B}} < 0.001$.

	H_0	$Y \sim X$	$\widehat{\mathcal{B}}_1$	p_Z	$p_{\mathcal{B}}$
1.	$\mathcal{B}_1 \ge 0.2$	Learning goals			
		$\sim \text{Mindset}$	0.098	0.015	0.013
2.	$\mathcal{B}_1 \leq -0.2$	Performance goals			
		$\sim \text{Mindset}$	-0.109	0.026	0.024
3.	$\mathcal{B}_1 \leq -0.2$	Performance avoidance goals			
		$\sim Mindset$	-0.039	< 0.001	< 0.001
4.	$\mathcal{B}_1 \leq -0.2$	Belief in talent alone			
		$\sim Mindset$	-0.061	0.002	0.001
5.	$\mathcal{B}_1 \geq 0.2$	Response to challenge			
		$\sim \text{Mindset}$	0.056	0.001	0.001
6.	$\mathcal{B}_1 \ge 0.2$	Raven failure score			
		$\sim \text{Mindset}$	-0.122	< 0.001	< 0.001
7.	$\mathcal{B}_1 \ge 0.2$	Raven failure score			
		$\sim Mindset +$	-0.055		< 0.001
		Cognitive ability			

Table 3. For each of the regression analyses fit by Burgoyne et al. (2020), p_B indicates the p-value for the equivalence test based on equation (6), and p_Z indicates the p-value for the equivalence test based on Fisher's Z transformation. Note that in order to conduct a non-inferiority test (a one-sided equivalence test), one defines an open ended equivalence margin (which we indicate by setting either $\Delta_{lower} = -\infty$ or setting $\Delta_{upper} = \infty$).

The impact -or lack thereof- of exposure to organic food on moral judgments

As a third practical example, we consider data from a replication study. Eskine (2013) showed that participants who had been exposed to organic food were harsher in their moral judgments relative to those not exposed, and Moery & Calin-Jageman (2016) attempted to replicate this finding in a preregistered replication study.

Lakens (2017) uses data from Moery & Calin-Jageman (2016)'s replication study to demonstrate how one can conduct an equivalence test for the difference between two independent means. We will use the same data to demonstrate how, alternatively, one can conduct an equivalence test for the standardized difference between the two independent means, and also how one can conduct an equivalence test for a standardized regression coefficient.

The outcome variable under consideration is the number of points given by an individual on a 7-point scale assessing morality. The data consist of a total sample of N = 184 observations with $N_1 = 95$ observations from a control group for which the observed outcome has a mean of 5.25 points and standard deviation of $\hat{\sigma}_1 = 0.91$, and $N_2 = 89$ observations from a group exposed to organic food where the observed out-

come has a mean of 5.22 points and standard deviation of $\hat{\sigma}_2 = 0.83$. The pooled standard deviation estimate is $\hat{\sigma}_p = 0.871 = \sqrt{((N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2)/(N_1 + N_2 - 2)}$.

Lakens (2017) defines the null hypothesis for the equivalence test as:

$$H_0: \mu_d \leq -0.48 \times \hat{\sigma}_p$$
, or: $\mu_d \geq 0.48 \times \hat{\sigma}_p$,

where μ_d is the parameter representing the difference between the mean number of morality points among those in the control group and the mean number of morality points among those in the organic food group. Note that this is a theoretically invalid hypothesis since it is defined in terms of the observed value $\hat{\sigma}_p$. Regardless, using the equivalence test for the difference between two independent means to evaluate this invalid hypothesis, one obtains a p-value = $\max(p^{lower}, p^{upper})$ = 0.0014, where:

$$p^{lower} = 1 - F_t(3.48, 182) = 1 - F_t\left(\frac{5.25 - 5.22 + 0.418}{0.871\sqrt{1/95 + 1/89}}, 182\right) = 0.0003; \text{ and}$$

$$p^{upper} = 1 - F_t(-3.03, 182) = 1 - F_t\left(-\frac{5.25 - 5.22 - 0.418}{0.871\sqrt{1/95 + 1/89}}, 182\right) = 0.0014; \quad (19)$$

see Lakens (2017) for details.²

A valid null hypothesis can instead be defined in terms of $\theta = \mu_d/\sigma$, the standardized difference between the two independent means, such that:

$$H_0: \theta \le -0.48$$
, or: $\theta \ge 0.48$.

The corresponding p-value can then be calculated based on the non-central t-distribution as p-value= $\max(p_d^{lower}, p_d^{upper})$ =0.0012, where:

$$p_d^{lower} = 1 - F_t \left(\frac{5.25 - 5.22}{0.871} \sqrt{\frac{184 \times 89}{184 + 89}}, 184 + 89 - 2, 0.48 \sqrt{\frac{184 \times 89}{184 + 89}} \right) = 0.0003; \text{ and } p_d^{upper} = 1 - F_t \left(-\frac{5.25 - 5.22}{0.871} \sqrt{\frac{184 \times 89}{184 + 89}}, 184 + 89 - 2, -0.48 \sqrt{\frac{184 \times 89}{184 + 89}} \right) = 0.0012;$$

$$(20)$$

²Note that Lakens (2017)'s calculation of the estimated pooled standard deviation, $\hat{\sigma}_p$, appears slightly different (reported as $\hat{\sigma}_p = 0.894$, perhaps due to rounding error), and the equivalence bound appears to be miscalculated, reported as $\Delta = 0.384$ (which notably does not equal $0.48 \times 0.894 = 0.429$). In related material, Lakens (2022) presents this calculation differently and explains that: "the equivalence bound is [a Cohen's d of] d = 0.48, which equals a difference of 0.429 on a 7-point scale given the sample sizes and a pooled standard deviation of 0.894."

see equation (25) in the Appendix for details.

Alternatively, one could consider the same data within a simple linear regression model where a single exposure variable, X_1 , is defined as a binary indicator of whether an individual is in the control group $(X_1 = 0)$, or in the organic food group $(X_1 = 1)$, and the outcome, Y, is defined as the number of morality points. A valid null hypothesis can then be defined in terms of the standardized regression coefficient, \mathcal{B}_1 , such that:

$$H_0: \mathcal{B}_1 \le -0.48 \times s_1 = 0.241$$
 or: $\mathcal{B}_1 \ge 0.48 \times s_1 = 0.241$.

Note that this test is considering a slightly different hypothesis than previously since $\mathcal{B}_1 = 0.48 \times s_1$ is only approximately equivalent to $\theta = 0.48$ (see Figure 1). A *p*-value can be calculated using equation (6) as *p*-value = $max(\mathfrak{p}_1^{lower}, \mathfrak{p}_1^{upper}) = 0.0009$, where:

$$\mathbb{p}_1^{lower} = F_t (-0.224, df = 182, ncp = -3.361) = 0.0009;$$
 and $\mathbb{p}_1^{upper} = F_t (0.224, df = 182, ncp = -3.361) = 0.0002.$

Factors in hominid brain evolution

As a final example, we follow Rouder & Morey (2012) (and Heck (2019)) in reanalyzing a dataset first presented by Bailey & Geary (2009). The dataset consists of a information from a sample of hominid crania aged between 10 thousand and 1.9 million years. The sample includes a total of N=175 observations and the outcome of interest is the cranium capacity, a continuous variable measured in cm³, which ranges from 475cm^3 to 1.880cm^3 . We also consider 4 covariates of interest:

- (1) the local climate variation (X_1 : "local climate", in degrees Celsius ranging between 8 and 47), which is measured as the difference between the highest mean monthly high and the lowest mean monthly low temperature, in degrees Celsius, for a location near to where the fossil cranium was discovered;
- (2) the global average temperature (X_2 : "global climate", in standard deviations, ranging between 0.21 and 0.43), as indicated by the standard deviation of the $\delta^{18}O$ value (a metric calculated based on the record of oxygen isotopes) for the 200,000

years before the fossil was dated;

- (3) the parasite load, $(X_3 : \text{``parasites''}, \text{ integer between 0 and 7})$, which is measured as the total number of types of parasites considered to be potentially harmful to humans in the region where the fossil cranium was discovered; and
- (4) the population density, $(X_4 : \text{"pop. density"}, \text{ integer between 13 and 141})$ of the group the hominid lived within, as measured by "a population density proxy using the number of individuals living in surrounding areas during the hominid's ancestral history."

We fit the data with a multivariable linear regression: $Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4, \sigma^2)$. Table 4 lists parameter estimates obtained by standard least squares estimation and Table 5 lists the p-values for each of the hypothesis tests we consider as well as Bayes factors. The Bayes factors are calculated using the "regressionBF" function from the BayesFactor package (with the default prior-scale $r = \sqrt{2}/4$). Note that the Bayes factors we obtain are rather different than those obtained by Rouder & Morey (2012). This is due to the fact that Rouder & Morey (2012) use a prior-scale of r = 1, whereas we use $r = \sqrt{2}/4$. With a prior-scale of r = 1, one obtains Bayes factors of 1/12.9, 9.4×10^7 , 1/4.4, and 6.3×10^{13} , for $k = 1, \ldots, 4$, respectively.³

With $\alpha=0.05$ and a Bayes factor threshold of 6, each of the frequentist CET conclusions match each of the Bayesian conclusions. Regardless of what approach is used, one can conclude that the data indicates evidence for the effect of population density $(p_{NHST}<0.001;\ BF_{01}=5.1\times10^{13}\ (\text{with }r=\sqrt{2}/4))$ and for the effect of global climate $(p_{NHST}<0.001;\ BF_{01}=8.9\times10^7\ (\text{with }r=\sqrt{2}/4))$, and evidence for a lack of effect of local climate $(p_{EQUIV}=0.014\ (\text{with }\Delta=0.10);\ BF_{01}=1/11\ (\text{with }r=\sqrt{2}/4))$. Results are inconclusive with regards to the effect (or lack thereof) of parasites $(p_{NHST}=0.144,\ p_{EQUIV}=0.276\ (\text{with }\Delta=0.10);\ BF_{01}=1/3.8\ (\text{with }r=\sqrt{2}/4))$.

³Note that the BF values published in Table 1 of Rouder & Morey (2012) are slightly different. The discrepancies for these four BF values are likely due to rounding errors; see Morey (2021).

\overline{k}	covariate	β_k	$SE(\hat{\beta}_k)$	\mathcal{B}_k	$SE(\widehat{\mathcal{B}_k})$
0	intercept	261.70	97.72	-	-
1	local climate	0.15	1.62	0.004	0.045
2	global climate	1871.75	271.82	0.360	0.052
3	parasites	-9.26	6.30	-0.073	0.049
4	pop. density	4.43	0.48	0.531	0.058
			$\hat{\sigma} = 168.7$		$R_{Y,X}^2 = 0.711$

Table 4. Parameter estimates obtained by standard least squares estimation for the hominid brain evolution multivariable linear regression model.

\overline{k}	\mathcal{B}_k	p_{NHST}	p_{EQUIV}	BF_{10}	CET conclusion	Bayesian conclusion
			$\Delta = 0.10$	$r = \sqrt{2}/4$	$\alpha = 0.05$	BF threshold $= 6$
1	0.004	0.927	0.014	1/11	Negative	Negative
2	0.360	< 0.001	1.000	8.9×10^{7}	Positive	Positive
3	-0.073	0.144	0.276	1/3.8	Inconclusive	Inconclusive
4	0.531	< 0.001	1.000	5.1×10^{13}	Positive	Positive

Table 5. Calculated values and conclusions for both frequentist and Bayesian testing for the hominid brain evolution multivariable linear regression model.

Conclusion

Researchers require statistical tools that allow them to reject the presence of meaning-ful effects. Indeed, such tools are essential to scientific progress; see Serlin et al. (1993), Altman & Bland (1995), and more recently Amrhein et al. (2019). In this paper we considered just such a tool: an equivalence test for standardized effect sizes in linear regression analyses.

Equivalence tests may improve current research practices by allowing researchers to falsify their predictions concerning the presence of an effect. In this sense, equivalence testing provides a more formal approach to the "good-enough principle" (Serlin et al., 1993). To be clear, effect sizes need not be dimensionless (or standardized) in order to be meaningful (Kelley & Preacher, 2012). However, expanding equivalence testing to standardized effect sizes can help researchers conduct equivalence tests by facilitating what is often a very challenging task: defining an appropriate equivalence margin. While the use of "default equivalence margins" based on standardized effect sizes cannot be whole-heartily recommended for all cases, their use is not unlike the use of "default priors" for Bayesian inference which have indeed proven useful to researchers in many scenarios.

As Rouder & Morey (2012) note when discussing default BFs: "Subjectivity

should not be reflexively feared. Many aspects of science are necessarily subjective. [...] Researchers justify their subjective choices as part of routine scientific discourse, and the wisdom of these choices are evaluated as part of routine review." The same sentiment applies to frequentist testing. Researchers using equivalence testing should be prepared to justify their choice for the equivalence margin based on what effect sizes are considered negligible. That being said, equivalence tests for standardized effects may help researchers in situations when what is "negligible" is particularly difficult to determine. They may also help establish generally acceptable standards for margins in the literature (Campbell & Gustafson, 2021b).

When it comes to comparing frequentist equivalence testing procedures to Bayesian procedures, we note that each approach has its merits and each is based on a fundamentally different underlying principle. If a researcher decides on which underlying principle they wish to subscribe to in tackling a given problem, then the method should follow naturally; see Campbell & Gustafson (2021a). We observed in two of the practical examples and in simulations, that both approaches may often provide one with similar results for testing standardized regression coefficients.

Note that our non-inferiority test for the increase in the squared multiple correlation coefficient ($\operatorname{diff} P_k^2$) in a standard multivariable linear regression is limited to comparing two models for which the difference in degrees of freedom is 1. In other words, the test is not suitable for comparing two nested models where the difference is more than a single variable. For example, with the salaries data we considered, we cannot use the proposed test to compare a "smaller model" with only "sex" as a covariate, with a "larger model" that includes "sex," "discipline" and "rank," as covariates. A more general equivalence test for comparing two nested models will be considered in future work; Tan Jr (2012) is an excellent resource for this undertaking.

We wish to emphasize that the use of equivalence/non-inferiority tests should not rule out the complementary use of confidence intervals. Indeed, confidence intervals can be extremely useful for highlighting the stability (or lack of stability) of a given estimator, whether that be the β_k , \mathcal{B}_k , $diff P_k^2$ or any other statistic (Fidler et al., 2004). One major strength of confidence intervals is that, not only can they indicate if the effect of interest is trivial, but they can also indicate how small the effect may be. Perhaps one advantage of equivalence/non-inferiority testing over confidence intervals may be that testing can improve the interpretation of null results (Parkhurst, 2001; Hauck & Anderson, 1986). By clearly distinguishing between what is a "negative" versus an "inconclusive" result, equivalence testing serves to simplify the long "series of searching questions" necessary to evaluate a "failed outcome" (Pocock & Stone, 2016). The best interpretation of data might be obtained when using both tools together. Those wishing to calculate a confidence interval for a standardized regression coefficient to present alongside an equivalence test should see Kelley (2007) for formulas and code.

Finally, note that we only considered equivalence tests using TOST derived from inverting the noncentrality parameter confidence intervals. It would certainly be worthwhile in future research to consider equivalence tests based on alternative approximations for the sampling variability of standardized regression coefficients (Jones & Waller, 2013; Yuan & Chan, 2011). We also note that the TOST approach is not necessarily optimal in the sense that other procedures may have slightly higher power (Möllenhoff et al., 2022). Finally, there is certainly potential to expand equivalence testing for standardized regression coefficients in logistic regression models and time-to-event models. Such work will help to further "extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" (Wellek, 2017).

References

Acock, A. C. (2008). A gentle introduction to Stata, 4th ed. College Station, TX: Stata Press.

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

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 7748 (567), 305–307. doi: https://doi.org/10.1038/d41586-019-00857-9.

- Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12, http://dx.doi.org/10.1037/met0000051.
- Azen, R., & Budescu, D. (2009). Applications of multiple regression in psychological research. In R.E. Millsap, A. Maydeu-Olivares (Eds.), The SAGE handbook of quantitative methods in psychology (pp. 285–310). Sage Thousand Oaks, CA. https://doi.org/10.4135/9780857020994.
- Baguley, T. (2009). Standardized or simple effect size: What should be reported? British Journal of Psychology, 100(3), 603–617. doi: 10.1348/000712608X377117.
- Bailey, D. H., & Geary, D. C. (2009). Hominid brain evolution. *Human Nature*, 20(1), 67–79. https://doi.org/10.1007/s12110-008-9054-0.
- Ball, L. C., Cribbie, R. A., & Steele, J. R. (2013). Beyond gender differences: Using tests of equivalence to evaluate gender similarities. *Psychology of Women Quarterly*, 37(2), 147–154. https://doi.org/10.1177/0361684313480483.
- Barten, A. (1962). Note on unbiased estimation of the squared multiple correlation coefficient. Statistica Neerlandica, 16(2), 151–164. doi:10.1111/J.1467-9574.1962.TB01062.X.
- Berger, J. O. (1985). Statistical decision theory and Bayesian analysis. New York, NY: Springer Science & Business Media. doi:10.1007/978-1-4757-4286-2.
- Briggs, W. M., Nguyen, H. T., & Trafimow, D. (2019). The replacement for hypothesis testing. In *International Conference of the Thailand Econometrics Society* (pp. 3–17).
- Bring, J. (1994). How to standardize regression coefficients. *The American Statistician*, 48(3), 209–213. doi: 10.1080/00031305.1994.10476059.
- Burgoyne, A. P., Hambrick, D. Z., & Macnamara, B. N. (2020). How firm are the foundations of mind-set theory? The claims appear stronger than the evidence. *Psychological Science*, 31(3), 258–267. doi: 10.1177/0956797619897588.

- Campbell, H., & Gustafson, P. (2018). Conditional equivalence testing:

 An alternative remedy for publication bias. *PLoS ONE*, 13(4), e0195145.

 https://doi.org/10.1371/journal.pone.0195145.
- Campbell, H., & Gustafson, P. (2021a). re: Linde et al.(2021)—the Bayes factor, HDI-ROPE and frequentist equivalence testing are actually all equivalent. arXiv preprint arXiv:2104.07834.
- Campbell, H., & Gustafson, P. (2021b). What to make of equivalence testing with a post-specified margin? *Meta-Psychology*, 5. doi: https://doi.org/10.15626/MP.2020.2506.
- Campbell, H., & Lakens, D. (2021). Can we disregard the whole model? Omnibus non-inferiority testing for R2 in multi-variable linear regression and in ANOVA. British Journal of Mathematical and Statistical Psychology, 74(1), 64–89. https://doi.org/10.1111/bmsp.12201.
- Cohen, J. (1988). Statistical power analysis for the behavioral sciences. New York, NY: Routledge Academic.
- Consonni, G., & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23(3), 332–353. doi: 10.1214/08-STS258.
- Cramer, J. S. (1987). Mean and variance of R2 in small and moderate samples. *Journal of Econometrics*, 35(2-3), 253–266. https://doi.org/10.1016/0304-4076(87)90027-3.
- Dixon, P. M., Saint-Maurice, P. F., Kim, Y., Hibbing, P., Bai, Y., & Welk, G. J. (2018). A primer on the use of equivalence testing for evaluating measurement agreement. *Medicine and Science in Sports and Exercise*, 50(4), 837. doi: 10.1249/MSS.0000000000001481.
- Dudgeon, P. (2017). Some improvements in confidence intervals for standardized regression coefficients. *Psychometrika*, 82(4), 928–951. doi: 10.1007/s11336-017-9563-z.
- Eskine, K. J. (2013). Wholesome foods and wholesome morals? Organic foods reduce prosocial behavior and harshen moral judgments. *Social Psychological and*

- Personality Science, 4(2), 251-254. https://doi.org/10.1177/1948550612447114.
- Etz, A. (2015). Using Bayes factors to get the most out of linear regression: A practical guide using R. *The Winnower*. Retrieved from https://tinyurl.com/2p9a5ymr
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15(2), 119–126. doi: 10.1111/j.0963-7214.2004.01502008.x.
- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., . . . Graves, S. (2012). Package "car". Vienna: R Foundation for Statistical Computing.
- Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, 9(10), e109019. https://doi.org/10.1371/journal.pone.0109019.
- Fruehauf, L. M., Fair, J. E., Liebel, S. W., Bjornn, D., & Larson, M. J. (2021). Cognitive control in obsessive-compulsive disorder (ocd): Proactive control adjustments or consistent performance? *Psychiatry Research*, 298, 113809; https://doi.org/10.1016/j.psychres.2021.113809.
- Ghashim, E., & Boily, P. (2018). A ggplot2 primer. Data Action Lab Data Science Report Series. Retrieved from https://tinyurl.com/2p9brz9a
- 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.
- Goodman, E., Huang, B., Wade, T. J., & Kahn, R. S. (2003). A multilevel analysis of the relation of socioeconomic status to adolescent depressive symptoms: Does school context matter? *The Journal of Pediatrics*, 143(4), 451–456. https://doi.org/10.1067/S0022-3476(03)00456-6.
- Hartung, J., Cottrell, J. E., & Giffin, J. P. (1983). Absence of evidence is not evidence of absence. Anesthesiology: The Journal of the American Society of Anesthesiologists, 58(3), 298–299. https://doi.org/10.1097/00000542-198303000-00033.

- Hauck, W. W., & Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, 5(3), 203–209. doi: 10.1002/sim.4780050302.
- Heck, D. W. (2019). A caveat on the Savage–Dickey density ratio: the case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72(2), 316–333. https://doi.org/10.1111/bmsp.12150.
- Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *Journal of Counseling Psychology*, 29, 58–65. https://doi.org/10.1037/0003-066X.58.1.78.
- Hung, H., Wang, S.-J., & O'Neill, R. (2005). A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal*, 47(1), 28–36. https://doi.org/10.1002/bimj.200410084.
- Jeffreys, H. (1961). The Theory of Probability. Oxford, UK: Oxford University Press.
- Jones, J. A., & Waller, N. G. (2013). Computing confidence intervals for standardized regression coefficients. *Psychological Methods*, 18(4), 435. doi: 10.1037/a0033269.
- Keefe, R. S., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., ... Leon, A. C. (2013). Defining a clinically meaningful effect for the design and interpretation of randomized controlled trials. *Innovations in Clinical Neuroscience*, 10(5-6 Suppl A), 4S–19S. PMCID: PMC3719483.
- Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20(8), 1–24. doi: 10.18637/jss.v020.i08.
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17(2), 137–152. https://doi.org/10.1037/a0028086.
- Koh, A., & Cribbie, R. (2013). Robust tests of equivalence for k independent groups. British Journal of Mathematical and Statistical Psychology, 66(3), 426–434. doi: 10.1111/j.2044-8317.2012.02056.x.
- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, 9(9), e105825. https://doi.org/10.1371/journal.pone.0105825.

- 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. (2022). Introduction to equivalence testing with TOSTER. https://cran.rstudio.com/web/packages/TOSTER/vignettes/IntroductionToTOSTER.html.
- 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. https://doi.org/10.1177/2515245918770963.
- Leonidaki, V., & Constantinou, M. P. (2021). A comparison of completion and recovery rates between first-line protocol-based cognitive behavioural therapy and non-manualized relational therapies within a UK psychological service. *Clinical Psychology & Psychotherapy*, 1—13. doi: 10.1002/cpp.2669.
- 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 Associ*ation, 103(481), 410–423. https://doi.org/10.1198/016214507000001337.
- Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, 112(2), 331–348. doi: 10.2466/03.11.PMS.112.2.331-348.
- Moery, E., & Calin-Jageman, R. J. (2016). Direct and conceptual replications of eskine (2013) organic food exposure has little to no effect on moral judgments and prosocial behavior. *Social Psychological and Personality Science*, 7(4), 312–319. https://doi.org/10.1177/1948550616639649.
- Möllenhoff, K., Loingeville, F., Bertrand, J., Nguyen, T. T., Sharan, S., Zhao, L., ... Mentré, F. (2022). Efficient model-based bioequivalence testing. *Biostatistics*, 23(1), 314–327. doi: 10.1093/biostatistics/kxaa026.
- Moon, K.-W. (2017). Learn "ggplot2" Using Shiny App. Cham, Switzerland: Springer International Publishing. doi: 10.1007/978-3-319-53019-2.

- Morey, R. D. (2021). Replicate Table 1 of Rouder and Morey (2012). RPubs by R Studio. Retrieved from https://rpubs.com/richarddmorey/RouderMorey2012
- 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.
- R. J. Ν., T., D. Morey, D., Rouder, Jamil, & Morey, Μ. R. Package 'BayesFactor'. CRAN repository. (2015).Retrieved from https://cran.r-project.org/web/packages/BayesFactor/index.html
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. https://doi.org/10.1002/sim.8086.
- Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691–692. https://doi.org/10.1093/biomet/78.3.691.
- Parkhurst, D. F. (2001). Statistical significance tests: Equivalence and reverse tests should reduce misinterpretation. *Bioscience*, 51(12), 1051–1057. https://doi.org/10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2.
- Pocock, S. J., & Stone, G. W. (2016). The primary outcome fails -what next? New England Journal of Medicine, 375(9), 861–870. doi: 10.1016/j.jacc.2021.06.024.
- R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/
- Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47(6), 877–903. https://doi.org/10.1080/00273171.2012.734737.
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. doi: 10.3758/s13423-017-1230-y.

- Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, 3(4), 403–411. https://doi.org/10.1037/1082-989X.3.4.403.
- Serlin, R. C., Lapsley, D. K., Keren, G., & Lewis, C. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 199–228). Hillsdale, NJ: Psychology Press.
- Tan Jr, L. (2012). Confidence intervals for comparison of the squared multiple correlation coefficients of non-nested models. Thesis submitted in partial fulfillment of the requirements for the degree in Master of Science; The University of Western Ontario.
- Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, 24(6), 774–795. doi: 10.1037/met0000221.
- Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, 6(3), 190–213. https://doi.org/10.1080/19312458.2012.703834.
- Wellek, S. (2010). Testing statistical hypotheses of equivalence and noninferiority.

 Boca Raton, FL: Chapman and Hall/CRC.
- Wellek, S. (2017). A critical evaluation of the current "p-value controversy". *Biometrical Journal*, 59(5), 854–872. doi: 10.1002/bimj.201700001.
- West, S. G., Aiken, L. S., Wu, W., & Taylor, A. B. (2007). Multiple regression: Applications of the basics and beyond in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality* psychology (pp. 573–601). The Guilford Press.
- Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. Controlled Clinical Trials, 23(1), 2–14. doi: 10.1016/s0197-2456(01)00196-9.

Yuan, K.-H., & Chan, W. (2011). Biases and standard errors of standardized regression coefficients. *Psychometrika*, 76(4), 670–690. https://doi.org/10.1007/s11336-011-9224-6.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, 227(3), 617–628. https://doi.org/10.1148/radiol.2273011499.

Appendix

Additional formulas and notation

Least squares estimation

For completeness, we provide details and notation for least squares estimation in a standard linear regression model. We define:

$$\hat{\beta}_k = ((X^T X)^{-1} X^T y)_k$$
, for k in 1,..., K ; and (21)

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (\hat{\epsilon}_i^2) / (N - K - 1)},$$
(22)

where $\hat{\epsilon}_i = \hat{y}_i - y_i$, and $\hat{y}_i = X_{i \times}^T \hat{\beta}$, for i in 1,..., N. A standard NHST for the k-th covariate, X_k , is stated as:

 $H_0: \beta_k = 0$, vs.

 $H_1: \beta_k \neq 0.$

Typically one conducts one of two different (yet mathematically identical) tests. Most commonly a t-test is done to calculate a p-value as follows:

$$p\text{-value}_k = 2 \times F_t \left(\frac{|\widehat{\beta_k}|}{SE(\widehat{\beta_k})}, N - K - 1 \right), \text{ for } k \text{ in } 0, ..., K,$$
 (23)

where we use $F_t(\ \cdot\ ;df)$ to denote the cdf of the t-distribution with df degrees of freedom, and where: $\widehat{SE(\beta_k)} = \hat{\sigma}\sqrt{[(X^TX)^{-1}]_{kk}}$. Alternatively, we can conduct an

F-test and, for k in 1,...,K, we will obtain the very same p-value with:

$$p\text{-value}_k = p_F \left((N - K - 1) \frac{\text{diff} R_k^2}{1 - R_{YX}^2}, 1, N - K - 1 \right),$$
 (24)

where $p_f(\cdot ; df_1, df_2)$ is the cdf of the F-distribution with df_1 and df_2 degrees of freedom, and where: $\text{diff}R_k^2 = R_{YX}^2 - R_{YX_{-k}}^2$. Regardless of whether the t-test or the F-test is employed, if p-value $_k < \alpha$, we reject the null hypothesis of $H_0: \beta_k = 0$ against the alternative $H_0: \beta_k \neq 0$.

A valid equivalence test for the standardized difference between two independent means

A valid equivalence test for the standardized difference between two independent means, θ , can be defined by the following null and alternative hypotheses (see Weber & Popova (2012)):

$$H_0: \theta \leq \Delta_{lower}$$
 or: $\theta \geq \Delta_{upper}$, vs.

$$H_1: \theta > \Delta_{lower}$$
 and: $\theta < \Delta_{upper}$,

where $\theta = \mu_d/\sigma$ and the equivalence margin is $(\Delta_{lower}, \Delta_{upper})$. A *p*-value for this test can then be calculated as *p*-value=max $(p_d^{lower}, p_d^{upper})$, where:

$$p_d^{lower} = 1 - F_t \left(\frac{\hat{\mu}_d}{\hat{\sigma}_p} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}, N_1 + N_2 - 2, \Delta_{lower} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \right), \text{ and } (25)$$

$$p_d^{upper} = 1 - F_t \left(-\frac{\hat{\mu}_d}{\hat{\sigma}_p} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}, N_1 + N_2 - 2, -\Delta_{upper} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \right),$$

where N_1 is the number of observations in the first sample, N_2 is the number of observations in the second sample, where $\hat{\mu}_d$ is the difference between the two sample means, and $\hat{\sigma}_p$, the pooled standard deviation estimate, is calculated from the two

samples as:

$$\hat{\sigma}_p = \sqrt{\frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{N_1 + N_2 - 2}},\tag{26}$$

where $\hat{\sigma}_1$ is the estimated standard deviation of the first sample, and $\hat{\sigma}_2$ is the estimated standard deviation of the second sample.

An equivalence test for correlations based on Fisher's Z transformation

A p-value from the equivalence test for correlations based on Fisher's Z transformation is calculated as $p_Z = max(\mathbf{p}_Z^{lower}, \mathbf{p}_Z^{upper})$, where:

$$p_Z^{lower} = 1 - F_Z \left(\frac{\sqrt{N-3}}{2} \ln \left(\left(\frac{1+\widehat{\mathcal{B}}_1}{1-\widehat{\mathcal{B}}_1} \right) - \left(\frac{1+\Delta_{lower}}{1-\Delta_{lower}} \right) \right) \right), \tag{27}$$

and:

$$\mathbf{p}_{Z}^{upper} = 1 - F_{Z} \left(\frac{\sqrt{N-3}}{2} \ln \left(\left(\frac{1+\widehat{\mathcal{B}}_{1}}{1-\widehat{\mathcal{B}}_{1}} \right) + \left(\frac{1+\Delta_{upper}}{1-\Delta_{upper}} \right) \right) \right),$$

where $F_Z()$ denotes the cdf of the standard normal distribution; see Goertzen & Cribbie (2010) for details.

Simulation Study 1 - alternative settings

We conducted an alternate version of Simulation Study 1 with correlated binary covariates.

Simple linear regression

For K = 1, we sampled X_1 as a binary covariate in such a way so that half of the X_1 values are equal to 1. We set $\beta = (-0.20, 0.10)^T$ to correspond to $\mathcal{B}_1 = 0.070$, 0.124, and 0.200, with $\sigma^2 = 0.50, 0.15$ and 0.05, respectively. These non-zero values correspond to relatively small effect sizes (Hemphill, 2003). We set $\beta = (-0.20, 0.00)^T$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

Figures 6 and 7 plot the results for the simple linear regression simulation with alternative settings. Results are similar to those obtained with the original settings.

Multivariable linear regression

For K=2, we sampled correlated binary variables in such a way so that $cor(X_1, X_2) = 0.40$, and so that half of the X_1 values are equal to 1 and only a quarter of the X_2 values are equal to 1. We set $\beta = (-0.20, 0.10, 0.19)^T$ to correspond to $\mathcal{B}_1 = 0.070, 0.124$, and 0.200, with $\sigma^2 = 0.50, 0.15$ and 0.05, respectively. These non-zero values correspond to relatively small effect sizes (Hemphill, 2003). We set $\beta = (-0.20, 0.00, 0.19)^T$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

With K=4, we sampled correlated binary variables in such a way that the correlation between the four variables was:

$$cor(X_1, X_2, X_3, X_4) = \begin{pmatrix} 1 & 0.4 & 0.3 & 0 \\ 0.4 & 1 & 0.4 & 0.3 \\ 0.3 & 0.4 & 1 & 0.4 \\ 0 & 0.3 & 0.4 & 1 \end{pmatrix},$$
(28)

and so that half of the X_1 and the X_4 values are equal to 1, and only a quarter of the X_2 and the X_3 values are equal to 1. We set $\beta = (0.20, 0.10, 0.14, -0.12, -0.14)^T$ to correspond to $\mathcal{B}_1 = 0.070, 0.124$, and 0.200, with $\sigma^2 = 0.50, 0.15$ and 0.05, respectively. We set $\beta = (0.20, 0.00, 0.14, -0.12, -0.14)^T$ to correspond to $\mathcal{B}_1 = 0.000$, with $\sigma^2 = 0.50$.

Figures 8 and 9 plot the results for the multivariable linear regression simulation with alternative settings. Results are similar to those obtained with the original settings. Note however, that the approximate power numbers obtained from equation (8) overestimate the true power for N = 180, for all values of Δ_1 .

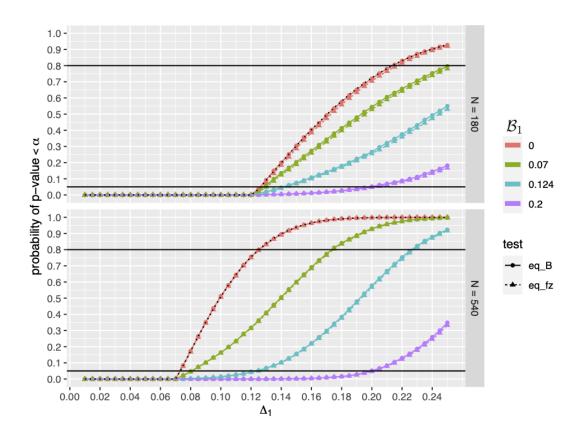


Figure 6. Simulation Study 1 (alternative settings) - Results for simple linear regression with K=1. Upper panel shows results for N=540. Two different equivalence tests were applied to each simulated dataset: (1) "eq_B" the proposed equivalence test for standardized regression coefficients (i.e., equation (6)); and (2) "eq_fz" the equivalence test for correlations based on Fisher's Z transformation proposed by Goertzen & Cribbie (2010) (see equation (27) and details in the Appendix). The lower solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$ and the upper solid horizontal black line indicates the reference power of 0.80. The curved dotted black lines correspond to numbers obtained using the proposed formula for approximate power calculation (equation (8)). Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

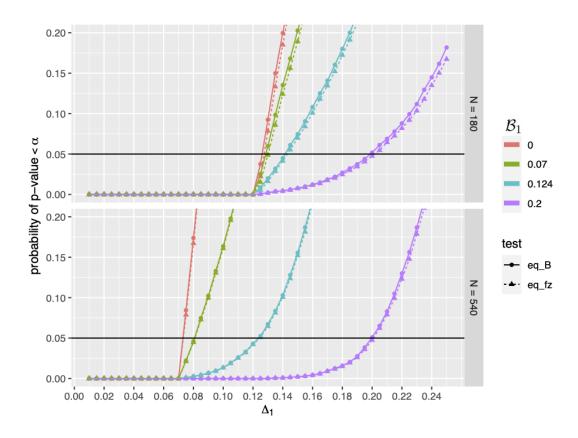


Figure 7. Simulation Study 1 (alternative settings) - Note that this Figure is an "inset" (i.e., a "magnified portion") of Figure 6. Results for simple linear regression with K=1. Upper panel shows results for N=180; Lower panel shows results for N=540. Two different equivalence tests were applied to each simulated dataset: (1) "eq_B" the proposed equivalence test for standardized regression coefficients (i.e., equation (6)); and (2) "eq_fz" the equivalence test for correlations based on Fisher's Z transformation proposed by Goertzen & Cribbie (2010) (see equation (27) and details in the Appendix). Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$. Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

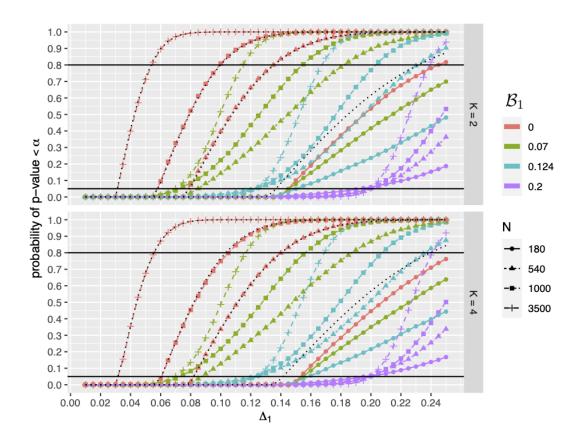


Figure 8. Simulation Study 1 (alternative settings) - Upper panel shows results for K=2; Lower panel shows results for K=4. The proposed equivalence test for standardized regression coefficients (i.e., "eq.B", see equation (6)) was applied to each simulated dataset. The solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$ and the upper solid horizontal black line indicates the reference power of 0.80. The curved dotted black lines correspond to numbers obtained using the proposed formula for approximate power calculation (equation (8)). Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

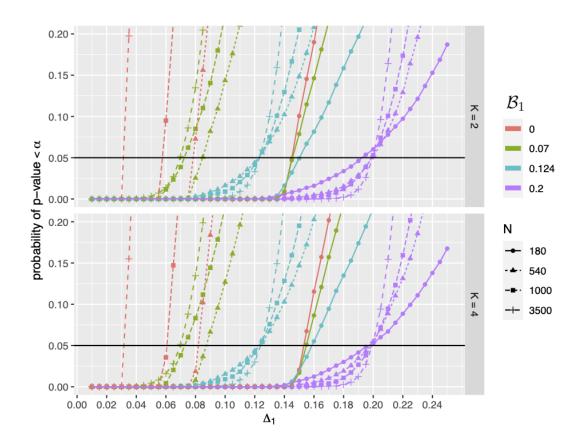


Figure 9. Simulation Study 1 (alternative settings) - Note that this Figure is an "inset" (i.e., a "magnified portion") of Figure 8. Upper panel shows results for K=2; lower panel shows results for K=4. The proposed equivalence test for standardized regression coefficients (i.e., "eq_B", see equation (6)) was applied to each simulated dataset. Both plots are presented with a restricted vertical-axis to better show the type 1 error rates. The solid horizontal black line indicates the desired type 1 error of $\alpha=0.05$. Note that the maximum type 1 error rate should not exceed $\alpha=0.05$. As such, when $\Delta_1 \leq \mathcal{B}_1$, the probability of a p-value less than 0.05 should not exceed 0.05. When $\Delta_1 > \mathcal{B}_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

Simulation Study 2

We simulated data in order to compare the operating characteristics of two equivalence tests for the difference between two independent means:

- (1) the invalid test (i.e., the test proposed by Lakens (2017)), with null hypothesis $H_0: |\mu_d| \ge \Delta \times \hat{\sigma}$; and
- (2) the valid test (see equation (25)), with null hypothesis $H_0: |\theta| \geq \Delta$, where $\theta = \mu_d/\sigma$.

We considered 6 different values for the total sample size, N, ranging from 54 to 3500, and 4 different values for the upper bound of a symmetric equivalence margin. Δ , ranging from 0.2 to 1.0. We simulated data from a Normal distribution such that the true Cohen's d was either equal to 0 or equal to Δ .

For each of the different configurations within the simulation study, we simulated 2,000,000 unique datasets and calculated a p-value with each of the two equivalence tests. We then calculated the proportion of these p-values less than $\alpha=0.05$. We specifically chose to conduct 2,000,000 simulation runs so as to keep computing time within a reasonable limit while also reducing the amount of Monte Carlo standard error to a very negligible amount (for looking at type 1 error with $\alpha=0.05$, Monte Carlo SE will be approximately $0.00015 \approx \sqrt{0.05(1-0.05)/2,000,000}$; see Morris et al. (2019)). Results are displayed in Table 1 and suggest that, in practice, using the invalid test can lead to a higher than advertised type 1 error when sample sizes are large and a minor loss of efficiency when sample sizes are small.

N	Δ	$\Pr(p\text{-value} < 0.05 d = \Delta)$		Pr(p-value < 0.05 d = 0)	
		invalid test	valid test	invalid test	valid test
54	0.20	0.000	0.000	0.000	0.000
54	0.50	0.024	0.029	0.129	0.152
54	0.75	0.047	0.050	0.715	0.726
54	1.00	0.049	0.050	0.949	0.950
80	0.20	0.000	0.000	0.000	0.000
80	0.50	0.045	0.047	0.431	0.444
80	0.75	0.049	0.050	0.905	0.907
80	1.00	0.051	0.050	0.994	0.993
180	0.20	0.000	0.000	0.000	0.000
180	0.50	0.049	0.050	0.909	0.910
180	0.75	0.051	0.050	0.999	0.999
180	1.00	0.055	0.050	1.000	1.000
540	0.20	0.048	0.049	0.501	0.503
540	0.50	0.051	0.050	1.000	1.000
540	0.75	0.053	0.050	1.000	1.000
540	1.00	0.057	0.050	1.000	1.000
1000	0.20	0.050	0.050	0.870	0.870
1000	0.50	0.051	0.050	1.000	1.000
1000	0.75	0.054	0.050	1.000	1.000
1000	1.00	0.058	0.050	1.000	1.000
3500	0.20	0.050	0.050	1.000	1.000
3500	0.50	0.052	0.050	1.000	1.000
3500	0.75	0.055	0.050	1.000	1.000
3500	1.00	0.059	0.050	1.000	1.000

Table 6. Results from Simulation study 2. Note that the maximum type 1 error rate should not exceed $\alpha = 0.05$. As such, when $\Delta = d$, the probability of a *p*-value less than 0.05 should not exceed 0.05. When $\Delta > d$, the probability of a *p*-value less than 0.05 corresponds to the test's statistical power.

Simulation Study 3

By means of a simple simulation study, we compared a CET frequentist testing scheme (based on NHST and equivalence testing of standardized regression coefficients) to the Bayesian approach based on Rouder & Morey (2012)'s default Bayes factors (BF). The hypothesis in question is:

 $H_0: |\mathcal{B}_1| \geq \Delta_1$, vs.

 $H_1: |\mathcal{B}_1| < \Delta_1.$

In the simulation study, frequentist conclusions are based on the CET procedure by setting Δ_1 equal to either 0.05, or 0.10, or 0.25; and with α =0.05. Bayesian conclusions are based on an evidence threshold of 6. A threshold of 6 can be considered "strong evidence" (Jeffreys, 1961). All priors required for calculating the BF are set by simply selecting the default settings of the regressionBF() function (with $r = \sqrt{2}/4$); see Morey et al. (2015).

We simulated datasets for 36 unique scenarios. We varied over the following:

- one of twelve sample sizes: N = 20, N = 33, N = 55, N = 90, N = 149, N = 246, N = 406, N = 671, N = 1,109, N = 1,832, N = 3,027, or N = 5,000;
- one of two designs with K=4 continuous covariates (normally distributed with mean=0 and sd=0.25), with either $\beta = (0.20, 0.10, 0.14, -0.10, -0.10)^T$ or $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)^T$;
- one of two variances: $\sigma^2 = 0.50$, or $\sigma^2 = 1.00$.

Note that for the $\beta = (0.20, 0.00, 0.14, -0.10, -0.10)^T$ design, we only consider one value for $\sigma^2 = 1.00$. The outcome variable was simulated from a Normal distribution such that $Y_i \sim \text{Normal}(X_{i\times}^T\beta, \sigma^2), \forall i = 1, ..., N$. Depending on the particular design and σ^2 , the true standardized regression coefficient, \mathcal{B}_1 , for these data is either $\mathcal{B}_1 = 0.00$, $\mathcal{B}_1 = 0.05$, or $\mathcal{B}_1 = 0.07$. These values for \mathcal{B}_1 were chosen so as to show a range of potential conclusions from a majority of positive conclusions to a majority of negative conclusions.

The simulation study was done in R statistical software using the default simulation routines (R Core Team, 2020). For each simulated dataset, we obtained frequentist p-values, BFs, and declared the result to be positive, negative or inconclusive, accordingly. Results are presented in Figure 10 and are based on 500 distinct simulated datasets per scenario.

Based on our comparison of BFs and frequentist tests, we can confirm that, given the same data, both approaches will often provide one with the same overall conclusion. In certain scenarios however, we do see notable differences. For example, the probability that the CET result is negative (the correct conclusion) far exceeds that of a negative Bayesian conclusion when $\Delta_1 = 0.25$ and $\mathcal{B}_1 = 0$ (see Fig 10, panel 3). However, with small non-zero effect sizes and large sample sizes (see Figure 10, panels 5,6, 8, and 9), the CET procedure is more likely to incorrectly obtain a positive conclusion because the first test (the NHST) has sufficient power to declare a positive

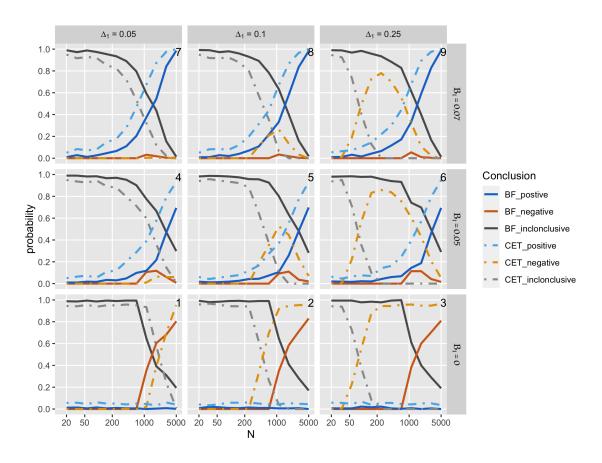


Figure 10. Simulation Study 3, complete results for BF threshold of 6. The probability of obtaining each conclusion by Bayesian testing scheme (JZS-BF with fixed sample size design, BF threshold of 6:1) and CET ($\alpha = 0.05$). Each panel displays the results of simulations with for different values of Δ_1 and \mathcal{B}_1 . Note that all solid lines and the dashed blue line do not change for different values of Δ_1 .

effect.

The level of agreement between the frequentist and Bayesian approaches is highly sensitive to the choice of Δ . Averaging over all 36 scenarios and over the 500 Monte Carlo simulations per scenario, the Bayesian and frequentist methods reached the very same conclusion 86% of the time (with $\Delta=0.05$), 77% of the time (with $\Delta=0.10$), and 48% of the time (with $\Delta=0.25$) (numbers based on average over $36\times500=18,000$ unique datasets). While we did not consider the impact of selecting different priors or different Bayes factor evidence thresholds, it is reasonable to assume that the level of agreement between the Bayesian and frequentist tests will also be rather sensitive to the chosen priors and the particular evidence threshold; see Berger (1985) and more recently Campbell & Gustafson (2021a).

Five useful R-functions for calculating p-values and Bayes factors

The "equivBeta" function can be used to calculate the p-values from equation (3).

```
##### equivBeta #####
equivBeta <- function(Y = rnorm(100),</pre>
                       Xmatrix = cbind(rnorm(100), rnorm(100)),
                       DELTA_upper = 0.1,
                       DELTA_lower = -0.1){
  if(is.na(DELTA_lower)[1]){DELTA_lower <-(-DELTA_upper)}</pre>
  Xmatrix <- cbind(Xmatrix)</pre>
  X <- cbind(1, Xmatrix)</pre>
  N <- dim(cbind(X[,-1]))[1]</pre>
  K <- dim(cbind(X[,-1]))[2]</pre>
  if(length(DELTA_lower)==1){DELTA_lower <- rep(DELTA_lower, K+1)}</pre>
  if(length(DELTA_upper)==1){DELTA_upper <- rep(DELTA_upper, K+1)}</pre>
  lmmod <- summary(lm(Y~X[,-1]))</pre>
  beta_hat <- lmmod$coef[,1]</pre>
  SE_beta_hat <- lmmod$coef[,2]</pre>
pval <- p_lower <- p_upper <- rep(0,K)</pre>
  for(k in 1:(K+1)){
    p_lower[k] <- pt((beta_hat[k] - DELTA_lower[k])/SE_beta_hat[k], N-K-1, 0, lower.tail=FALSE)</pre>
    p_upper[k] <- pt((-beta_hat[k] + DELTA_upper[k])/SE_beta_hat[k], N-K-1, 0, lower.tail=FALSE)</pre>
    pval[k] <- max(c(p_lower[k],p_upper[k]))</pre>
  }
  names(beta_hat) <- paste("beta", c(1:dim(X)[2])-1, sep="_")
  return(list(beta = beta_hat, pval = pval, DELTA = cbind(DELTA_lower, DELTA_upper)))
}
The "equivstandardBeta" function can be used to calculate the p-values from equation
(6).
##### equivstandardBeta #####
equivstandardBeta <- function(Y = rnorm(100),
                                Xmatrix = cbind(rnorm(100), rnorm(100)),
                                DELTA_upper = 0.1,
                                DELTA_lower = -0.1){
```

```
if(is.na(DELTA_lower)[1]){DELTA_lower <-(-DELTA_upper)}</pre>
Xmatrix <-cbind(Xmatrix)</pre>
X <- cbind(1, Xmatrix)</pre>
N <- dim(cbind(X[,-1]))[1]</pre>
K <- dim(cbind(X[,-1]))[2]</pre>
if(length(DELTA_lower)==1){DELTA_lower <- rep(DELTA_lower, K+1)}</pre>
if(length(DELTA_upper)==1){DELTA_upper <- rep(DELTA_upper, K+1)}</pre>
unstandard_beta <- lm(Y ~ cbind(X[,-1]))$coefficients[-1]</pre>
sigma2_Y <- var(Y)
sigma2_X <- apply(cbind(X[,-1]),2, var)</pre>
standard_beta<-unstandard_beta*(c(sqrt(sigma2_X))/c(sqrt(sigma2_Y)))</pre>
p_lower <- p_upper <- pval <- rep(0,K)</pre>
R2YdotXmink <- R2YdotX <- R2XkdotXminK <- lambda_U2 <- lambda_L2 <- rep(0,K)
upperCI2 <- lowerCI2 <-upperCI <- lowerCI <- SE_beta_FIX <- rep(0,K)
lambda_U <- lambda_L <- rep(0,K)</pre>
b_vec <- standard_beta
for(k in 1:K){
  if(K>1){ Xmink <- cbind(cbind(X[,-1])[,-k])}</pre>
  if(K==1){ Xmink <- rep(1,N)}</pre>
  R2YdotXmink[k] <- summary(lm(Y ~ Xmink))$r.squared</pre>
  R2XkdotXminK[k] \leftarrow (summary(lm(cbind(X[,-1])[,k]^ Xmink)))r.squared
  R2YdotX[k] <- (summary(lm(Y~ cbind(X[,-1]))))$r.squared</pre>
  SE\_beta\_FIX[k] \leftarrow sqrt( (1-R2YdotX[k])/( (1-R2XkdotXminK[k])*(N-K-1) )) \# see Kelley2007 eq80.
}
for(k in 1:K){
  ncp_d_lower = 1 - ((1-R2XkdotXminK[k])*DELTA_lower[k]^2 + R2YdotXmink[k])
  ncp_lower = DELTA_lower[k]*sqrt(N*(1-R2XkdotXminK[k]))/sqrt(ncp_d_lower)
  ncp_d_upper = 1 - ((1-R2XkdotXminK[k])*DELTA_upper[k]^2 + R2YdotXmink[k])
  ncp_upper = -DELTA_upper[k]*sqrt(N*(1-R2XkdotXminK[k]))/sqrt(ncp_d_upper)
  p_lower[k] <- pt(b_vec[k]/SE_beta_FIX[k], N-K-1, ncp=ncp_lower, lower.tail=FALSE)</pre>
  p_upper[k] <- pt(-b_vec[k]/SE_beta_FIX[k], N-K-1, ncp=ncp_upper, lower.tail=FALSE)</pre>
  pval[k] <- max(c(p_lower[k], p_upper[k]), na.rm=TRUE)</pre>
}
return(list(standard_beta = standard_beta, pval = pval, DELTA = cbind(DELTA_lower, DELTA_upper)))
```

}

The "BFstandardBeta" function can be used to calculate the "default" Bayes factors.

```
##### BFstandardBeta #####
BFstandardBeta <- function(Y = rnorm(100), Xmatrix = cbind(rnorm(100), rnorm(100)),</pre>
                              BFthres = 3, kvec=c(1:K)){
  require("BayesFactor")
  Xmatrix<-cbind(Xmatrix)</pre>
  K<-dim(Xmatrix)[2]; print(K)</pre>
  mydata<-data.frame(Y, Xmatrix)</pre>
  colnames(mydata)<- c(c("yvector"), paste("X",1:K,sep=""))</pre>
  head(mydata)
  BFmod <- regressionBF(yvector~. , data = mydata)</pre>
  BF<-result <- 0*kvec
  if(K>1){
    for(k in kvec){
      whichk<-paste("X",k,sep="")
      BF_without_k <-BFmod[!grepl(whichk,names(BFmod)$numerator)][</pre>
        which.max(nchar(names(BFmod) $numerator[!grepl(whichk,names(BFmod) $numerator)]))]
      BF_full <- BFmod[which.max(nchar(names(BFmod)$numerator))]</pre>
      BF[k] <- exp(as.numeric(</pre>
        slot(BF_without_k, "bayesFactor")[1]))/exp(as.numeric(slot(BF_full, "bayesFactor")[1]))
    }
  if(K==1){ BF<-exp(slot(regressionBF(yvector~. , data = mydata), "bayesFactor")[1]) }</pre>
  for(k in kvec){
    if(BF[k] <= 1/BFthres){result[k] <- "positive" }</pre>
    if(BF[k]>BFthres){result[k]<-"negative"}</pre>
    if(BF[k]> 1/BFthres & BF[k] < BFthres) {result[k] <- "inconclusive"}</pre>
  }
  return(list(BF = c(BF), BFthres = c(BFthres), conclusion = result))
}
The "powerestimate" function can be used to calculate the approximate power using
equation (8).
##### powerestimate #####
powerestimate <- function(Delta, K, N, R2XkdotXminK, R2YdotXmink){</pre>
require("extraDistr")
```

```
ncp_d = 1 - ((1-R2XkdotXminK)*Delta^2 + R2YdotXmink)
ncp = Delta*sqrt(N*(1-R2XkdotXminK))/sqrt(ncp_d)
Tstatstar <- qt(0.05, N-K-1, ncp=ncp)
power = pht(Tstatstar, N-K-1, lower.tail=TRUE)
return(power)
}</pre>
```

The "equiv_corrZ" function (based on code for "equiv_corr" provided by Goertzen & Cribbie (2010)) can be used to calculate a p-value for the equivalence test for correlations based on Fisher's Z transformation.

```
equiv_corrZ<-function(var1, var2, delta_upper, delta_lower = NA) {
   if(is.na(delta_lower)){delta_lower <- (-delta_upper)}
   corxy <- cor(var1,var2)
   n <- length(var1)
   zei_lower <- log((1-delta_lower)/(1+delta_lower))/2
   zei_upper <- log((1+delta_upper)/(1-delta_upper))/2
   zcorxy <- log((1+corxy)/(1-corxy))/2
   equivt1_fz <- (zcorxy+ zei_lower)/(1/sqrt(n-3))
   pvalue1_fz <- 1-pnorm(equivt1_fz)
   equivt2_fz <- (zcorxy- zei_upper)/(1/sqrt(n-3))
   pvalue2_fz <- pnorm(equivt2_fz)
   the_results <- c(pvalue_equiv_z=max(c(pvalue1_fz, pvalue2_fz), na.rm=TRUE))
   return(the_results)
}</pre>
```

The "TOST₋d" function can be used to calculate a *p*-value for the valid (and invalid) equivalence test or differences between two independent means.

```
TOST_d <- function(outcome, group_id, delta_upper, delta_lower=NA, include_invalid=FALSE) {
   var1<-c(outcome);var2<-c(group_id);
   if(is.na(delta_lower)){delta_lower <-(-delta_upper)}
   zero <- unique(var2)[1]; one <- unique(var2)[2];
   n1 <- length(var1[var2==zero])
   n2 <- length(var1[var2==one])
   n <- n1 + n2
   m1 <- mean(var1[var2==zero])
   m2 <- mean(var1[var2==one])
   sd1 <- sd(var1[var2==zero])
   sd2 <- sd(var1[var2==one])</pre>
```

```
sd_p \leftarrow sqrt(((n1-1)*sd1^2 + (n2-1)*sd2^2)/(n1+n2-2))
t1 <-(m1-m2-delta_lower*sd_p)/(sd_p*sqrt(1/n1+1/n2))
t2 <- (m1-m2-delta_upper*sd_p)/(sd_p*sqrt(1/n1+1/n2))
# p-values for invalid test:
pval_lower <- pt(t1, n1+n2-2, lower.tail=FALSE)</pre>
pval_upper <- pt(-t2, n1+n2-2, lower.tail=FALSE)</pre>
# identical result is obtained using TOSTER package:
# toster <- TOSTtwo(m1=m1, m2=m2, sd1=sd1, sd2=sd2, n1=n1, n2=n2,
                      low_eqbound_d=delta_lower, high_eqbound_d=delta_upper,
                       alpha=0.05, plot=FALSE, verbose=FALSE, var.equal=TRUE)
# p-values for valid test:
 \texttt{cp1} < \texttt{-pt((m1-m2)/(sd_p*sqrt(1/n1+1/n2)), n1+n2-2, delta\_lower*sqrt(n1*n2/(n1+n2)), lower.tail=FALSE) } 
 \texttt{cp2} < -\texttt{pt(-(m1-m2)/(sd_p*sqrt(1/n1+1/n2)), n1+n2-2, -delta\_upper*sqrt(n1*n2/(n1+n2)), lower.tail=FALSE) } 
if(include_invalid) {output <- c(invalid=max(c(pval_lower,pval_upper)), valid=max(c(cp1,cp2)))}</pre>
if(!include_invalid) {output <- max(c(cp1,cp2))}</pre>
return(output)}
```

R-code for calculating equivalence testing p-values

Consider any random data:

library(mvtnorm)

pval

```
set.seed(123)
Sx <- matrix(c(4,2,2,3), ncol=2)
X <- cbind(1, rmvnorm(100, c(0,0), Sx))
y <- rnorm(100);
In R, we can obtain the p-value from equation (23) as follows:
lmmod <- summary(lm(y~X[,-1]));
N <- length(y);
K <- dim(cbind(X[,-1]))[2];
beta_hat <- lmmod$coef[,1];
SE_beta_hat <- lmmod$coef[,2];
pval <- 2*pt(abs(beta_hat/SE_beta_hat), N-K-1, 0, lower.tail=FALSE);</pre>
```

```
0.21936250 0.87118022 0.07558993
and the p-value from equation (24) as follows:
R2 <- lmmod$r.squared;
\label{eq:diffR2k} $$ $$ \dim(x) = \lim_{k \to \infty} (c(2:(K+1)), \ function(k) \ \{R2-summary(lm(y^X[,-k])) : squared\})); $$
pval \leftarrow pf((N-K-1)*(diffR2k/(1-R2)), 1, N-K-1, 0, lower.tail = FALSE);
# 0.87118022 0.07558993
We can obtain the p-values from equation (3) as follows:
equivBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.56906829 0.07356655 0.60764740
We can obtain the estimated standardized regression coefficients (equation (5)) in R
as follows:
b_vec <- (beta_hat*(apply(X,2,sd)/sd(y)))[-1];</pre>
b vec
# 0.01815061 -0.20050872
and obtain the p-values from equation (6) in R with the following code:
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.2263376 0.8121366
Finally, one can calculate the p-value from equation (11) in R with the following code:
K \leftarrow dim(X)[2] - 1
N <- length(y)
DELTA_P \leftarrow rep(0.01, K)
pval <- ncp <- R2YdotXmink <- R2XkdotXminK <- rep(0, K)</pre>
R2 <- summary(lm(y ~ X))$r.squared
diffR2k <- unlist(lapply(c(2:(K+1)), function(k)</pre>
{R2 - summary(lm(y ~ X[,-k]))$r.squared}));
pval <- rep(0,K);</pre>
for(k in 1:K){
  if(K>1)\{ \ Xmink <- \ cbind(cbind(X[,-1])[,-k])\} \\
  if(K==1){ Xmink <- rep(1,N)}
  R2YdotXmink[k] <- summary(lm(y ~ Xmink))$r.squared
  R2XkdotXminK[k] <- (summary(lm(cbind(X[,-1])[,k] ~ Xmink)))$r.squared;</pre>
  ncp_d = 1 - ((1-R2XkdotXminK[k])*DELTA_P[k] + R2YdotXmink[k])
  ncp[k] = sqrt(DELTA_P[k]*N*(1-R2XkdotXminK[k]))/sqrt(ncp_d)
```

```
pval[k] <- 1-pt(sqrt((N-K-1)*diffR2k[k])/sqrt(1-R2), N-K-1, ncp = ncp[k] , lower.tail = FALSE)
}
pval
# 0.2263376 0.8121366</pre>
```

R-code for "salaries" example

Results for the salaries analysis example can be obtained with the following R-code:

```
## Salaries example
library(carData)
### simple linear regression:
y <- Salaries$salary
X <- model.matrix(lm(salary ~ sex, data=Salaries))</pre>
equivBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 5000, DELTA_lower = -5000)$pval[2]
# 0.9632451
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.7804196
library(BayesFactor)
sdata <- data.frame(salary = Salaries$salary, sex = as.numeric(Salaries$sex) - 1)</pre>
regressionBF(salary ~ sex, data = sdata)
# 4.525
linearReg.R2stat(N = 397, p = 1, R2 = summary(lm(salary ~ sex, data=Salaries))$r.squared, simple = TRUE)
# 4.525
lmBF(salary ~ sex, data = Salaries)
# 6.177
lmBF(salary ~ sex, data = sdata)
# 4.525
### multivariable linear regression:
y <- Salaries$salary
X <- model.matrix(lm(salary ~ sex + yrs.since.phd + yrs.service + discipline + rank, data=Salaries))
# NHST p-values
summary(lm(salary ~ sex + yrs.since.phd + yrs.service + discipline + rank, data=Salaries))$coef[-1,4]
# 2.158412e-01 2.697855e-02 2.142543e-02 1.878412e-09 1.983251e-03 2.296130e-23
# EQUIV p-values
```

```
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.07552343 0.89174973 0.88518372 0.99980626 0.86805158 1.00000000

# Bayes Factors
BFs <- (BFstandardBeta(Y= y, Xmatrix=X[,-1])$BF)
BFs
# 3.860594e+00 7.352492e-01 6.036516e-01 1.540123e-07 7.331954e-02 5.592721e-21</pre>
```

R-code for "six key premises of mindset theory" example

Results presented in Table 3 can be obtained with the following R-code:

```
# Six key premises of mindset theory example
mind <- read.csv("Mindset Premise Study Data-Privacy Protected.csv")
mind[,"cog"] <- rowMeans(cbind(scale(mind[,"Cattell.Score"]), scale(mind[,"Letter.Sets.Score"])))</pre>
# Testing Premise 1: people with growth mind-sets hold learning goals
coefficients(summary(lm(X1.Learning.Goal~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X1.Learning.Goal"], mind[,"Mindset.Score"], 0.20, -Inf)
B <- equivstandardBeta(mind[,"X1.Learning.Goal"], mind[,"Mindset.Score"], 0.20, -Inf)[1:2]
res1 <- c(unlist(B)[1], Z, unlist(B)[2])
# Testing Premise 2: people with fixed mind-sets hold performance goals
coefficients(summary(lm(X2.Performance.Goal~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X2.Performance.Goal"], mind[,"Mindset.Score"], Inf, -0.2)
B <- equivstandardBeta(mind[,"X2.Performance.Goal"], mind[,"Mindset.Score"], Inf, -0.20)
res2 <- c(unlist(B)[1], Z, unlist(B)[2])
# Testing Premise 3: people with fixed mind- sets hold performance-avoidance goals
coefficients(summary(lm(X3.Performance.Avoidance.Goal~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X3.Performance.Avoidance.Goal"], mind[,"Mindset.Score"], Inf, -0.2)
B <- equivstandardBeta(mind[,"X3.Performance.Avoidance.Goal"], mind[,"Mindset.Score"], Inf, -0.20)
res3 <- c(unlist(B)[1], Z, unlist(B)[2])
# Testing Premise 4: people with fixed mind-sets believe that talent alone without effort creates success
coefficients(summary(lm(X4.Belief.in.Talent~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X4.Belief.in.Talent"], mind[,"Mindset.Score"], Inf, -0.2)
B <- equivstandardBeta(mind[,"X4.Belief.in.Talent"], mind[,"Mindset.Score"], Inf, -0.20)
res4 <- c(unlist(B)[1], Z, unlist(B)[2])
```

```
# Testing Premise 5: people with growth mind-sets persist to overcome challenges
coefficients(summary(lm(X5.Response.To.Challenge~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X5.Response.To.Challenge"], mind[,"Mindset.Score"], 0.20, -Inf)
B <- equivstandardBeta(mind[,"X5.Response.To.Challenge"], mind[,"Mindset.Score"], 0.20, -Inf)
res5 <- c(unlist(B)[1], Z, unlist(B)[2])
# Testing Premise 6: people with growth mind-sets are more resilient following failure
coefficients(summary(lm(X6.Raven.Test.Score~Mindset.Score, data=mind)))[2,]
Z <- equiv_corrZ(mind[,"X6.Raven.Test.Score"], mind[,"Mindset.Score"], 0.20, -Inf)
B <- equivstandardBeta(mind[,"X6.Raven.Test.Score"], mind[,"Mindset.Score"], 0.20, -Inf)
res6 <- c(unlist(B)[1], Z, unlist(B)[2])
# Testing Premise 6a: people with growth mind-sets are more resilient
# following failure when controlling for cognitive ability
B <- equivstandardBeta(mind[,"X6.Raven.Test.Score"], as.matrix(mind[,c("Mindset.Score", "cog")]), 0.20, -Inf)
res6a <- c(unlist(B)[1], Z, unlist(B)[3])
mindset_results <- round(rbind(res1, res2, res3, res4, res5, res6, res6a), digits=5)
mindset_results
       standard_beta.cbind(X[, -1]) pvalue_equiv_z
                                                      pval
                           0.09774
                                           0.01451 0.01331
#res1
                                           0.02577 0.02392
                           -0.10895
#res2
                                           0.00032 0.00028
#res3
                           -0.03914
                                           0.00159 0.00140
#res4
                           -0.06122
                                           0.00110 0.00096
#res5
                           0.05587
                           -0.12177
                                           0.00000 0.00000
#res6
#res6a
                           -0.05546
                                                NA 0.00000
```

R-code for "exposure to organic food" example

```
# Exposure to organic food example
dat1 <- read.csv("eskine - study 2 - mturk study - data_1.csv")
thedata <- dat1[dat1[,"F_all"]%in%1 & dat1[,"Condition"]%in%c(0,2),]

TOST_d(thedata[,"Morality"], thedata[,"Condition"]==2,0.48,include_invalid=TRUE)
# invalid valid
# 0.001401506 0.001223013</pre>
```

the test for the standardized regression coefficient:

```
s1 <- sd(as.numeric(thedata[,"Condition"]==2))
equivstandardBeta(thedata[, "Morality"], (thedata[,"Condition"]==2),
    DELTA_upper = 0.48*s1, DELTA_lower = -0.48*s1)$pval
# 0.0008520027</pre>
```

R-code for "hominid brain" example

Results for the hominid brain evolution example analysis example can be obtained with the following R-code:

```
## Hominid brain evolution example
### simple linear regression:
bailey <- read.csv("bailey2009.csv")</pre>
y <- bailey$cranium_capacity
X <- model.matrix(lm(cranium_capacity ~ temp_variation + isosd + parasites + population_dens, data=bailey))
# NHST p-values
summary(lm(y ~ temp_variation + isosd + parasites + population_dens, data=bailey))$coef[-1,4]
# 9.273608e-01
                 1.062033e-10 1.436866e-01
                                                1.287237e-16
# EQUIV p-values
equivstandardBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.01405927 0.99999931 0.27566257 1.00000000
# Bayes Factors
BFs <- (BFstandardBeta(Y= y, Xmatrix=X[,-1])$BF)
# 1.104066e+01 1.122479e-08 3.827830e+00 1.979351e-14
```