ORIGINAL RESEARCH PAPER

Equivalence testing for linear regression

Harlan Campbell

University of British Columbia, Department of Statistics

Vancouver, British Columbia, Canada

harlan.campbell@stat.ubc.ca

ARTICLE HISTORY

Compiled January 21, 2023

Abstract

We introduce equivalence testing procedures for linear regression analyses. Such tests can be very useful for confirming the lack of a meaningful association between a continuous outcome and a continuous or binary predictor. Specifically, we propose an equivalence test for unstandardized regression coefficients and an equivalence test for semipartial correlation coefficients. We review how to define valid hypotheses, how to calculate p-values, and how these tests compare to an alternative Bayesian approach with applications to various examples in the literature.

KEYWORDS

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

Author notes - 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 $\mathrm{H}_0 : \theta = 0$, 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 $\mathrm{H}_0 : \theta \notin (-\Delta, \Delta)$. 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 $(-\Delta, \Delta)$ is known as the "equivalence margin" and represents the range of values for which θ can be considered negligible.

Statistical methods for equivalence testing have their origins in the 1970s and 1980s (e.g., Westlake (1972), Schuirmann (1987), Anderson & Hauck (1983)). In psychology research and in the social sciences more broadly, the practice of equivalence testing is relatively new but is "rapidly expanding" (Koh & Cribbie, 2013). Recent examples of equivalence testing in the applied psychological research literature include Fruehauf et al. (2021) who use equivalence testing to study cognitive control in obsessive-compulsive disorder, and Leonidaki & Constantinou (2021) who use equiva-

lence testing in a study of cognitive behavioural therapy.

Statistical software for equivalence testing is also rapidly expanding. For researchers using R, the packages "equivalence" and "TOST" (Robinson & Robinson, 2016; Lakens, 2017) provide many accessible functions. For researchers using SAS, STATA, or SPSS, there are also many available resources; Batterham et al. (2016) (in their Appendix) provide a summary with examples.

In the psychological research methods literature, one early appearance of equivalence testing methods is Rogers et al. (1993) who discuss equivalence testing for group mean differences. More recent examples include Goertzen & Cribbie (2010) who highlight the importance of using equivalence tests to establish the independence of two different variables (i.e., for establishing negligible correlations) and Counsell et al. (2020) who consider equivalence testing methods for measurement invariance. Outside of psychology there is also a growing literature on equivalence testing methods. Two recent examples are Marcoulides & Yuan (2017) who consider equivalence testing for assessing structural equation models, and Leday et al. (2022) who review multivariate equivalence testing methods for food safety assessment.

Conspicuously absent is any published research on methods for equivalence testing in linear regression, with the notable exception of Dixon & Pechmann (2005) who propose equivalence tests for establishing negligible population trends in ecology. This is rather surprising since linear regression is arguably one of the most commonly used methods for statistical analysis. The first objective of this paper is therefore to address this research gap by outlining a general equivalence testing method for establishing negligible regression coefficients. In the first section of this paper ("Equivalence testing for unstandardized regression coefficients"), we review how to define valid hypotheses, calculate p-values, and establish "equivalence confidence intervals" (Seaman & Serlin, 1998) for equivalence tests of unstandardized regression coefficients.

Despite becoming more common and despite the fact that available software has made it more accessible, 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; 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 (e.g., Likert scales), researchers will often prefer to work with standardized effect sizes to aid with interpretation (Wilkinson, 1999; 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. For linear regression analyses, reporting standardized effect sizes is quite common (Bring, 1994; West et al., 2007) and the semipartial correlation coefficient is a standardized effect size that can be easily interpreted (Dudgeon, 2016). Therefore, our objective in the second section of this paper ("Equivalence testing for a standardized effect size in linear regression") is to establish an equivalence test for the semipartial correlation coefficient.

Several Bayesian methods (e.g., Morey & Rouder (2011), Rouder & Morey (2012), Bedrick & Hund (2018)) have been proposed for establishing equivalence. While the focus of this paper is frequentist equivalence testing, in the third section of this paper ("A Bayesian alternative for establishing equivalence in a linear regression"), we briefly review one of the proposed Bayesian alternatives for establishing equivalence in linear regression analyses.

Finally, in the fourth section of this paper ("Practical Examples"), we demon-

strate how all of the different testing methods can be applied in practice with a number of practical examples. We then conclude with some general recommendations on how to perform equivalence testing for linear regression. In the supplemental material, R code is available to implement all of the calculations and analyses.

2. Equivalence testing for unstandardized regression coefficients

Consider a multiple linear regression where Y is the outcome variable and X is the $N \times (K+1)$ fixed predictor 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 predictor.

Note that the regression may include both categorical and continuous predictors. 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 predictor corresponding to the student's age in years; and X_3 might be a continuous predictor 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, for i=1,...,N:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + \epsilon_i, \text{ and} \tag{1}$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2), \tag{2}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ 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 $\widehat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_K})^T$, and $\hat{\sigma}^2$; see equations (22) and (23) which are provided in

the supplemental material 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 given a fixed age and 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:

$$\begin{aligned} &\mathrm{H}_0 : \beta_k \leq \Delta_{k,lower} \quad \text{or} \quad \beta_k \geq \Delta_{k,upper}, \quad \text{vs.} \\ &\mathrm{H}_1 : \beta_k > \Delta_{k,lower} \quad \text{and} \quad \beta_k < \Delta_{k,upper}, \end{aligned} \tag{3}$$

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 situations, 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 predictors, $k = 2$ 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?

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 inverting a confidence interval. For example, we will reject the above null hypothesis ($\mathrm{H}_0 : \beta_k \leq \Delta_{k,lower}$ or $\beta_k \geq \Delta_{k,upper}$), at a α significance level, whenever a (1 - 2α)% 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:

$$p_k^{lower} = 1 - F_t\left(\frac{\hat{\beta}_k - \Delta_{k,lower}}{\mathrm{SE}(\hat{\beta}_k)}, N - K - 1\right), \quad \text{and}$$
$$p_k^{upper} = 1 - F_t\left(\frac{\Delta_{k,upper} - \hat{\beta}_k}{\mathrm{SE}(\hat{\beta}_k)}, N - K - 1\right), \tag{4}$$

for k in 0,...,K; where $F_t(\quad \cdot \quad ; df)$ denotes the cumulative distribution function (cdf) of the t-distribution with df degrees of freedom, and where $\mathrm{SE}(\hat{\beta}_k) = \hat{\sigma}\sqrt{[(X^T X)^{-1}]_{kk}}$. In order to reject the equivalence test null hypothesis ($\mathrm{H}_0 : \beta_k \leq \Delta_{k,lower}$ or $\beta_k \geq \Delta_{k,upper}$), 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: $p\text{-value}_k = \max(p_k^{lower}, p_k^{upper})$.

An *a priori* sample size calculation for this equivalence test can be performed using the following analytic formula to obtain a reasonable approximation of the equivalence test's statistical power (Zhang, 2003):

$$power = F_t\Big(\frac{\Delta_{k,lower} - \beta_k}{\mathrm{SE}(\hat{\beta}_k)} - t^*_{1-\alpha}, N - K - 1\Big) - F_t\Big(\frac{\Delta_{k,upper} - \beta_k}{\mathrm{SE}(\hat{\beta}_k)} + t^*_{1-\alpha}, N - K - 1\Big), \tag{5}$$

where $t^*_{1-\alpha}$ is the $(1-\alpha)$th percentile of a t-distribution with $N - K - 1$ degrees of freedom, and β_k and $\mathrm{SE}(\hat{\beta}_k)$ are set to whatever values are assumed to be true *a priori*. Note that if X_k is uncorrelated with the other predictors, $\mathrm{SE}(\hat{\beta}_k) = \sigma/(\sigma_k\sqrt{N})$, where σ_k is the standard deviation of the k-th predictor, X_k, for k in 1,...,K.

To exemplify the above power calculation and TOST procedure, we return to

our hypothetical anxiety study example. The parameter of primary interest in this example is β_1, the effect of counselling services on the anxiety score. As noted earlier, suppose the researcher has defined $\Delta_{1,lower} = -2$ and $\Delta_{1,upper} = 2$. Suppose also that the researcher has decided to collect data from $N = 40$ participants, randomly assigning half to receive counselling and the other half to not receive counselling. Finally, based on what is known about the typical variability of scores obtained with the anxiety assessment questionnaire, suppose the researcher reasonably assumes, *a priori*, that $\sigma = 2$. Then, to approximate the power of the equivalence test, we can set $\text{SE}(\hat{\beta}_1) = \sigma/(\sigma_1\sqrt{N}) = 2/(0.5 \times \sqrt{40}) = 0.63$, and, using equation (5), calculate that the study will have about 85% power to reject the equivalence null hypothesis ($\text{H}_0 : \beta_1 \notin (-2, 2)$) if the true value of β_1 is in fact zero (i.e., if the counselling services truly have no effect on anxiety scores):

$$power = F_t\Big(\frac{-2-0}{0.63} - 1.69, 40-3-1\Big) - F_t\Big(\frac{2-0}{0.63} + 1.69, 40-3-1\Big) = 0.85. \quad (6)$$

Now suppose the researcher conducts the study (see full dataset in Table 6 of the supplemental material) and obtains the following results:

$$\begin{aligned}
\hat{\beta}_0 = 3.69; \quad \text{SE}(\widehat{\beta_0}) = 3.34; \quad & 90\%\text{CI} = (-1.94, 9.31); \\
& 95\%\text{CI} = (-3.08, 10.45); \\
\hat{\beta}_1 = -0.57; \quad \text{SE}(\widehat{\beta_1}) = 0.66; \quad & 90\%\text{CI} = (-1.68, 0.54); \\
& 95\%\text{CI} = (-1.90, 0.76); \\
\hat{\beta}_2 = 0.84; \quad \text{SE}(\widehat{\beta_2}) = 0.20; \quad & 90\%\text{CI} = (0.49, 1.18); \\
& 95\%\text{CI} = (0.42, 1.24); \\
\hat{\beta}_3 = -1.93 \times 10^{-5}; \quad \text{SE}(\widehat{\beta_3}) = 1.77 \times 10^{-5}; \quad & 90\%\text{CI} = (-4.91, 1.05) \times 10^{-5}; \\
& 95\%\text{CI} = (-5.51, 1.65) \times 10^{-5};
\end{aligned}$$

These results suggest that, on average, older students obtain higher anxiety scores ($\hat{\beta}_2 = 0.84$), and students from wealthier households obtain lower scores ($\hat{\beta}_3 = -1.93 \times 10^{-5}$; about 2 points lower for every \$100,000 increase in household income). In order

to test whether β_1 is at most negligible, one calculates, from equation (4):

$$p_1^{lower} = 1 - F_t\left(\frac{\widehat{\beta_1} - (-2)}{\text{SE}(\widehat{\beta_1})}, N - K - 1\right) = 1 - F_t\left(2.16, 36\right) = 0.0178, \quad \text{and}$$
$$p_1^{upper} = 1 - F_t\left(\frac{2 - \widehat{\beta_1}}{\text{SE}(\widehat{\beta_1})}, N - K - 1\right) = 1 - F_t\left(3.92, 36\right) = 0.0002,$$

such that:

$$p - \text{value}_1 = \max(p_1^{lower}, p_1^{upper}) = 0.0178.$$

If a nominal significance level of $\alpha = 0.05$ is to be used for the equivalence test, then we can reject the equivalence null hypothesis ($\text{H}_0 : \beta_1 \notin (-2, 2)$) and plausibly conclude that, when controlling for age and wealth, any difference on the anxiety score between those students receiving counselling and those not receiving counseling is smaller than 2 points (p-value=0.0178).

The "least equivalent allowable difference" or "equivalence confidence interval"

Seaman & Serlin (1998) and Meyners (2007) suggest that researchers, instead of only reporting an equivalence test p-value, should also report the smallest possible absolute value at which one could have claimed equivalence. Meyners (2007) calls this the "least equivalent allowable difference" (LEAD), while Seaman & Serlin (1998) refer to this as the "equivalence confidence interval." To illustrate the concept, we return one again to our hypothetical anxiety study.

Instead of simply concluding that any difference on the anxiety score between those students receiving counselling and those not receiving counseling is smaller than 2 points with p-value = 0.0178, we could also report that any difference less than 1.68 could also have been ruled out (at the nominal significance level of $\alpha = 0.05$). Indeed, prior to having observed the data, had we defined the equivalence margin to be (-1.68, 1.68) instead of (-2, 2), then we could have rejected the equivalence null hypothesis with

p-value=0.05. In this case, the LEAD is equal to 1.68 and the "equivalence confidence interval" is simply (-1.68,1.68).

Note that the 90% CI for β_1 is (-1.68, 0.54). The -1.68 lower bound is no coincidence. The LEAD can be calculated as the maximum of the absolute value of the bounds of the $(1 - 2\alpha)$%CI. For instance, in our example, we have that LEAD= $\max(|-1.68|, |0.54|) = 1.68$. Despite it's simplicity, Meyners (2007) argues that the LEAD is worth reporting since it depends only on the α significance level and thereby enables readers to draw their own conclusions irrespective of the equivalence margin that a particular researcher might choose.

3. Equivalence testing for a standardized effect size in linear regression

Unstandardized regression coefficients are often difficult to interpret since both the predictors and the outcome can be measured on arbitrary units with no objective meaning. As a result, researchers may prefer to report standardized effect sizes. 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 example, 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. To explain why, let us consider a two-sample equivalence test for the difference in means.

Suppose that, 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 supplemental material 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 supplemental material, we show results from a small simulation study (Simulation Study 1) 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. This is likely the result of failing to account for the uncertainty involved in estimating the standard deviation.

The most commonly used standardized effect sizes for linear regression analyses are the standardized regression coefficient and the semipartial correlation coefficient (Courville & Thompson, 2001; Dudgeon, 2016). However, as Dudgeon (2016) notes, the popularity of the standardized regression coefficient "is arguably a product of convention rather than any perceived intrinsic merit of the standardized regression coefficient as an effect size." Indeed, many researchers argue that the standardized regression coefficient is difficult to interpret (and problematic when it comes comparing effect sizes across different studies) since it does not appropriately partition variance when predictors are correlated (Kanetkar et al., 1995; Tonidandel & LeBreton, 2011; Aloe & Becker, 2012). Levine et al. (2008) explain as follows: "the common practice of interpreting [the standardized regression coefficient] as analogous to [the correlation]

[...] can be misleading because [the standardized regression coefficient] can be near zero even when the predictor explains a substantial amount of the variance in the outcome variable when other predictors correlated with the predictor claim the shared variance." For a more detailed explanation see Disbato (2016) and Darlington & Hayes (2017) who argue that the semipartial correlation coefficient is a much better effect size. With this in mind, we propose that researchers use an equivalence test for the semipartial correlation coefficient.

An equivalence test for the semipartial correlation coefficient

The semipartial correlation coefficient, sr_k, is a parameter taking values between -1 and 1, that measures the strength of the association between the outcome, Y, and the predictor, X_k, that is independent of any linear relationship between X_k and the other predictors in the model, for k in 1,...,K. We define sr_k as follows, for k in 1,...,K:

$$sr_k = \left(\beta_k \frac{\sigma_k}{\sigma_Y}\right) \times \sqrt{1 - R^2_{X_k X_{-k}}}, \tag{7}$$

where σ_Y is the standard deviation of Y, σ_k is the standard deviation of X_k, and $R^2_{X_k X_{-k}}$ is the coefficient of determination from the linear regression of X_k predicted from the other $K-1$ predictors (see full details in supplemental material). Note that $\mathcal{B}_k = (\beta_k \sigma_k / \sigma_Y)$ is equal to the k-th standardized regression coefficient, and that $1 - R^2_{X_k X_{-k}}$ is known as the k-th "tolerance", representing the proportion of variation in X_k that is linearly unrelated to all the other predictors in the regression model. To be clear, when $K = 1$ (or when X_k and X_{-k} are perfectly uncorrelated), the tolerance will equal 1, and therefore we have that: $sr_k = \mathcal{B}_k = \text{cor}(Y, X_k)$.

Note that the *squared* semipartial correlation, sr_k^2, can be understood as the amount of variance in Y that is uniquely explained by the k-th predictor, X_k, since:

$$sr_k^2 = \left(\beta_k \frac{\sigma_k}{\sigma_Y}\right)^2 \times \left(1 - R^2_{X_k X_{-k}}\right) = R^2_{YX} - R^2_{Y X_{-k}}, \tag{8}$$

where R^2_{YX} is the coefficient of determination from the linear regression of Y predicted

from X, and where $R^2_{YX_{-k}}$ is the coefficient of determination from the linear regression of Y predicted from all but the k-th predictor.

J. Cohen (1988)'s well known rules of thumb for interpreting the magnitude of correlation coefficients (small=0.1, medium=0.3, large=0.5) can be applied for interpreting semipartial correlation coefficients. However, we note that P. Cohen et al. (2013) offer alternative values that are slightly larger: small effects may be defined as $sr_k = 0.14$ (or equivalently $sr_k^2 = 0.02$), medium effects as $sr_k = 0.39$ (or $sr_k^2 = 0.15$), and large effects as $sr_k = 0.59$ (or $sr_k^2 = 0.35$). Determining what exact values are ideal is beyond the scope of this paper, but interested readers are referred to Hemphill (2003), Funder & Ozer (2019), and Lovakov & Agadullina (2021).

Dudgeon (2016) propose using the adjusted Aloe-Becker large-sample confidence interval for sr_k which can be calculated, for k in 1,...,K, as:

$$(1-\alpha)\%\text{CI for } sr_k = [\widehat{sr_k} - t^*_{1-\alpha/2}\text{SE}(\widehat{sr_k}), \quad \widehat{sr_k} + t^*_{1-\alpha/2}\text{SE}(\widehat{sr_k})], \tag{9}$$

where:

$$\widehat{sr_k} = (\hat{\beta}_k \frac{\hat{\sigma}_k}{\hat{\sigma}_Y}) \times \sqrt{1 - \hat{R}^2_{X_k X_{-k}}}, \tag{10}$$

and

$$\text{SE}(\widehat{sr_k}) = \sqrt{\frac{\hat{R}^4_{YX} - 2\hat{R}^2_{YX} + \hat{R}^2_{YX_{-k}} + 1 - \hat{R}^4_{YX_{-k}}}{N-K-1}}, \tag{11}$$

where $\hat{R}^2_{YX}$, $\hat{R}^2_{YX_{-k}}$, and $\hat{R}^2_{X_k X_{-k}}$ are estimates for R^2_{YX}, $R^2_{YX_{-k}}$, and $R^2_{X_k X_{-k}}$, respectively, obtained from the observed data.

An equivalence test for the k-th semipartial correlation coefficient can be defined by the following null and alternative hypotheses:

$\text{H}_0 : sr_k \leq \Delta_{k,lower}$ or: $sr_k \geq \Delta_{k,upper}$, vs.

$\text{H}_1 : sr_k > \Delta_{k,lower}$ and: $sr_k < \Delta_{k,upper}$,

where the equivalence margin is $(\Delta_{k,lower}, \Delta_{k,upper})$, for k in $1, \ldots, K$. By inverting the adjusted Aloe-Becker large-sample confidence interval, we can conduct two one-sided t-tests (TOST) with the following p-values, for k in 1,...,K:

$$\mathbb{p}_k^{lower} = 1 - F_t\left(\frac{\widehat{sr}_k - \Delta_{k,lower}}{\mathrm{SE}(\widehat{sr}_k)}; df = N - K - 1\right), \quad \text{and}$$
$$\mathbb{p}_k^{upper} = 1 - F_t\left(\frac{\Delta_{k,upper} - \widehat{sr}_k}{\mathrm{SE}(\widehat{sr}_k)}; df = N - K - 1\right). \tag{12}$$

Therefore, for the k-th predictor, the null hypothesis, $\mathrm{H}_0 : sr_k \leq \Delta_{k,lower}$ or: $sr_k \geq \Delta_{k,upper}$, is rejected if and only if the p-value, $p\text{-value}_k = \max(\mathbb{p}_k^{lower}, \mathbb{p}_k^{upper})$, is less than α.

An *a priori* sample size calculation for this equivalence test can be performed using the following analytic formula to obtain a reasonable approximation of the equivalence test's statistical power:

$$power = F_t\Big(\frac{\Delta_{k,lower} - sr_k}{\mathrm{SE}(\widehat{sr}_k)} - t^*_{1-\alpha}, N-K-1\Big) - F_t\Big(\frac{\Delta_{k,upper} - sr_k}{\mathrm{SE}(\widehat{sr}_k)} + t^*_{1-\alpha}, N-K-1\Big), \tag{13}$$

where sr_k and $\mathrm{SE}(\widehat{sr}_k)$ are whatever values are assumed to be true *a priori*. Note that if one assumes that $sr_k = 0$, then $\mathrm{SE}(\widehat{sr}_k) = \sqrt{1/(N-K-1)}$.

In the supplemental material, we conduct two small simulation studies, Simulation Study 2 and Simulation Study 3, to investigate the proposed methods. The first shows that, when $K = 1$, the proposed equivalence test for semipartial correlation coefficients and a commonly used equivalence test for correlations (based on Fisher's Z transformation) provide, as expected, very similar results. The second simulation study confirms that the type 1 error obtained with the proposed equivalence test for semipartial correlation coefficients is correct and that the proposed formula for power calculation (equation (13)) provides a reasonable approximation of the true statistical power. The results also suggest that large sample sizes (much larger than those typically encountered in psychological studies (Kühberger et al., 2014; Fraley & Vazire, 2014; Marszalek et al., 2011)) are required for the equivalence tests to have non-negligible statistical power. Goertzen & Cribbie (2010) reached a similar conclu-

sion.

4. A Bayesian alternative for establishing equivalence in a linear regression

As noted in the Introduction, there are a several different Bayesian methods available for establishing equivalence. Rouder & Morey (2012)'s proposed "default" Bayes factor (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 factor 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:

$$\text{BF}_{10} = \frac{\Pr(Data \,|\, Model\ 1)}{\Pr(Data \,|\, Model\ 0)} = \frac{\Pr(\ Model\ 1|Data\ ) \times \Pr(\ Model\ 1)}{\Pr(\ Model\ 0|Data\ ) \times \Pr(\ Model\ 0)}, \tag{14}$$

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, with equal prior model probabilities, a BF_{10} equal to 0.20 indicates that the null model is five times more likely than the alternative model. Going forward, we suppose that equal prior model probabilities ($\Pr(\text{Model } 0) = \Pr(\text{Model } 1) = 0.5$) are always assumed, as is often (implicitly) done in practice; but see Tendeiro & Kiers (2019), and Campbell & Gustafson (2022a) for discussion of this practice.

Bayesian methods require one to define appropriate prior distributions for all model parameters (Consonni & Veronese, 2008) and Rouder & Morey (2012) suggest using Jeffreys-Zellner-Siow (JZS) "objective priors". A version of this prior setup, whereby the so-called r scale parameter is set equal to a specific value, allows one to specify prior beliefs about the magnitude of standardized regression coefficients (i.e.,

specify the *a priori* distribution of $\mathcal{B}_k = (\beta_k \sigma_k / \sigma_Y)$, for k in 1,...,K). For instance, the BayesFactor R package uses the scaled-JZS prior setup with a default of $r = \sqrt{2}/4 = 0.354$, corresponding to a prior belief in a 50% probability that $|\mathcal{B}_k| > 0.354$, for k in 1,...,K. While perhaps "computationally convenient", researchers who have difficulty interpreting the magnitude of standardized regression coefficients (e.g., Disbato (2016)) will no doubt be particularly challenged when it comes to defining an appropriate value for r for such a prior.

To test the k-th regression coefficient, in a multiple linear regression model, one computes a Bayes factor for a model that includes the k-th predictor against a model that does not, such that:

$$\text{Model } 0: \quad Y_i \sim \text{Normal}(X_{i,-k}^T \beta_{-k}, \sigma^2), \qquad \forall i = 1, ..., N; \tag{15}$$

$$\text{Model } 1: \quad Y_i \sim \text{Normal}(X_{i\times}^T \beta, \sigma^2), \qquad \forall i = 1, ..., N; \tag{16}$$

where β_{-k} ($X_{i,-k}$) is the vector (matrix) of regression coefficients (predictors), with the k-th coefficient (predictors) omitted.

If this Bayes factor were to be above a certain threshold (e.g., if $\text{BF}_{10} > 6$), one would conclude with a "positive" finding 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 $\text{BF}_{10} < 1/6$), one would conclude with a "negative" finding that there is evidence for $\beta_k = 0$ (i.e., evidence in support of Model 0). Finally, if this Bayes factor were neither above or bellow the certain threshold (e.g., if $1/6 < \text{BF}_{10} < 6$), one would conclude with a "inconclusive" finding that there is insufficient evidence to support either model.

Campbell & Gustafson (2018) discuss a similar frequentist way to categorize one's results as "positive", "negative", or "inconclusive". Testing a parameter θ under the so-called "conditional equivalence testing" (CET) scheme would proceed as follows. If a first p-value, p_{NHST}, obtained from testing $\text{H}_0 : \theta = 0$, is less than the type 1 error α-threshold (e.g., if $p_{NHST} < 0.05$), one concludes with a "positive" finding: θ is significantly different than 0. On the other hand, if the first p-value, p_{NHST}, is

greater than α and a second p-value, p_{EQUIV}, obtained from testing $\mathrm{H}_0 : |\theta| > \Delta$, 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. Finally, if both p-values are larger than α, the result is "inconclusive". In two of the practical examples we review in the next section we will have the opportunity to see how the BF and CET based categorizations compare.

5. 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 multiple 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 sex (X_1, where "0" corresponds to "female," and "1" corresponds to "male."). Standard least squares estimation results in the following parameter estimates: $\widehat{\beta}_0 = 101002$, $\text{SE}(\widehat{\beta}_0) = 4809$, and $\widehat{\beta}_1 = 14088$, $\text{SE}(\widehat{\beta}_1) = 5065$; $\widehat{\sigma} = 30034.61$; $\widehat{sr}_1 = 0.14$, $\text{SE}(\widehat{sr}_1) = 0.05$.

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, $\text{H}_0 : |\beta_1| \geq 5000$ vs. $\text{H}_1 : |\beta_1| < 5000$, can be calculated following equation (4). We obtain a $p - \text{value}_1 = \max(p_1^{lower}, p_1^{upper}) = \max(0.00009, 0.963) = 0.963$ 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 semipartial correlation coefficient, sr_1. Suppose we consider anything less than "small" to be negligible and therefore define the equivalence margin as (-0.10, 0.10). Then we can calculate a p-value for $\text{H}_0 : |sr_1| \geq 0.10$ vs. $\text{H}_1 : |sr_1| < 0.10$ as per equation (12). We obtain: p-value $= \max(\mathbb{p}_1^{lower}, \mathbb{p}_1^{upper}) = 0.783$, where:

$$\mathbb{p}_k^{lower} = 1 - F_t\left(\frac{\widehat{sr}_k - \Delta_{k,lower}}{\text{SE}(\widehat{sr}_k)}; df = N - K - 1\right), \tag{17}$$

$$= 1 - F_t\left(\frac{0.14 + 0.10}{0.05}; df = 397 - 1 - 1\right), \tag{18}$$

$$= 1 - F_t\left(2.78; df = 395\right), \tag{19}$$

$$< 0.001 \tag{20}$$

and:

$$\begin{aligned}
\mathbb{p}_k^{upper} &= 1 - F_t\left(\frac{\Delta_{k,upper} - \widehat{sr}_k}{\text{SE}(\widehat{sr}_k)}; df = N - K - 1\right), \\
&= 1 - F_t\left(\frac{0.10 - 0.14}{0.05}; df = 397 - 1 - 1\right), \\
&= 1 - F_t\left(-2.78; df = 395\right), \\
&= 0.783
\end{aligned}$$

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 $\text{BF}_{10} = 4.5$ 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 $\text{BF}_{10} = 6.2$ 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]

Multiple linear regression

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

$$Y \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6, \sigma^2), \qquad (21)$$

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 ($X_5 = 0$, $X_6 = 0$)

[1]The 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 supplemental material) 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.

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 LEAD values for the semipartial correlation coefficients (with $\alpha = 0.05$). Table 2 lists the p-values for each of the hypothesis tests we consider as well as Bayes factors. An equivalence margin of (-0.10, 0.10) is used for the equivalence testing of the semipartial correlation coefficients and 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 CET would also be "inconclusive" (for $\alpha = 0.05$ 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 CET and Bayes factor approaches do not entirely agree for the other predictors, see Table 2. For both the "years since Ph.D" ($k = 2$) and the "years of service" ($k = 3$) predictors, the frequentist CET obtains a positive result whereas the Bayes factor obtains an inconclusive result.

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

k	predictor	β_k	SE($\hat{\beta}_k$)	$\widehat{sr}_k$	SE($\widehat{sr}_k$)	LEAD($\widehat{sr}_k$)
0	intercept	65955.23	4588.60	-	-	-
1	sex (male)	4783.49	3858.67	0.046	0.037	0.108
2	years since Ph.D.	535.06	240.99	0.083	0.037	0.145
3	years of service	-489.52	211.94	-0.086	0.037	0.148
4	discipline (applied)	14417.63	2342.88	0.230	0.037	0.291
5	rank (Asst. Prof.)	12907.59	4145.28	0.116	0.037	0.178
6	rank (Prof.)	45066.00	4237.52	0.398	0.036	0.457
				$\hat{\sigma} = 22538.65$		$R^2_{Y,X} = 0.455$

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

k	$\widehat{sr}_k$	p_{NHST}	p_{EQUIV} $\Delta = 0.10$	BF_{10} $r = \sqrt{2}/4$	CET conclusion $\alpha = 0.05$	Bayesian conclusion BF threshold = 6
1	0.046	0.216	0.076	1/3.9	Inconclusive	Inconclusive
2	0.083	0.027	0.325	1.4	Positive	Inconclusive
3	-0.086	0.021	0.358	1.7	Positive	Inconclusive
4	0.230	< 0.001	1.000	6.5×10^6	Positive	Positive
5	0.116	0.002	0.670	13.6	Positive	Positive
6	0.398	< 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 multiple linear regression model.

"mindset score" and used a non-inferiority test (i.e., a one-sided equivalence test) to determine whether the correlation 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). Burgoyne et al. (2020) justify this choice of margin by citing Richard et al. (2003) and explaining that "effects described as profound should at least meet the mean effect size in social-psychological research."

Burgoyne et al. (2020) used the test for correlations proposed by Goertzen & Cribbie (2010) based on Fisher's Z transformation (see details of this test in supplemental material). We calculated p-values for each of the six regressions based instead on our proposed test for semipartial correlation coefficients (recall that when $K = 1$, $sr_k = \mathcal{B}_k = \text{cor}(Y, X_k)$). In Table 3, the p-values calculated based on equation (12) 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.

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 multiple linear regression and Goertzen & Cribbie (2010)'s test for correlations is therefore not applicable. Our proposed test for semipartial correlation 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 (12) as $p_{sr} < 0.001$.

	H_0	$Y \sim X$	$\widehat{sr}_1$	p_Z	p_{sr}
1.	$sr_1 \geq 0.2$	Learning goals ∼ Mindset	0.098	0.015	0.016
2.	$sr_1 \leq -0.2$	Performance goals ∼ Mindset	-0.109	0.026	0.027
3.	$sr_1 \leq -0.2$	Performance avoidance goals ∼ Mindset	-0.039	<0.001	<0.001
4.	$sr_1 \leq -0.2$	Belief in talent alone ∼ Mindset	-0.061	0.002	0.002
5.	$sr_1 \geq 0.2$	Response to challenge ∼ Mindset	0.056	0.001	0.001
6.	$sr_1 \geq 0.2$	Raven failure score ∼ Mindset	-0.122	<0.001	<0.001
7.	$sr_1 \geq 0.2$	Raven failure score ∼ Mindset + Cognitive ability	-0.055	—	<0.001

Table 3. For each of the regression analyses fit by Burgoyne et al. (2020), p_{sr} indicates the p-value for the equivalence test based on equation (12), and p_Z indicates the p-value for the equivalence test based on Fishers 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 $\Delta_{upper} = \infty$).

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 cm^3, which ranges from $475cm^3$ to $1{,}880cm^3$. We also consider 4 predictors of interest: (1) the local climate variation (X_1 : "local climate", in degrees Celsius ranging between 8 and 47); (2) the global average temperature (X_2 : "global climate", in standard deviations, ranging between 0.21 and 0.43); (3) the parasite load, (X_3 : "parasites", integer between 0 and 7); and (4) the population density, (X_4 : "pop. density", integer between 13 and 141). Details are available in Bailey & Geary (2009).

We fit the data with a multiple 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 LEAD values for the semipartial correlation coefficients (with $\alpha = 0.05$). Table 5 lists the p-values for each of the hypothesis tests we consider as well as Bayes factors. An equivalence margin of (-0.10, 0.10) is used for equivalence testing of the semipartial correlation coefficients. 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 $k = 1, 2, 3$, and 4, respectively.[2]

With $\alpha = 0.05$ and a Bayes factor threshold of 6, each of the frequentist CET conclusions match each of the Bayesian conclusions. Therefore, regardless of what approach is used, one can conclude that the data indicate evidence for the effect of population density ($p_{NHST} < 0.001$; $\text{BF}_{01} = 5.1 \times 10^{13}$ (with $r = \sqrt{2}/4$)) and for the effect of global climate ($p_{NHST} < 0.001$; $\text{BF}_{01} = 8.9 \times 10^7$ (with $r = \sqrt{2}/4$)), and evidence for a lack of effect of local climate (p_{EQUIV}=0.010 (with Δ=0.10); $BF_{01} = 1/11$ (with $r = \sqrt{2}/4$)). Results are inconclusive with regards to the effect (or lack thereof) of parasites (p_{NHST}= 0.144, p_{EQUIV}=0.171 (with Δ=0.10); $BF_{01} = 1/3.8$ (with $r = \sqrt{2}/4$)).

k	predictor	β_k	SE($\hat{\beta}_k$)	$\widehat{sr}_k$	SE($\widehat{sr}_k$)	LEAD($\widehat{sr}_k$)
0	intercept	261.70	97.72	-	-	-
1	local climate	0.15	1.62	0.004	0.041	0.072
2	global climate	1871.75	271.82	0.284	0.043	0.355
3	parasites	-9.26	6.30	-0.061	0.041	0.129
4	pop. density	4.43	0.48	0.379	0.044	0.452
				$\hat{\sigma} = 168.7$		$R^2_{Y,X} = 0.711$

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

[2]Note 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).

k	$\widehat{sr}_k$	p_{NHST}	p_{EQUIV} $\Delta = 0.10$	BF_{10} $r = \sqrt{2}/4$	CET conclusion $\alpha = 0.05$	Bayesian conclusion BF threshold = 6
1	0.004	0.927	0.010	1/11	Negative	Negative
2	0.284	< 0.001	1.000	8.9×10^7	Positive	Positive
3	-0.061	0.144	0.171	1/3.8	Inconclusive	Inconclusive
4	0.379	< 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 multiple linear regression model.

Conclusion

Researchers require statistical tools that allow them to reject the presence of meaningful 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 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).

The use of equivalence/non-inferiority tests should not rule out the complementary use of confidence intervals (whereas the use of Bayes factors may indeed rule out the use of certain credible intervals; see Campbell & Gustafson (2022b)). Indeed, confidence intervals can be extremely useful for highlighting the stability (or lack of stability) of a given estimator (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, or perhaps by reporting the "least equivalent allowable difference" (LEAD or "equivalence confidence interval") as recommended by Meyners (2007).

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 wholeheartily recommended for all cases, their use is not unlike the use of "default priors" for defining Bayes factors which have indeed proven useful to researchers in many scenarios.

In the practical examples we showed that testing with Bayes factors and testing with frequentist equivalence tests will often, but not always, lead to similar conclusions. The pros and cons of frequentist versus Bayesian testing methods are a topic of great debate (Campbell & Gustafson, 2021a).

Note that our proposed equivalence tests are limited to comparing two models for which the difference in degrees of freedom is 1. In other words, the tests are 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 tests to compare a smaller model with only "sex" as a predictor, with a larger model that includes "sex," "discipline" and "rank," as predictors. 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 also note that the TOST approach we proposed is not necessarily optimal in the sense that other procedures may have slightly higher power. For instance, Anderson & Hauck (1983) proposed the so-called "power method" as an alternative to the TOST approach (but note that Frick (1987) and Müller-Coors (1990) expressed concerns that the actual type I error rate of the power method may exceed the nominal level). More recently, Romano (2005) proposed what they call the "optimal equivalence test" (based on the folded-Normal distribution) as a more powerful alternative to TOST (see also Möllenhoff et al. (2022)).

Finally, there is certainly potential to expand equivalence testing for other analyses including for logistic regression and time-to-event models. These are objectives for

future research.and will help to further "extend the arsenal of confirmatory methods rooted in the frequentist paradigm of inference" (Wellek, 2017).

Available Code - All the code used in this paper and relevant materials are made available in an OSF repository: DOI 10.17605/OSF.IO/5YR92

References

Aloe, A. M., & Becker, B. J. (2012). An effect size for regression predictors in meta-analysis. *Journal of Educational and Behavioral Statistics*, *37*(2), 278–297.

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

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.

Batterham, M. J., Van Loo, C., Charlton, K. E., Cliff, D. P., & Okely, A. D. (2016). Improved interpretation of studies comparing methods of dietary assessment: combining equivalence testing with the limits of agreement. *British Journal of Nutrition*, *115*(7), 1273–1280.

Bedrick, E. J., & Hund, L. (2018). An approach for quantifying small effects in regression models. *Statistical Methods in Medical Research*, *27*(4), 1088–1098.

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 (accepted for publication in Psychological Methods)*.

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., & Gustafson, P. (2022a). Bayes factors and posterior estimation: Two sides of the very same coin. *The American Statistician*, *0*(0), 1-11. doi:

Campbell, H., & Gustafson, P. (2022b). Defining a credible interval is not always possible with "point-null" priors: A lesser-known consequence of the Jeffreys-Lindley paradox. *arXiv preprint arXiv:2210.00029*.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York, NY: Routledge Academic.

Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences.* Psychology press.

Consonni, G., & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, *23*(3), 332–353. doi: 10.1214/08-STS258.

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research*, *55*(2), 312–328.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: $\mathcal{B}$ is not enough. *Educational and Psychological Measurement*, *61*(2), 229–248.

Darlington, R. B., & Hayes, A. F. (2017). Regression analysis and linear models. *New York, NY: Guilford*, 603–611.

Disbato, D. (2016). *On effect sizes in multiple regression.* http://www.daviddisabato.com/blog/2016/4/8/on-effect-sizes-in-multiple-regression.

Dixon, P. M., & Pechmann, J. H. (2005). A statistical test to show negligible trend. *Ecology*, *86*(7), 1751–1756.

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. (2016). A comparative investigation of confidence intervals for independent variables in linear regression. *Multivariate Behavioral Research*, *51*(2-3), 139–153.

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.

Frick, H. (1987). On level and powee of Anderson and Hauck's peocedure for testing equivalence in compative bioavailability. *Communications in Statistics-Theory and Methods*, *16*(9), 2771–2778.

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.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

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.

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.

Kanetkar, V., Evans, M. G., Everell, S. A., Irvine, D., & Millman, Z. (1995). The effect of scale changes on meta-analysis of multiplicative and main effects models. *Educational and Psychological Measurement*, *55*(2), 206–224.

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., & 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., 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.

Leday, G. G., Engel, J., Vossen, J. H., de Vos, R. C., & van der Voet, H. (2022). Multivariate equivalence testing for food safety assessment. *Food and Chemical Toxicology*, *170*, 113446.

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.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*(2), 188–209.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. https://doi.org/10.1198/016214507000001337.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504.

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 148–153.

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.

Meyners, M. (2007). Least equivalent allowable differences in equivalence testing. *Food Quality and Preference*, *18*(3), 541–547.

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.

Morey, R. D., Rouder, J. N., Jamil, T., & Morey, M. R. D. (2015). Package 'BayesFactor'. *CRAN repository*. 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.

Müller-Coors, J. (1990). The power of the Anderson-Hauck test and the double-test. *Biometrical Journal*, *32*, 259–266.

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/

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). *One hundred years of social psychology quantitatively described.* Educational Publishing Foundation.

Robinson, A., & Robinson, M. A. (2016). *CRAN repository: Package 'equivalence'.* https://tinyurl.com/34brwssa.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553.

Romano, J. P. (2005). Optimal testing of equivalence hypotheses. *The Annals of Statistics*, *33*(3), 1036–1047.

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.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*(6), 657–680.

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.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*, *26*(1), 1–9.

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.

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

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.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594.

Zhang, P. (2003). A simple formula for sample size calculation in equivalence studies. *Journal of Biopharmaceutical Statistics*, *13*(3), 529–538.

Supplemental Material

Additional formulas, notation and tables

Least squares estimation

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

$$\widehat{\beta}_k = ((X^T X)^{-1} X^T y)_k, \text{for } k \text{ in 1,..., } K, \text{ and} \tag{22}$$

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (\hat{\epsilon}_i^2)/(N-K-1)}, \tag{23}$$

where $\hat{\epsilon}_i = \widehat{y}_i - y_i$, and $\widehat{y}_i = X_{i\times}^T \widehat{\beta}$, for i in 1,..., N. We also define R^2_{YX}, the coefficient of determination from the linear regression of Y predicted from X:

$$R^2_{YX} = \frac{\sigma_{XY}^T \Sigma_X^{-1} \sigma_{XY}}{\sigma_Y^2}, \tag{24}$$

where $\sigma_Y^2 = (\beta^T \text{Cov}(X)\beta + \sigma^2)$ is the unconditional variance of Y, (note that: $\sigma_Y^2 \geq \sigma^2$); σ_{XY} is the vector of population covariances between the K different predictors and Y; and Σ_X is the population covariance matrix of the K different predictors. The $\hat{R}^2_{YX}$ statistic estimates the parameter R^2_{YX} from the observed data:

$$\hat{R}^2_{YX} = 1 - \frac{\sum_{i=1}^{N} \hat{\epsilon}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{25}$$

where $\bar{y} = \sum_{i=1}^{N} y_i/N$.

A standard NHST for the k-th predictor, X_k, is stated as:

$\text{H}_0 : \beta_k = 0$, vs.

$\mathrm{H}_1 : \beta_k \neq 0$.

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:

$$p\text{-value}_k = 2 \times F_t \left(\frac{|\widehat{\beta}_k|}{\mathrm{SE}(\widehat{\beta}_k)}, N - K - 1 \right), \text{for } k \text{ in } 0,...,K, \tag{26}$$

where we use $F_t(\quad \cdot \quad ; df)$ to denote the cdf of the t-distribution with df degrees of freedom, and where: $\mathrm{SE}(\widehat{\beta}_k) = \hat{\sigma}\sqrt{[(X^T X)^{-1}]_{kk}}$. Alternatively, we can conduct an F-test and, for k in 1,...,K, we will obtain the very same p-value with:

$$p\text{-value}_k = p_F \left((N - K - 1) \frac{\widehat{sr}_k^2}{1 - \hat{R}_{YX}^2}, 1, N - K - 1 \right), \tag{27}$$

where $p_f(\cdot \quad ; df_1, df_2)$ is the cdf of the F-distribution with df_1 and df_2 degrees of freedom, and where: $\widehat{sr}_k^2 = \hat{R}_{YX}^2 - \hat{R}_{YX_{-k}}^2$. Regardless of whether the t-test or the F-test is employed, if $p\text{-value}_k < \alpha$, we reject the null hypothesis of $\mathrm{H}_0 : \beta_k = 0$ against the alternative $\mathrm{H}_0 : \beta_k \neq 0$.

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 Serlin et al. (1993), Weber & Popova (2012)):

$$\begin{aligned} &\mathrm{H}_0 : \theta \leq \Delta_{lower} \quad \text{or:} \quad \theta \geq \Delta_{upper}, \quad \text{vs.} \\ &\mathrm{H}_1 : \theta > \Delta_{lower} \quad \text{and:} \quad \theta < \Delta_{upper}, \end{aligned}$$

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:

$$p_d^{lower} = 1 - F_t \left(\frac{\hat{\mu}_d}{\hat{\sigma}_p} \sqrt{\frac{N_1 N_2}{N_1 + N_2}}, N_1 + N_2 - 2, \Delta_{lower} \sqrt{\frac{N_1 N_2}{N_1 + N_2}} \right), \quad \text{and} \tag{28}$$

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

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:

$$\hat{\sigma}_p = \sqrt{\frac{(N_1-1)\hat{\sigma}_1^2 + (N_2-1)\hat{\sigma}_2^2}{N_1+N_2-2}}, \tag{29}$$

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 $p_Z = max(\mathbb{p}_Z^{lower}, \mathbb{p}_Z^{upper})$, where:

$$\mathbb{p}_Z^{lower} = 1 - F_Z\left(\frac{\sqrt{N-3}}{2}\ln\left(\left(\frac{1+\widehat{sr}_1}{1-\widehat{sr}_1}\right) - \left(\frac{1+\Delta_{lower}}{1-\Delta_{lower}}\right)\right)\right), \tag{30}$$

and:

$$\mathbb{p}_Z^{upper} = 1 - F_Z\left(\frac{\sqrt{N-3}}{2}\ln\left(\left(\frac{1+\widehat{sr}_1}{1-\widehat{sr}_1}\right) + \left(\frac{1+\Delta_{upper}}{1-\Delta_{upper}}\right)\right)\right),$$

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

i	X_1 Received counselling (yes=1; no=0)	X_2 Age (years)	X_3 Household income ($)	Y Anxiety score points
1	0	13	75593	12.1
2	0	15	57954	15.5
3	0	13	61336	13.3
4	1	14	47628	14.3
5	0	14	46564	12.3
6	1	12	74071	13.8
7	1	17	76964	14.5
8	1	15	69060	11.6
9	0	13	86445	12.8
10	0	17	109002	17.5
11	1	16	58179	14.7
12	1	14	21817	16.8
13	1	17	88115	12.5
14	0	17	53816	15.8
15	1	16	54240	17.3
16	0	16	88511	15.8
17	1	16	62305	16.2
18	0	15	43586	13.6
19	0	14	71626	12.3
20	0	14	65222	12.0
21	0	14	68115	14.6
22	1	15	75706	13.2
23	0	13	60587	12.8
24	0	19	80888	16.1
25	0	17	63590	20.1
26	0	13	74636	12.3
27	1	14	89937	15.2
28	1	14	76704	15.0
29	0	16	61481	13.2
30	1	15	90976	15.0
31	0	15	87870	17.9
32	1	15	78968	16.3
33	0	15	72775	14.9
34	1	17	55442	15.6
35	1	15	95213	10.4
36	0	18	55995	19.0
37	0	12	111747	9.5
38	0	16	98652	16.7
39	0	15	63286	19.3
40	1	15	47472	12.3

Table 6. The hypothetical anxiety study dataset. 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 predictor corresponding to the student's age in years; and X_3 might be a continuous predictor corresponding to the student's household income in dollars.

Simulation Study 1

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 $H_0 : |\mu_d| \geq \Delta \times \hat{\sigma}$; and

(2) the valid test (see equation (28)), 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 (values representative of sample sizes in large and very large psychological studies (Kühberger et al., 2014; Fraley & Vazire, 2014; Marszalek et al., 2011)), 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 equal to 0, or equal to Δ, or equal to 0.15.

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)).

The simulation study was done using the R statistical software with default simulation routines (R Core Team, 2020). Results are displayed in Table 5 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-val $< 0.05 \mid d = \Delta$)		Pr(p-val $< 0.05 \mid d = 0$)		Pr(p-val $< 0.05 \mid d = 0.15$)	
		invalid test	valid test	invalid test	valid test	invalid test	valid test
54	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	0.000	0.000	0.000	0.000	0.000	0.000
80	0.10	0.000	0.000	0.000	0.000	0.000	0.000
80	0.15	0.000	0.000	0.000	0.000	0.000	0.000
80	0.20	0.016	0.022	0.081	0.107	0.076	0.101
180	0.10	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	0.038	0.041	0.273	0.286	0.237	0.249
180	0.20	0.046	0.049	0.694	0.707	0.629	0.642
540	0.10	0.048	0.048	0.499	0.503	0.345	0.348
540	0.15	0.048	0.050	0.935	0.937	0.817	0.820
540	0.20	0.047	0.049	0.998	0.998	0.982	0.983
1000	0.10	0.049	0.050	0.870	0.872	0.597	0.599
1000	0.15	0.048	0.050	0.998	0.998	0.968	0.969
1000	0.20	0.048	0.050	1.000	1.000	1.000	1.000
3500	0.10	0.050	0.050	1.000	1.000	0.972	0.972
3500	0.15	0.049	0.050	1.000	1.000	1.000	1.000
3500	0.20	0.049	0.050	1.000	1.000	1.000	1.000

Table 7. Results from Simulation Study 1. 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 2

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

(1) the proposed equivalence test for semipartial correlation coefficients (see equation (12)) ("sr test"); and

(2) the equivalence test for correlations based on Fisher's Z transformation (see equation (30)) ("Z test").

Both tests are valid for testing the lack of an association between Y and X when $K = 1$ (i.e., for simple linear regression). We considered 6 different values for the total sample size, N, ranging from 54 to 3500 (values representative of sample sizes in large and very large psychological studies (Kühberger et al., 2014; Fraley & Vazire, 2014; Marszalek et al., 2011)), and 3 different values for the upper bound of a symmetric equivalence margin, Δ, ranging from 0.1 to 0.20. We simulated data from a bivariate Normal distribution such that the true value of sr_1 was equal to 0, or equal to Δ, or

equal to 0.05.

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)).

The simulation study was done using the R statistical software with default simulation routines (R Core Team, 2020). Results are displayed in Table 5 and suggest that, in practice, both equivalence tests obtain very similar values for the type 1 error and statistical power.

N	Δ	$\Pr(p < 0.05 \mid sr_1 = \Delta)$		$\Pr(p < 0.05 \mid sr_1 = 0)$		$\Pr(p < 0.05 \mid sr_1 = 0.05)$	
		sr test	Z test	sr test	Z test	sr test	Z test
54	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	0.000	0.000	0.000	0.000	0.000	0.000
80	0.10	0.000	0.000	0.000	0.000	0.000	0.000
80	0.15	0.000	0.000	0.000	0.000	0.000	0.000
80	0.20	0.016	0.022	0.081	0.107	0.076	0.101
180	0.10	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	0.038	0.041	0.273	0.286	0.237	0.249
180	0.20	0.046	0.049	0.694	0.707	0.629	0.642
540	0.10	0.048	0.048	0.499	0.503	0.345	0.348
540	0.15	0.048	0.050	0.935	0.937	0.817	0.820
540	0.20	0.047	0.049	0.998	0.998	0.982	0.983
1000	0.10	0.049	0.050	0.870	0.872	0.597	0.599
1000	0.15	0.048	0.050	0.998	0.998	0.968	0.969
1000	0.20	0.048	0.050	1.000	1.000	1.000	1.000
0	0.00	0.000	0.000	0.000	0.000	0.000	0.000
3500	0.15	0.049	0.050	1.000	1.000	1.000	1.000
3500	0.20	0.049	0.050	1.000	1.000	1.000	1.000

Table 8. Results from Simulation Study 2. Note that the maximum type 1 error rate should not exceed $\alpha = 0.05$. When $\Delta > sr_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power. The two tests under consideration are the proposed equivalence test for semipartial correlation coefficients ("sr test") and the equivalence test for correlations based on Fisher's Z transformation ("Z test").

Simulation Study 3

We conducted a simple simulation study in order to better understand the operating characteristics of the proposed equivalence test for the semipartial correlation and to confirm that the proposed formula for approximating statistical power (equation (13)) is accurate. The equivalence test in the simulation study targeted sr_1 and considered a symmetric equivalence margin, $(-\Delta, \Delta)$, such that the hypothesis test in question can be stated as: $\mathrm{H}_0 : |sr_1| \geq \Delta$, vs. $\mathrm{H}_1 : |sr_1| < \Delta$.

We considered 4 different values for the total sample size, N, ranging from 54 to 3500 (values representative of sample sizes in large and very large psychological studies (Kühberger et al., 2014; Fraley & Vazire, 2014; Marszalek et al., 2011)), and 3 different values for the upper bound of the symmetric equivalence margin, Δ, ranging from 0.10 to 0.20. We considered two values for K, the number of predictors: $K = 2$ or $K = 4$; and simulated the predictors from a multivariate Normal distribution with a correlation matrix in which all off-diagonal elements were equal to either $\rho_X = 0.1$ or to $\rho_X = 0.2$. Finally, the outcome data, Y, was simulated such that the true value of sr_1 was equal to 0, or equal to Δ, or equal to 0.05.

For each of the different configurations within the simulation study, we simulated 500,000 unique datasets and calculated a p-value with the proposed equivalence test. We then calculated the proportion of these p-values less than $\alpha = 0.05$. We also used the proposed formula for approximating statistical power for each scenario to calculate the approximate power. We specifically chose to conduct 500,000 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 $0.0003 \approx \sqrt{0.05(1 - 0.05)/500,000}$; see Morris et al. (2019)).

The simulation study was done using the R statistical software with default simulation routines (R Core Team, 2020). Results are displayed in Table 5 and suggest that, in practice, the proposed test ("sr test") has correct type 1 error and that the proposed formula for estimating statistical power ("approx pwr.") is reasonably accurate.

N	Δ	K	ρ_X	$\Pr(p < 0.05 \mid sr_1 = \Delta)$		$\Pr(p < 0.05 \mid sr_1 = 0)$		$\Pr(p < 0.05 \mid sr_1 = 0.05)$	
				sr test	approx pwr.	sr test	approx pwr.	sr test	approx pwr.
54	0.10	2.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	2.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	2.0	0.10	0.004	0.000	0.000	0.000	0.000	0.000
180	0.10	2.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	2.0	0.10	0.044	0.046	0.303	0.301	0.225	0.229
180	0.20	2.0	0.10	0.049	0.050	0.718	0.713	0.601	0.598
540	0.10	2.0	0.10	0.049	0.049	0.528	0.527	0.284	0.285
540	0.15	2.0	0.10	0.049	0.050	0.945	0.942	0.756	0.754
540	0.20	2.0	0.10	0.049	0.050	0.998	0.998	0.970	0.968
3500	0.10	2.0	0.10	0.050	0.050	1.000	1.000	0.910	0.909
3500	0.15	2.0	0.10	0.050	0.050	1.000	1.000	1.000	1.000
3500	0.20	2.0	0.10	0.050	0.050	1.000	1.000	1.000	1.000
54	0.10	4.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	4.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	4.0	0.10	0.003	0.000	0.000	0.000	0.000	0.000
180	0.10	4.0	0.10	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	4.0	0.10	0.044	0.045	0.299	0.299	0.225	0.227
180	0.20	4.0	0.10	0.050	0.050	0.717	0.711	0.600	0.596
540	0.10	4.0	0.10	0.048	0.049	0.525	0.524	0.282	0.284
540	0.15	4.0	0.10	0.049	0.050	0.943	0.941	0.756	0.753
540	0.20	4.0	0.10	0.049	0.050	0.998	0.998	0.971	0.968
3500	0.10	4.0	0.10	0.049	0.050	1.000	1.000	0.909	0.909
3500	0.15	4.0	0.10	0.049	0.050	1.000	1.000	1.000	1.000
3500	0.20	4.0	0.10	0.049	0.050	1.000	1.000	1.000	1.000
54	0.10	2.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	2.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	2.0	0.25	0.007	0.010	0.000	0.000	0.000	0.000
180	0.10	2.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	2.0	0.25	0.045	0.047	0.303	0.301	0.227	0.231
180	0.20	2.0	0.25	0.049	0.050	0.718	0.713	0.602	0.600
540	0.10	2.0	0.25	0.049	0.049	0.528	0.527	0.285	0.286
540	0.15	2.0	0.25	0.049	0.050	0.945	0.942	0.757	0.756
540	0.20	2.0	0.25	0.049	0.050	0.998	0.998	0.971	0.969
3500	0.10	2.0	0.25	0.050	0.050	1.000	1.000	0.911	0.910
3500	0.15	2.0	0.25	0.050	0.050	1.000	1.000	1.000	1.000
3500	0.20	2.0	0.25	0.050	0.050	1.000	1.000	1.000	1.000
54	0.10	4.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
54	0.15	4.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
54	0.20	4.0	0.25	0.002	0.000	0.000	0.000	0.000	0.000
180	0.10	4.0	0.25	0.000	0.000	0.000	0.000	0.000	0.000
180	0.15	4.0	0.25	0.043	0.045	0.295	0.295	0.225	0.226
180	0.20	4.0	0.25	0.049	0.050	0.714	0.708	0.600	0.595
540	0.10	4.0	0.25	0.048	0.049	0.521	0.521	0.282	0.283
540	0.15	4.0	0.25	0.049	0.050	0.942	0.940	0.756	0.752
540	0.20	4.0	0.25	0.049	0.050	0.998	0.998	0.971	0.968
3500	0.10	4.0	0.25	0.049	0.050	1.000	1.000	0.909	0.908
3500	0.15	4.0	0.25	0.049	0.050	1.000	1.000	1.000	1.000
3500	0.20	4.0	0.25	0.049	0.050	1.000	1.000	1.000	1.000

Table 9. Results from Simulation Study 3. Note that the maximum type 1 error rate should not exceed $\alpha = 0.05$. When $\Delta > sr_1$, the probability of a p-value less than 0.05 corresponds to the test's statistical power.

R code

All data and code has been saved in csv format and is available in the OSF repository: DOI 10.17605/OSF.IO/5YR92.

```
##########################################

## Useful functions
```

```
##########################################
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)

  ##### Run a two t-test procedure for equivalance with Fisher's z transformation ####
  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)
}
##########################################

##########################################
equivBeta <- 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)}
  Xmatrix <- cbind(Xmatrix)
  X <- cbind(1, Xmatrix)
  N <- dim(cbind(X[,-1]))[1]
  K <- dim(cbind(X[,-1]))[2]
  if(length(DELTA_lower)==1){DELTA_lower <- rep(DELTA_lower, K+1)}
  if(length(DELTA_upper)==1){DELTA_upper <- rep(DELTA_upper, K+1)}

  lmmod <- summary(lm(Y~X[,-1]))
  beta_hat <- lmmod$coef[,1]
  SE_beta_hat <- lmmod$coef[,2]

  mysigma<-summary(lm(Y~X[,-1]))$sigma
```

```
  mysigma*sqrt(solve(t(X)%*%X)[k,k])

  pval <- p_lower <- p_upper <- rep(0,K)
  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)
    p_upper[k] <- pt((-beta_hat[k] + DELTA_upper[k])/SE_beta_hat[k], N-K-1, 0,
                     lower.tail=FALSE)
    pval[k] <- max(c(p_lower[k],p_upper[k]))
  }

  names(beta_hat) <- paste("beta", c(1:dim(X)[2])-1, sep="_")
  names(pval) <- paste("pval", c(1:dim(X)[2])-1, sep="_")
  DELTA = cbind(DELTA_lower, DELTA_upper)
  rownames(DELTA) <- paste("DELTA", c(1:dim(X)[2])-1, sep="_")
  return(list(beta = beta_hat, pval = pval, DELTA = DELTA))
}
##########################################

##########################################
equivBetaPower <- function(DELTA_upper, DELTA_lower, N, K, SEbetak, true_beta=0){
  ncp1 <- (DELTA_upper-true_beta)/SEbetak
  ncp2 <- (DELTA_lower-true_beta)/SEbetak
  Tstatstar <- qt(1-0.05, N-K-1)
  power = pt(+ncp1-Tstatstar, N-K-1, lower.tail=TRUE) - pt(+ncp2+Tstatstar,
              N-K-1, lower.tail=TRUE)
  return(power)
}
##########################################

##########################################
equivSR <- function(Y= rnorm(100),
         Xmatrix= cbind(rnorm(100),rnorm(100)),
         DELTA_upper= 0.1,
         DELTA_lower= NA){

  Xmatrix <-cbind(Xmatrix)
  X <- cbind(1,Xmatrix)
  N <- dim(Xmatrix)[1]
  K <- dim(Xmatrix)[2]
```

```
kvec= 1:K
if(is.na(DELTA_lower[1])){DELTA_lower <-(-DELTA_upper)}
if(length(DELTA_upper)!=K){DELTA_upper <- rep(DELTA_upper[1], K)}
if(length(DELTA_lower)!=K){DELTA_lower <- rep(DELTA_lower[1], K)}

lmmod <- summary(lm(Y~X[,-1]))
R2 <- lmmod$r.squared
if(K==1){R2Xkmink <- 0; diffR2k<-R2; R2Ymink<-0}
if(K>1){
  R2Ymink <- apply(cbind(kvec),1,function(k)summary(lm(Y~Xmatrix[,-k]))$r.squared)
  R2Xkmink <- apply(cbind(kvec),1,function(k)summary(lm(Xmatrix[,k]~Xmatrix[,-k]))$r.squared)
  diffR2k <- unlist(lapply(c(kvec), function(k) {R2-summary(lm(Y~Xmatrix[,-k]))$r.squared}))
}
lmmod_scale <- summary(lm(scale(Y)~scale(X[,-1])-1))
SPC <- lmmod_scale$coef[,1]*sqrt(1-R2Xkmink)
# should be equal in abs:
c(sqrt(diffR2k),SPC)

# equation (11) of Dudgeon (2016)
SIGMA2_SPC <- (R2^2 - 2*R2 + R2Ymink + 1 - R2Ymink^2)/(N-K-1)
SE_SPC <- sqrt(SIGMA2_SPC)

CI90_upper_squared <- CI95_upper <- CI95_lower <- CI90_upper <-
  CI90_lower <- pval <- pval1 <- pval2 <- rep(0, length(kvec))
for(k in kvec){
  pval1[k] <- pt((SPC[k]-DELTA_lower[k])/SE_SPC[k], N-K-1, lower.tail=FALSE)
  pval2[k] <- pt((DELTA_upper[k]-SPC[k])/SE_SPC[k], N-K-1, lower.tail=FALSE)
  CI90_upper[k] <- SPC[k] - qt(0.05,df=N-K-1)* SE_SPC[k]
  CI90_lower[k] <- SPC[k] + qt(0.05,df=N-K-1)* SE_SPC[k]

  CI95_upper[k] <- SPC[k] - qt(0.025,df=N-K-1)* SE_SPC[k]
  CI95_lower[k] <- SPC[k] + qt(0.025,df=N-K-1)* SE_SPC[k]

  CI90_upper_squared[k] <- (SPC[k] - qt(0.1,df=N-K-1)* SE_SPC[k])^2
  pval[k] <- max(c(pval1[k], pval2[k]))
}

CI90 <- cbind(CI90_lower, CI90_upper)
CI95 <- cbind(CI95_lower, CI95_upper)
LEAD <- apply(abs(CI90),1,max)
```

```
  names(SPC) <- paste("sr", 1:K, sep="_")
  names(SE_SPC) <- paste("SE_sr", 1:K, sep="_")
  names(pval) <- paste("pval", 1:K, sep="_")
  rownames(CI90) <- paste("CI90", 1:K, sep="_")
  rownames(CI95) <- paste("CI95", 1:K, sep="_")
  names(LEAD) <- paste("LEAD", 1:K, sep="_")
  return(list(sr = unname(SPC), SE_sr= SE_SPC , pval= pval,
              CI90=CI90, CI95=CI95, LEAD=LEAD))
}
##########################################

##########################################
equivSRPower <- function(DELTA_upper, DELTA_lower, N, K, SESRk, true_SR=0){
  ncp1 <- (DELTA_upper-true_SR)/SESRk
  ncp2 <- (DELTA_lower-true_SR)/SESRk
  Tstatstar <- qt(1-0.05, N-K-1)
  power = pt(+ncp1-Tstatstar, N-K-1, lower.tail=TRUE) - pt(+ncp2+Tstatstar,
            N-K-1, lower.tail=TRUE)
  return(power)
}
##########################################

##########################################
BFstandardBeta<-function(Y= yvec, Xmatrix= Xmat, BFthres=3, random=FALSE){

  K<-dim(Xmatrix)[2]
  mydata<-data.frame(Y, Xmatrix)
  colnames(mydata)<- c(c("yvector"),paste("X",1:K,sep=""))
  head(mydata)
  BFmod <- regressionBF(yvector~. , data= mydata)

  BF<-result<-rep(0,K)
  for(k in 1:K){

    whichk<-paste("X",k,sep="")
    BF_without_k <-BFmod[!grepl(whichk,names(BFmod)$numerator)][
      which.max(nchar(names(BFmod)$numerator[!grepl(whichk,names(BFmod)$numerator)]))]
    BF_full <- BFmod[which.max(nchar(names(BFmod)$numerator))]
```

```
    BF[k] <- exp(as.numeric(slot(BF_without_k,
                "bayesFactor")[1]))/exp(as.numeric(slot(BF_full,"bayesFactor")[1]))

    if(BF[k]<= 1/BFthres){result[k]<-"positive" }
    if(BF[k]>BFthres){result[k]<-"negative"}
    if(BF[k]> 1/BFthres & BF[k]<BFthres){result[k]<-"inconclusive"}
  }
  return(list(BF=c(BF), BFthres=c(BFthres),conclusion= result))
}
##########################################

##########################################
## Hypothetical anxiety study example:
##########################################
equivBetaPower(DELTA_upper=2, DELTA_lower=-2, N=40, K=3, SEbetak=0.63)
#0.8540956

# Note: hypothetical data was created with the following code:
#set.seed(123)
#anx <- cbind(1,sample(c(0,1),40,TRUE),round(rnorm(40,15,1.8)), round(rnorm(40,68000,20000)))
#score <- round(anx%*%c(8,0.6,0.5,-0.00001) + rnorm(40,0,2.3),1)

anxiety_data <- read.csv("anxiety.csv")[,-1]
score <- anxiety_data[,1]
anx <- as.matrix(anxiety_data[,-1])

# Study results:
coefficients(summary(lm(score~anx-1)))[,1:2]
confint((lm(score~anx-1)), level=0.95)
confint((lm(score~anx-1)), level=0.90)

# Equivalence test for regression coef:
equivBeta(Y = score, Xmatrix = anx[,-1], DELTA_lower = -2, DELTA_upper = 2)

# or "by hand":
beta_hat <- coefficients(summary(lm(score~anx-1)))[2,1]
SE_beta_hat <- coefficients(summary(lm(score~anx-1)))[2,2]
DELTA_lower <- -2
DELTA_upper <- 2
```

```
N <- 40
K <- 3
p_lower <- pt((beta_hat - DELTA_lower)/SE_beta_hat, N-K-1, 0, lower.tail=FALSE)
p_upper <- pt((-beta_hat + DELTA_upper)/SE_beta_hat, N-K-1, 0, lower.tail=FALSE)
pval <- max(c(p_lower,p_upper))
pval
# 0.017733

# LEAD (or "equivalence confidence interval"):
# 90% CI:
CI90 <- c(beta_hat-qt(1-0.05,N-K-1)*SE_beta_hat, beta_hat+qt(1-0.05,N-K-1)*SE_beta_hat)
LEAD <- max(abs(CI90))
LEAD
#  1.674596
equivBeta(Y = score, Xmatrix = anx[,-1], DELTA_lower = -LEAD, DELTA_upper = LEAD)$pval[2]
#  0.05

##########################################
## Salaries example
##########################################
Salaries <- read.csv("Salaries.csv")[,-1]
# or obtain data from:
# library(carData)

### simple linear regression:
y <- Salaries$salary
X <- model.matrix(lm(salary ~ sex, data=Salaries))

summary(lm(salary ~ sex, data=Salaries))$coef[1:2,1:2]
# Estimate Std. Error
# (Intercept) 101002.41   4809.386
# sexMale      14088.01   5064.579
summary(lm(salary ~ sex, data=Salaries))$sigma
# 30034.61
salaries_equiv <- equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.5, DELTA_lower = -0.5)
salaries_equiv["sr"]
# 0.1386102
salaries_equiv["SE_sr"]
# 0.04934876
equivBeta(Y = y, Xmatrix = X[,-1], DELTA_upper = 5000,  DELTA_lower = -5000)$pval[2]
```

```
# 0.9632451

# the same equivalence test done "by hand":
beta_hat <- coefficients(summary(lm(salary~sex,data=Salaries)))[2,1]
SE_beta_hat <- coefficients(summary(lm(salary~sex,data=Salaries)))[2,2]
DELTA_lower <- -5000
DELTA_upper <- 5000
N <- 397
K <- 1
p_lower <- pt((beta_hat - DELTA_lower)/SE_beta_hat, N-K-1, 0, lower.tail=FALSE)
p_upper <- pt((-beta_hat + DELTA_upper)/SE_beta_hat, N-K-1, 0, lower.tail=FALSE)
pval <- max(c(p_lower,p_upper))
pval

equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.7827741
# the same equivalence test done "by hand":
p_lower <- pt((sr1- (-0.10))/SEsr1, 397-1-1, lower.tail=FALSE)
p_upper <- pt((0.10-sr1)/SEsr1,397-1-1, lower.tail=FALSE)
pval <- max(c(p_lower,p_upper))
pval

library(BayesFactor)
sdata <- data.frame(salary = Salaries$salary,
            sex = as.numeric(as.factor(Salaries$sex)) - 1)
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

### multiple linear regression:
y <- Salaries$salary
X <- model.matrix(lm(salary ~ sex + yrs.since.phd  + yrs.service + discipline + rank,
                     data=Salaries))
```

```
# NHST p-values
mod1 <- summary(lm(salary ~ sex + yrs.since.phd  + yrs.service + discipline +
           rank, data=Salaries))
mod1$coef[-1,4]
# 2.158412e-01  2.697855e-02  2.142543e-02  1.878412e-09  1.983251e-03  2.296130e-23
mod1$coef[,1:2]

SPCobj <- equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)
round(SPCobj$pval,4)
# 0.0761 0.3249 0.3577 0.9997 0.6699 1.0000
round(SPCobj$LEAD,4)
# 0.1080 0.1446 0.1480 0.2913 0.1780 0.3999
mod1$sigma
# 22538.65
mod1$r.squared
# 0.4546766

# Bayes Factors
BFs <- (BFstandardBeta(Y= y, Xmatrix=X[,-1])$BF)
BFs
# 3.860594e+00 7.352492e-01 6.036516e-01 1.540123e-07 7.331954e-02 5.592721e-21
round(cbind(mod1$coef[-1,4],SPCobj$pval,BFs,1/BFs),3)
#                              BFs
#sexMale       0.216 0.076 3.861 2.590000e-01
#yrs.since.phd 0.027 0.325 0.735 1.360000e+00
#yrs.service   0.021 0.358 0.604 1.657000e+00
#disciplineB   0.000 1.000 0.000 6.492987e+06
#rankAsstProf  0.002 0.670 0.073 1.363900e+01
#rankProf      0.000 1.000 0.000 1.138788e+15
##########################################
## Mindset Theory example
##########################################
mind <- read.csv("Mindset.csv")
mind[,"cog"]<-rowMeans(cbind(scale(mind[,"Cattell.Score"]), scale(mind[,"Letter.Sets.Score"])))

# 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"], 0.2, -Inf)
S <- equivSR(mind[,"X1.Learning.Goal"],
        mind[,"Mindset.Score"], DELTA_upper=0.2, DELTA_lower=-Inf)
```

```
res1 <- c((S$sr)[1], Z, S$pval)
res1
# 0.09774130     0.01450918     0.01582211

# 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)
S <- equivSR(mind[,"X2.Performance.Goal"],
      mind[,"Mindset.Score"], DELTA_upper=Inf, DELTA_lower=-0.20)
res2 <- c((S$sr)[1], Z, S$pval)
res2
#  -0.10894803     0.02576879     0.02749891

# 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)
S <- equivSR(mind[,"X3.Performance.Avoidance.Goal"],
      mind[,"Mindset.Score"], DELTA_upper=Inf, DELTA_lower=-0.20)
res3 <- c((S$sr)[1], Z, S$pval)
res3
#  -0.0391385402   0.0003229064   0.0004179885

# Testing Premise 4: people with fixed mind-sets believe
# that talent alone without effortcreates 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)
S <- equivSR(mind[,"X4.Belief.in.Talent"],
       mind[,"Mindset.Score"], DELTA_upper=Inf, DELTA_lower=-0.20)
res4 <- c((S$sr)[1], Z, S$pval)
res4
#  -0.061215284    0.001588981    0.001906768

# 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"], 0.2, -Inf)
S <- equivSR(mind[,"X5.Response.To.Challenge"],
      mind[,"Mindset.Score"], DELTA_upper=0.20, DELTA_lower=-Inf)
res5 <- c((S$sr)[1], Z, S$pval)
res5
```

```
#  0.055873780    0.001100171    0.001343043

# 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"], 0.2, -Inf)
S <- equivSR(mind[,"X6.Raven.Test.Score"],
      mind[,"Mindset.Score"], DELTA_upper=0.20, DELTA_lower=-Inf)
res6 <- c((S$sr)[1], Z, S$pval)
res6
#  -1.217740e-01   5.976098e-12   1.525199e-11

# Testing Premise 6a: people with growth mind-sets are
# more resilient following failure when controlling for cognitive ability
summary(lm(X6.Raven.Test.Score~Mindset.Score+ cog, data=mind))
Z1 <- NA
S <- equivSR(mind[,"X6.Raven.Test.Score"],
       as.matrix(mind[,c("Mindset.Score", "cog")]), DELTA_upper=0.20, DELTA_lower=-Inf)
res6a <- c(S$sr[1], Z1, S$pval[1])
res6a
# -5.492464e-02            NA  1.109186e-09

mindset_results <- round(rbind(res1, res2, res3, res4, res5, res6, res6a), digits=3)
mindset_results
#              pvalue_equiv_z pval_1
# res1   0.098          0.015  0.016
# res2  -0.109          0.026  0.027
# res3  -0.039          0.000  0.000
# res4  -0.061          0.002  0.002
# res5   0.056          0.001  0.001
# res6  -0.122          0.000  0.000
# res6a -0.055             NA  0.000

##########################################
## Hominid brain evolution example
##########################################
### simple linear regression:
bailey <- read.csv("bailey2009.csv")

y <- bailey$cranium_capacity
X <- model.matrix(lm(cranium_capacity ~ temp_variation + isosd  + parasites +
```

```
                        population_dens, data=bailey))

summary(lm(cranium_capacity ~ temp_variation + isosd  + parasites +
     population_dens, data=bailey))$sigma

summary(lm(cranium_capacity ~ temp_variation + isosd  + parasites +
             population_dens, data=bailey))$r.squared

round(equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$sr,3)
round(equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$SE_sr,3)
# NHST p-values
summary(lm(y ~  temp_variation + isosd  + parasites + population_dens,
           data=bailey))$coef[-1,4]
# 9.273608e-01    1.062033e-10    1.436866e-01    1.287237e-16
# EQUIV p-values
equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$pval
# 0.01039188 0.99998331 0.17085167 1.00000000

equivSR(Y = y, Xmatrix = X[,-1], DELTA_upper = 0.1, DELTA_lower = -0.1)$LEAD
#0.07196695 0.35533726 0.12895891 0.45207451

# Bayes Factors
BFs <- (BFstandardBeta(Y= y, Xmatrix=X[,-1])$BF)
BFs
# 1.104066e+01 1.122479e-08 3.827830e+00 1.979351e-14
```