DETERMINING NEGLIGIBLE ASSOCIATIONS IN REGRESSION

UDI ALTER*
York University
udialter@yorku.ca
https://orcid.org/0000-0003-3133-839X

ALYSSA COUNSELL
Toronto Metropolitan University
a.counsell@torontomu.ca
https://orcid.org/0000-0001-9449-6630

* Corresponding author

Abstract

Psychological research is rife with inappropriately concluding "no effect" between predictors and outcome in regression models following statistically nonsignificant results. However, this approach is methodologically flawed because failing to reject the null hypothesis using traditional, difference-based tests does not mean the null is true. Using this approach leads to high rates of incorrect conclusions that flood psychological literature. This paper introduces a novel, methodologically sound alternative. In this paper, we demonstrate how an equivalence testing approach can be applied to multiple regression (which we refer to here as "negligible effect testing") to evaluate whether a predictor (measured in standardized or unstandardized units) has a negligible association with the outcome. In the first part of the paper, we evaluate the performance of two equivalence-based techniques and compared them to the traditional, difference-based test via a Monte Carlo simulation study. In the second part of the paper, we use examples from the literature to illustrate how researchers can implement the recommended negligible effect testing methods in their own work using open-access and user-friendly tools (`negligible` R package and Shiny app). Finally, we discuss how to report and interpret results from negligible effect testing and provide practical recommendations for best research practices based on the simulation results. All materials, including R code, results, and additional resources, are publicly available on the Open Science Framework (OSF): https://osf.io/w96xe/.

Determining Negligible Associations in Regression

Psychologists are often interested to determine if an individual predictor in a multiple regression model is negligible or practically insignificant. For example, Proudfoot and Kay (2018) tested whether feelings of personal control moderate the effect of perceived organizational stability on participants' tendency to identify with their organization, with higher stability associated with greater organization identification only for those in the "control threat" condition. To support the hypothesis, they sought to demonstrate that *no such relationship*, or effect, existed for participants feeling a lack of control. In another study, Seli et al. (2017) investigated the relationship between obsessive-compulsive disorder (OCD) symptomatology and mind wandering. Because intrusive thoughts are a shared symptom of both spontaneous mind wandering and OCD, the authors reasoned that spontaneous, *but not* deliberate, mind wandering would be positively associated with OCD symptomatology.

The problem is that researchers employ the same tools to test for a negligible association as they do when testing for a meaningful one. A common, but methodologically inappropriate, practice in the literature is to draw inferences of "no relationship" between the independent variable and the dependent variable following a statistically nonsignificant result (e.g., $p \geq \square$) from null hypothesis significance tests (NHST). That is, if a particular test statistic results in a sufficiently large p value (e.g., $p \geq 0.05$), researchers conclude "no association" and accept the null hypothesis (e.g., H_0: $\beta = 0$). But, p values are not - and, cannot be - an indication of the accuracy or probability of a hypothesis (Lakens, 2022). In the NHST framework, the null hypothesis is already assumed to be true, and p values are only indicative of the probability of observing the data we obtained. As Cohen (1994) explains, p values represent the probability of

obtaining data as or more extreme than that observed given the null hypothesis - P(data | H_0) - which is entirely different from P(H_0 | data), the probability of the null hypothesis, given the data.

What can one say then when encountered with a nonsignificant *p* value? If $p \geq \square$, one can merely conclude that there is insufficient evidence to reject the null. And, as the old expression goes: "absence of evidence is not evidence of absence." Of course, it might be that the null hypothesis is indeed true, but we simply cannot infer this from a nonsignificant *p* value: The statistical tests commonly used in psychology are difference-based tests which are designed to detect the *presence of an effect* – a difference or a relationship – not a lack thereof (Goertzen & Cribbie, 2010). Consequently, any conclusion about the accuracy of the null hypothesis is inappropriate, "no matter how large the *p* value" (Quintana, 2018).

Another issue with concluding "no association" following statistically nonsignificant results is that the probability of finding statistical significance increases as the sample size increases (unless the true effect is *exactly* zero, but see next section). Hence, the likelihood of finding nonsignificant results – which are commonly, but incorrectly, interpreted as a "negligible association" – decreases with larger sample sizes (Goertzen & Cribbie, 2010). Put differently, even with truly negligible effects (that are not perfectly zero), the probability of concluding "negligible association" is highest with small sample sizes and lowest with large sample sizes. This inverse relationship between sample size and statistical power to support the researcher's hypothesis of a negligible relationship is counterintuitive, misleading, and therefore outright inappropriate.

Defining Negligible Association

Realistically, the probability that the true effect is *exactly zero* (i.e., the null hypothesis) is infinitely small (Berkson, 1938; Cohen, 1990; Cohen, 1994; Thompson, 1992; Tukey, 1991). For

most purposes, however, associations or effects need only be small enough to be regarded as *practically zero*. Using an *equivalence testing* approach, "small enough" is defined by a prespecified value (indicated by δ) which represents the threshold of practical interest used to create an equivalence (or negligible effect) interval (-δ, δ). Here, an effect that falls within the equivalence interval's bounds is considered negligible, or *practically zero*. Note that δ may also be called the smallest effect size of interest (SESOI) or minimally meaningful effect size (MMES; we use δ and SESOI interchangeably thereafter). It is important to emphasize here that equivalence interval bounds, or the SESOI value, should be carefully planned a priori with concrete justification and independently from the data. In this paper, we only briefly discuss selecting a SESOI with examples, but see Anvari and Lakens (2021) and Lakens et al. (2018) for how to justify SESOI decisions.

To illustrate the equivalence interval conceptually, let us consider an example from the literature borrowed from Quintana (2018): Kupats and colleagues (2018) examined the lack of relationship between symptoms of generalized anxiety and cardiovascular autonomic dysfunction, which is measured by heart-rate variability (HRV). According to Quintana (2016), about 75% of HRV effect sizes in anxiety studies are at $d = 0.26$ or above. Quintana (2018) suggested that this value ($d = 0.26$) should therefore be set as δ, or the SESOI, in this example. Thus, the equivalence interval lower and upper bounds will be $d = -0.26$ and $d = 0.26$, respectively. To conclude a negligible association between generalized anxiety symptoms and HRV, the magnitude of the resulting association needs neither be larger than $d = 0.26$ nor smaller than $d = -0.26$; if the observed relationship's effect size and its associated uncertainty are contained within the equivalence interval ($-0.26 < d < 0.26$), a negligible effect should be concluded.

Testing for Negligible Association Using Equivalence Tests

Early appearances of testing for a negligible effect in psychological research used equivalence testing to determine similarity (i.e., negligible difference) between two group means (see Rogers et al., 1993). Since then, several other negligible effect testing methods have been developed. For example, Goertzen and Cribbie (2010) demonstrated how tests of equivalence can be used to determine negligible effects in simple correlations. Beribisky et al. (2020) showed how to test whether an indirect effect is negligible for a substantial mediation. Yuan and Chan (2016) and Counsell et al. (2020) proposed to use equivalence testing to assess negligible effects for measurement invariance. Campbell and Lakens (2021) illustrated how to test whether an ANOVA or a linear regression model, as a whole, does not account for a meaningful proportion of the outcome variable. Most recently, in a preprint, Campbell (2022) demonstrated how equivalence testing can be used on regression coefficients to test for a lack of meaningful association.

The work in Campbell (2022) represents a significant contribution to the equivalence testing literature but includes a few areas which we seek to address and supplement in this paper. First, the focus in the preprint (Campbell, 2022) is on *standardized* regression coefficients. In this paper, we shift the focus to *unstandardized* regression coefficients: Unstandardized regression coefficients are the most prevalent effect size reported in psychological research (Farmus et al., 2022) and are often the default output from statistical software packages. Similarly, the standard error provided by software output and reported in published articles is almost always tied to the *unstandardized* effect. Furthermore, we agree with Pek and Flora (2018) that reporting and using unstandardized effects is typically preferred because they are often tied to a more meaningful metric (e.g., reaction time, number of correct responses, etc.), are

easier to interpret, and more directly aid researchers to answer questions about practical implications. We acknowledge, however, that many unstandardized effect sizes from psychological scales are arbitrary such that the magnitude of effects will change depending on choices such as reporting a mean or a total score. Nevertheless, we believe that mean or sum scores from Likert-type items can oftentimes be more intuitive than thinking about standard deviation units, and even more so with well-known and established scales such as the Beck Depression Inventory (BDI-II; Beck et al., 1996) or the Minnesota Multiphasic Personality Inventory (MMPI-II; Butcher et al., 2001).

Second, most papers evaluating the equivalence testing approach in psychology, use Schuirmann's (1987) Two One-Sided Tests (TOST) technique. This paper presents and compares an additional, but less familiar, equivalence testing technique - Anderson and Hauck's (1983) procedure - which was previously found to demonstrate greater statistical power than the TOST at smaller sample sizes (e.g., Counsell & Cribbie, 2015). The current paper further introduces functions from the `negligible` R package (Cribbie et al., 2022) followed by an online Shiny application to assess negligible associations between a given predictor and outcome variables - measured in either standardized or unstandardized units - in a linear regression model.

While there are several recommended equivalence testing R packages available such as `TOSTER` (Caldwell, 2022; Lakens, 2017), the `negligible` package contributes to the arsenal of online tools by introducing equivalence-based adaptations of numerous statistical methods such as multiple regression, mediation analyses, structural equation modelling and fit indices, correlations, interaction between continuous variables, association between categorical variables, etc. The `negligible` package also provides paper-ready graphical output and helpful guidelines for users on how to interpret test results. Finally, we demonstrate, using examples

from the literature, how to determine if certain predictors are in fact negligible, and provide practical recommendations for researchers.

Note that the term “equivalence testing” is often referred to as ”non-inferiority testing” or “negligible effect testing” (equivalence/similarity can be thought of as a negligible difference). By the same token, the term “equivalence interval” is commonly called “negligible effect interval.” In this paper, we will use “negligible effect” terminology when referring to determining negligible associations, whereas “equivalence” terminology applies more broadly to a variety of methods, many of which are discussed in the next section.

Applying Equivalence Testing to Conclude Negligible Association in Multiple Regression

The most popular equivalence testing methods are Schuirmann’s TOST (1987) and the Anderson and Hauck procedure (AH; 1983; Hauck & Anderson, 1984). Although the two methods have the same purpose and hypotheses, each adheres to a different set of mathematical procedures. Furthermore, some research has suggested that the TOST is slightly more conservative than the AH method - with lower power and Type I error rates (Berger & Hsu, 1996; Brown et al., 1997). Statistical power and Type I error in the context of equivalence testing will be explained using a more convenient terminology of correct and incorrect negligible association conclusions which is defined in the Method section. In this section, we describe in detail how the TOST and AH procedures can be used to determine if an effect size of an unstandardized or standardized regression coefficient can be considered both practically and statistically negligible.

Schuirmann’s Two One-Sided Tests (TOST)

The popular TOST method was originally developed to evaluate the equivalence of two group means (Schuirmann, 1987). Applying the TOST to individual predictors in multiple

regression requires replacing the difference between group means with the regression slope of interest (β), which could be in unstandardized or standardized units. As its name suggests, the TOST consists of two directional *t* tests, each of which has a unique null hypothesis. The first *t* test's null hypothesis states that the magnitude of the effect (i.e., regression coefficient) is *equal to, or less than, the lower bound* of the SESOI, whereas the second *t* test's null hypothesis states that the magnitude of the same effect is *equal to, or greater than, the upper bound* of the *SESOI*. Thus, the null hypotheses are

H_{01}: $\beta \leq -\delta$

H_{02}: $\beta \geq \delta$

Where, again, $-\delta$ and δ are the lower and upper SESOI bounds, respectively. Because β could be in standardized or unstandardized units, it is crucial that δ is on the same metric as β. The alternative hypotheses then follow

H_{11}: $\beta > -\delta$

H_{12}: $\beta < \delta$

If the first null hypothesis (H_{01}) is rejected, there is evidence that our regression coefficient is *greater than the lower SESOI bound* (i.e., H_{11}). By the same token, if the second null hypothesis is rejected (H_{02}), there is evidence that the regression coefficient is *smaller than the upper SESOI bound* (i.e., H_{12}). It then follows that if *both* null hypotheses are rejected at the nominal Type I error rate, we can conclude that the regression coefficient is *simultaneously* greater than the lower SESOI bound *and* smaller than the upper SESOI bound; the regression coefficient is completely contained within the negligible effect interval, such that $-\delta < \beta < \delta$.

Both sets of hypotheses can be tested using two traditional one-tailed Student's *t* tests, where the first set is tested with its corresponding *t* statistic:

$$t_1 = \frac{\hat{\beta} - (-\delta)}{se_{\hat{\beta}}} = \frac{\hat{\beta} + \delta}{se_{\hat{\beta}}} \quad (1)$$

and the second set is tested with a similar formula, appropriately adjusting the numerator:

$$t_2 = \frac{(+\delta) - \hat{\beta}}{se_{\hat{\beta}}} = \frac{\delta - \hat{\beta}}{se_{\hat{\beta}}} \quad (2)$$

where $\hat{\beta}$ is the estimated effect size of the predictor of interest, β, and $se_{\hat{\beta}}$is the standard error of the corresponding regression coefficient. H_{01} is rejected if $t_1 \geq t_{c(1-,df)}$ and H_{02} is rejected if $t_2 \geq t_{c(1-,df)}$, where $t_{c(1-,df)}$ is the critical *t* value associated with the prespecified and the corresponding degrees of freedom, $df = n-k-1$, where *n* is the sample size and *k* is the number of predictors in the regression model. If both null hypotheses are rejected, the predictor of interest can be considered practically and statistically negligible.

Symmetric Confidence Intervals (CIs) Approach. Analogous to the TOST, researchers can use CIs to test for negligible association (Westlake, 1972; 1976; Metzler, 1974). Much like in difference-based tests, a CI can be constructed around the parameter estimate of interest (e.g., $\hat{\beta}$) with a predefined level of confidence (e.g., 95%). If the resulting CI falls entirely within the SESOI bounds (i.e., -δ, δ), a researcher may conclude a negligible effect (Dunnet & Gent, 1977). CIs for equivalence testing have one notable difference from their difference-based counterparts, however; they should be constructed at the 100·(1 - 2α)% confidence level rather than 100·(1 - α)%. Although the overall Type I error rate remains □, if the 100·(1 - 2□)% CI associated with the observed effect is contained within the SESOI bounds, a negligible association can be concluded. To explain why we use 100·(1 - 2□)% (e.g., 90%) instead of 100·(1 - α)% (e.g., 95%) CI to reject the null at □, let us consider Seaman and Serlin's (1998) point: Because the two null hypotheses are mutually exclusive, each one-sided test is constructed at the nominal Type I error rate α. Each one-sided t test "occupies" one tail of the central *t* distribution. The

resulting *t* value of one test must fall anywhere above the left-hand (lower) critical *t* value (associated with α and the corresponding degrees of freedom), whereas the other test must simultaneously fall anywhere below the right-hand (upper) critical *t* value (with the same α and degrees of freedom) to reject both null hypotheses and conclude a negligible effect. The intersecting area of the corresponding *t* distribution for the two rejection regions is therefore 1 - 2α. For a detailed review, see Metzler (1974), Rogers et al. (1993), Seaman and Serlin (1998), or Westlake (1972, 1976, 1981).

Anderson and Hauck (AH)

Anderson and Hauck (1983; Hauck & Anderson, 1984) proposed an additional approach to testing the equivalence of two group means on a parameter of interest where the difference between the two groups is contrasted against the middle of the equivalence interval. To determine if a specific predictor's effect size is negligible, a researcher must compare the regression coefficient associated with the target predictor once again with the interval bounds. Thus, the regression-adjusted hypotheses are as follows:

H_0: $\beta \leq -\delta$ or $\beta \geq \delta$, equivalently $|\beta| \geq \delta$

H_1: $-\delta < \beta < \delta$, equivalently $|\beta| < \delta$

The accompanying *AH T* statistic measures how far the observed effect size - here, of the regression coefficient - is from the center of the equivalence interval. Thus, the alternative hypothesis is supported if $|T|$ is sufficiently small. The adjusted *T* statistic is:

$$T = \frac{\hat{\beta} - \frac{1}{2}(-\delta + \delta)}{se_{\hat{\beta}}} \quad (3)$$

If the lower and upper bounds of the equivalence interval have the same absolute value (which is often the case), the center of the equivalence interval equals zero. Then Equation 3 can be simplified to the following:

$$T = \frac{\hat{\beta} - \frac{1}{2}(-\delta + \delta)}{se_{\hat{\beta}}} = \frac{\hat{\beta}}{se_{\hat{\beta}}} \quad (4)$$

As indicated by Hauck and Anderson (1984), the *p* value can be calculated as:

$$p = t¿ \quad (5)$$

Where *t* is the distribution function for Student's *t* with *df* degrees of freedom which are calculated the same as shown under Equation 2. If the resulting *p* value is smaller than , such that $p <$, the null hypothesis is rejected. In this case, the alternative hypothesis that the regression coefficient of interest falls within the equivalence bounds is supported and thus a negligible effect can be concluded.

Current Study

The purpose of the current study is five-fold. We aim to: (1) Demonstrate, using computer simulations, that the use of the traditional NHST (i.e., difference-based) regression methods is both inappropriate and inaccurate when the goal is to determine negligible or no association. (2) Offer an appropriate alternative to concluding negligible association between a given predictor variable and outcome by introducing equivalence testing approaches in multiple regression. (3) Evaluate and compare the statistical performance of two equivalence-based tests (AH and TOST) across different conditions. (4) Illustrate how researchers can implement appropriate negligible effect testing techniques in their own work and provide practical recommendations. (5) And finally, demonstrate how researchers can employ functions from the `negligible` R package and accompanying Shiny application to determine negligible associations in multiple regression.

Method

Simulation Study

Objective

We constructed a Monte Carlo study to compare the two equivalence-based tests, TOST and AH, to one another as well as to the traditional difference-based test for detecting a negligible association between individual predictors and the outcome variable in multiple regression under common conditions encountered by psychology researchers.

Design

The study design is a 3 (test type) x 6 (sample size) x 5 (effect size) x 4 (correlations/covariance between predictors), resulting in 360 total unique conditions. The nominal significance level (α) was set at .05 for each analysis to mimic the common practice in the literature and the SESOI – or negligible effect threshold value – was set at $\delta = .15$ (measured in the same units as the predictors) – a value slightly less conservative than in similar studies (i.e., Counsell & Cribbie, 2015; Cribbie et al., 2004). The simulation parameters and values are summarized in Table 1.

Table 1

Summary of Simulation Parameters and Parameter Values

Parameter	Values
Testing approach	Equivalence-based: AH, TOST Difference-based: traditional predictor-level *t* test
n	50, 75, 100, 250, 500, 1000
β	0, .05, .1, .15, .2
ρ	0, 0.25, 0.5, 0.75
δ	.15

α	.05

Note. The manipulated variables in the simulation are testing approach (difference-based and equivalence tests), sample size (*n*), correlation/covariance between the predictor variables (ρ), and effect size of individual predictors measured in standardized regression coefficients (β). Constant parameters across all conditions include the smallest effect size of interest (SESOI; δ) and nominal significance level (Type I error rate; α). Equivalence testing procedures in the simulation are the Anderson-Hauck (AH; Anderson & Hauck, 1983) and the two one-sided tests (TOST; Schuirmann, 1987).

Procedure

We simulated a population-level multivariate normal dataset ($\mu = 0$, $\sigma = 1$) using the `SimDesign` package (Chalmers & Adkins, 2020) in R (R Core Team, 2021). The population-level data consisted of six parameters, one intercept and five slope coefficients to estimate the unique relationships between five predictors and one outcome variable. Because our simulation was not tied to any particular research context or effect (e.g., depression, reaction time, anxiety, etc.), we decided to measure the relationship between predictors and outcome in standardized units (i.e., β) for convenience and uniformity, however, these can be replaced with unstandardized coefficients for identical simulation results. Note, however, that, in practice, conversion from unstandardized to standardized or vice versa may produce a minor difference in Type I error rate (see Supplemental Materials for a brief commentary and Campbell, 2022, for a longer discussion).

The population-level intercept parameter was one while the population-level slope parameters were set at $\beta = 0$, .05, .1, .15, and .2. We further manipulated the strength of the relationship between the model predictors by specifying the correlation matrices according to which the predictor variables were simulated. Specifically, the predictor variables were correlated at 0, 0.25, 0.5, and 0.75 to represent a wide array of scenarios encountered in the field. For example, in the first condition, each pair of predictor variables have a correlation of 0 in the population. Parameters were estimated by creating a multiple regression model with five predictors and *n* random observations sampled from the population-level data. This estimation

was repeated 5000 times, each time with another random sample of the same size (*n*) and a different association magnitude between predictors. The relatively high number of levels is intended to uncover reliable trends in the data. In each regression model (from the 5000 repetitions × six sample size levels × four association magnitudes between predictors = 120000 models) there were five predictors, each of which is tested with both difference-based and equivalence tests. A graphical illustration of the simulation procedure can be found in Figure S1 in the Supplementary Materials along with the simulation code.

Evaluating the Performance of Statistical Tests

Test performance is typically evaluated through Type I error and power rates. However, the definitions of power and error rates change from difference-based to equivalence-based tests because the two methods have opposite null hypotheses. Therefore, this terminology cannot consistently be applied. Consequently, we use the language "correct" and "incorrect" conclusions of negligible association instead. Accordingly, we evaluated test performance by comparing the number of correct versus incorrect conclusions of negligible association divided by the number of iterations (i.e., 5000).

Correct and Incorrect Conclusions

The difference between correct and incorrect conclusions lies in the true effect of the individual predictor of interest, which refers to the population-level relationship between the predictor variable and the outcome variable. Of course, the population-level, or true, relationship is rarely (if ever) known. Thus, regardless of test type, correctly concluding negligible association occurs if results indicate "negligible association" between a predictor and outcome when the true (i.e., population) effect is *within the equivalence interval* (i.e., $-\delta < \beta < \delta$). In the current study, three effect size levels lie within the equivalence interval (true negligible

association): β_1, β_2, and β_3 represent the levels inside the equivalence interval ($-.15 < \beta < .15$). Therefore, any "negligible association" result drawn for these predictors is considered a correct negligible association conclusion. Similarly, incorrectly concluding negligible association occurs when results suggest "negligible association" between a predictor and outcome, but the true effect is *outside of the equivalence interval* (i.e., $\beta \leq -\delta$ or $\beta \geq \delta$): β_4 and β_5 represent the levels at or larger than the SESOI value ($\beta \geq .15$), which is outside the equivalence interval. Thus, any "negligible association" result drawn for these predictors is considered an incorrect negligible association conclusion. A summary of correct and incorrect negligible association conclusions can be found in Table 2.

Note that β_1 was set to 0, the middle of the equivalence interval, to reflect the highest rates of correct conclusions with tests of equivalence. The greater the difference between the estimated β from $|\delta|$, the more likely a researcher is to correctly conclude a negligible association (greater power). However, true effects are rarely (if ever) exactly zero. We, therefore, tested at other effect sizes contained inside the equivalence interval (i.e., β_2, and β_3). By the same token, β_4 was set to .15 (δ), the cusp of the equivalence interval, to determine the highest rates of incorrect conclusions with equivalence tests. The interval's bound (i.e., δ or $-\delta$) is the lowest possible value outside the equivalence interval. Outside the equivalence interval, the farther the estimated β from δ, the less likely it is to incorrectly conclude a negligible association (lower error).

Table 2

Correct and Incorrect Conclusions of Negligible Association

	Equivalence-based test		Difference-based test	
	Negligible association concluded ($p < \alpha$)	Association concluded ($p \geq \alpha$)	Negligible association concluded ($p \geq \alpha$)	Association concluded ($p < \alpha$)
True negligible association ($\beta < \delta$)	Correct decision: Reject H_0 (Power)	Incorrect decision: Fail to reject H_0 (Type II error)	Correct decision: Fail to reject H_0	Incorrect decision: Reject H_0 (Type I error)
True association ($\beta \geq \delta$)	Incorrect decision: Reject H_0 (Type I error)	Correct decision: Fail to reject H_0	Incorrect decision: Fail to reject H_0 (Type II error)	Correct decision: Reject H_0 (Power)

Note: Equivalence-based tests (H_0: there is an association as defined by the SESOI interval, $-\delta$, δ) include the two one-sided tests (TOST) and Anderson and Hauck (AH) test. Difference-based test (H_0: there is a negligible association between a predictor and outcome) includes the traditional multiple regression coefficient analysis. *p* refers to the resulting *p* value, α refers to the set Type I error rate, and β refers to the individual predictor regression coefficient. The light grey shaded boxes indicate correctly concluding negligible association. The dark grey shaded boxes indicate incorrectly concluding negligible association. The light and dark grey boxes are the conditions tested in the simulation study. Note, however, that "negligible association" conclusions cannot be drawn from statistically nonsignificant difference-based tests; this aspect of the simulation is meant to mimic the practices used in the field for the purpose of comparing such practices to methodologically sound alternatives.

Results

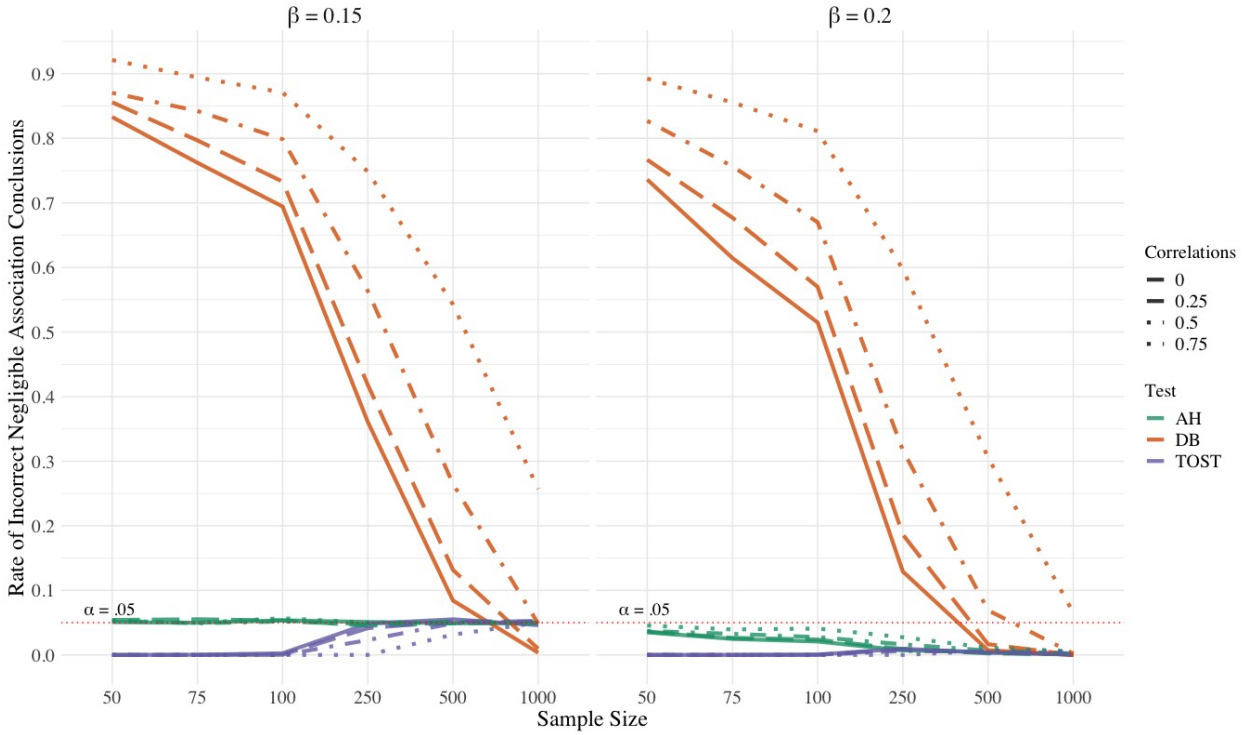
Simulation results are presented in Figure 2 and Figure 3. Note that the results, simulation code, and additional materials are also available on the Open Science Framework (OSF): https://osf.io/w96xe/.

Incorrectly Concluding Negligible Association

The probabilities of incorrectly concluding negligible association, when the predictor's true effect size falls outside the equivalence bounds ($\beta \geq \delta$), are illustrated in Figure 2.

Figure 2

Simulation Results: Incorrect Negligible Association Conclusions



Note. Incorrect negligible association conclusions are reflected by the effect size of predictors β_4 and β_5 which are presented at the top of each of the two graphs, respectively. Rates presented on the y-axis represent the proportion of incorrect conclusions for the traditional, difference-based (DB) test, two one-sided tests (TOST), and Anderson and Hauck's (AH) test. Different line types reflect the different relationship strength between predictor variables from completely independent predictors (solid line) to correlated at 0.75 (dotted line). The horizontal, dashed, red line indicates the nominal Type I error rate across all simulation conditions.

Difference-Based Approach

Simulation results from the difference-based approach suggest that the probability of incorrectly concluding negligible association is extremely high, especially for samples at or smaller than 100. These rates are even higher with stronger associations (i.e., correlation) between the predictor variables. For example, with a sample size of 100, a correlation of 0.25, and a true effect at β = .15, one has a 73.3% chance of *falsely* concluding a negligible association. Similarly, with the same sample size and correlation, one has a 57% chance of falsely concluding negligible association even with a true effect as large as β = .2. It is only with larger samples (n = 500, 1000) and weaker correlations (ρ = 0, 0.25), that the probability of incorrectly concluding negligible association comes somewhat close to the expected Type I error

rate (i.e., $\alpha = .05$). However, when the correlation between the predictors is strong (e.g., $\rho = 0.75$), even with large sample sizes (e.g., $n = 1000$), the rate of incorrect conclusion can still be very high (e.g., 25% when $\beta = \delta$).

Interestingly, although the performance of the difference-based test is affected by sample size and the magnitude of the effect, the traditional NHST is impervious to whether the effect is negligible. That is, for *any* association between a predictor and outcome variable (as long as it is not perfectly nil), the difference-based test will always be statistically significant (i.e., lower rates of incorrect negligible association conclusions) with a large enough sample size, irrespective of the presence or absence of a negligible effect. As evidence, we can see a similar, almost identical, downward slope pattern of the difference-based test in both Figure 2 (non-negligible effect) and Figure 3 (negligible effect), with the exception of when the effect is perfectly zero (which is discussed in the next section). This is not a characteristic we would like our test to possess if our goal is to determine a negligible association. It is not the case, however, with an equivalence-based approach.

Equivalence-Based Approach

Recall that a true effect of $\beta = .15$ is when we expect the highest rates of false rejections (see Simulation Conditions section). Simulation results demonstrate that the probability of falsely concluding negligible association when $\beta = .15$ stabilizes around the Type I error rate, or lower (i.e., $\leq 5\%$). These rates are very similar across all the correlation conditions for the TOST and virtually identical for the AH. These rates are the expected and appropriate error rates. When the true effect is larger than $\beta = .15$, however, the probability of falsely concluding negligible association approaches zero as the sample size increases. Again, these rates are essentially indistinguishable across the different correlation conditions for both equivalence tests.

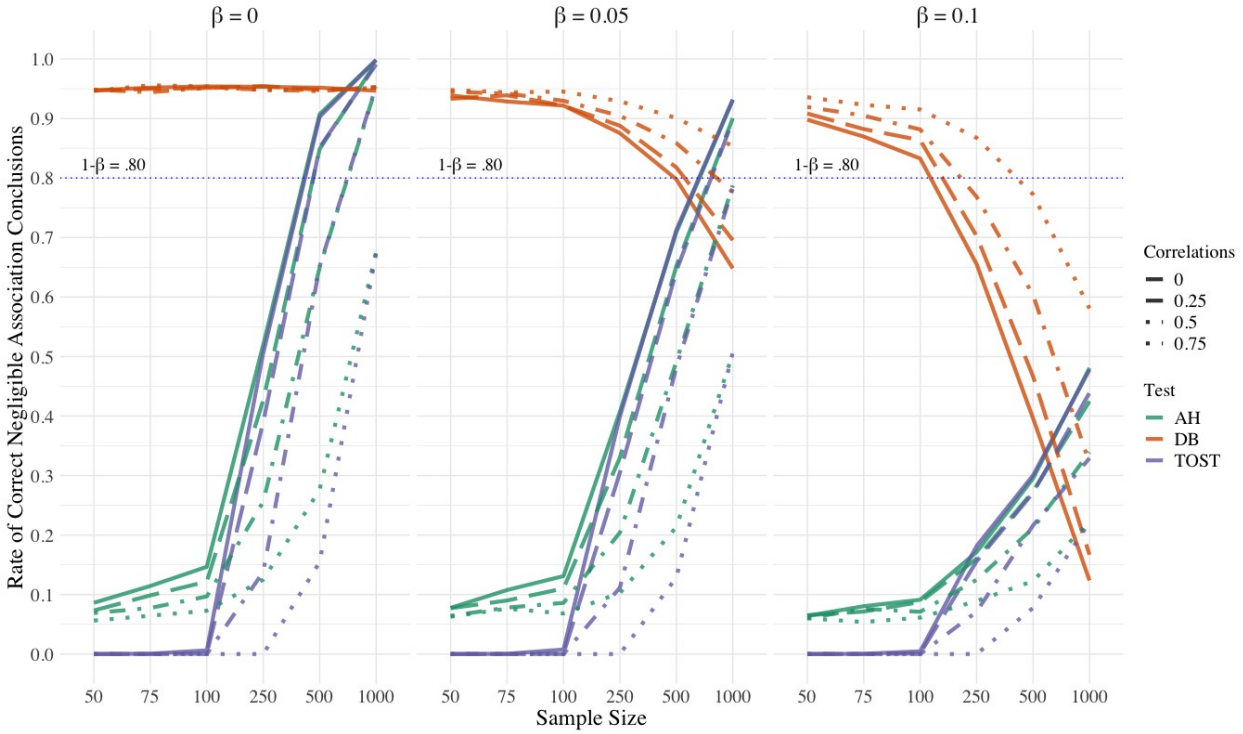
Anderson-Hauck (AH) versus Two One-Sided Tests (TOST). Simulation results from the two procedures suggest a similar pattern with slight differences. Here, the probability of incorrectly concluding negligible association using the AH procedure is slightly higher (precisely at the expected nominal Type I error rate, .05) than with the TOST, when $n \leq 100$. When the sample size is $n \geq 250$, however, the probabilities of incorrectly concluding negligible association using the two procedures converge and become practically the same.

Correctly Concluding Negligible Association

The probabilities of correctly concluding negligible association, when the predictor's true effect size falls within the equivalence bounds ($-\delta$, δ), are illustrated in Figure 3.

Figure 3

Simulation Results: Correct Negligible Association Conclusions



Note. Correct negligible association conclusions are reflected by the effect size of predictors β_1, β_2, and β_3, which are presented at the top of each of the three graphs, respectively. Rates presented on the y-axis represent the proportion of correct conclusions for the traditional, difference-based (DB) test, two one-sided tests (TOST), and Anderson and Hauck's (AH) test. Different line types reflect the different relationship strength between predictor variables from completely independent predictors (solid line) to correlated at 0.75 (dotted line). The blue, dashed line reflects the minimum desired statistical power in most psychological studies.

Difference-Based Approach

Simulation results from the difference-based approach suggest that the probability of correctly concluding negligible association decreases as the sample size increases. The inverse relationship between correct conclusions and sample size is prominent in small, non-zero effect sizes (see Figure 3), where higher rates are found with larger correlations. The exception is when the true effect is $\beta = 0$ (1st predictor, β_1) where the probability of correctly concluding negligible association using a difference-based approach, remains stable at around 95% across all *n* levels. However, this finding is expected because the $\beta = 0$ condition represents the case where the difference-based test's null hypothesis is perfectly true, i.e., where the probability of concluding

an association equals the pre-set Type I error rate, α (e.g., 5%). Because the difference-based results presented for our purposes are nonsignificant, we observe stable rates of concluding a negligible association around 1- α, or .95, regardless of sample size.

Notice that the probabilities of correct conclusions using the difference-based approach are high. However, this finding is deceiving and should not be considered in isolation from its associated error rates; the probabilities of correctly concluding negligible association using the difference-based approach are artificially inflated due to the extremely high rates of incorrect negligible association conclusions. For example, with a sample size of $n = 50$, a correlation of 0.25, and a true effect of $\beta = .1$, a negligible association is correctly concluded in about 90% of the cases, using the difference-based test. Nevertheless, with the same sample size, correlation, test, and similar effect size ($\beta = .15$), a negligible association is *falsely concluded in about 86% of the cases*.

Equivalence-Based Approach

In sharp contrast to the difference-based approach, both equivalence testing procedures demonstrate a strong, positive relationship between correct conclusions and sample size, where the probability of correctly concluding negligible association is significantly higher with larger samples and lower correlations between predictors. Importantly, with both the TOST and AH procedures, the probability of correctly concluding negligible association quickly increases when the sample size is greater than 100, regardless of effect size or correlation (except for TOST $\rho = 0.75$, which is when $n > 250$). However, the probabilities of correct negligible association conclusions are notably higher as the true effect is closer to zero, the middle of the equivalence interval, and with weaker correlations.

Anderson-Hauck (AH) versus Two One-Sided Tests (TOST). Simulation results from the two procedures reveal a similar pattern, albeit with some minor differences. The probability of correctly concluding negligible association using the AH procedure is slightly higher than with the TOST procedure, when the sample size is less than, or equal to 100, regardless of effect size or correlation. However, for any sample greater than 250, the probability of correctly concluding negligible association with the two equivalence-based procedures is virtually identical.

Determining Negligible Associations in Regression: A Practical Demonstration

Here, we re-analyze data from published studies using our proposed negligible association testing approach in multiple regression to illustrate the simplicity and necessity of the method. We also demonstrate how to perform and report these tests using the free and accessible `negligible` R package (Cribbie et al., 2022) and provide readers with an accompanying Shiny app (https://udialter.shinyapps.io/negreg-shiny/). To download the free, open-source software R, visit https://cran.r-project.org/. We also recommend downloading RStudio (https://www.rstudio.com/) for a more accessible interface.

Negligible effect testing can be applied when researchers have raw data or summary information from a regression table. The `negligible` package provides several functions designed to evaluate whether a negligible effect exists among variables in numerous statistical contexts such as between two means, among correlation coefficients, categorical data etc. In addition, the package provides graphics that help researchers interpret the results of the analyses. The package can be downloaded in R. The user must first download the package using the following command: `install.packages("negligible")`. To start using the functions

in the package, the user must then "call upon" - or load - the package by entering: `library(negligible)` in a line below.

To determine whether a certain predictor is practically and statistically negligible, we will use the `neg.reg` function found in the `negligible` package. There are two main approaches to using `neg.reg`. The first (and more recommended) is by entering a dataset (using the `data` argument) into the function. However, this function also accommodates cases where only summary statistics are available (e.g., coefficient value, sample size, SE, etc.) which are commonly found in the Results section of published articles. Next, we use examples from the literature to demonstrate how to use `neg.reg` to determine a negligible association in multiple regression and how to report and interpret the results.

Example 1: "Unlinking" Deliberate Mind Wandering and OCD Symptomatology

Our first example comes from Seli et al. (2017) where the authors investigated the relationship between everyday experiences of mind wandering and OCD symptomatology. Seli and colleagues make a clear distinction between two types of mind wandering: the first is an unintentional, *spontaneous* mind wandering (MW-S), and the second is a voluntary and *deliberate* off-task thought (MW-D). This distinction is important because the authors hypothesized that MW-S will show a meaningful relationship with each of the four dimensions of OCD (contamination, responsibility for harm and mistakes, unacceptable thoughts, and symmetry/completeness) whereas MW-D will demonstrate a *negligible, or no, association*.

Sample Details and Descriptive Statistics

Data come from 2636 undergraduate psychology students. The variables of interest for this demonstration are the two predictors, MW-S and MW-D, and the four dimensions of OCD

symptomatology, each of which serves as the outcome in four regression models (all four models have the same two predictors). Table 3 includes descriptive statistics and correlations.

Table 3

Means, standard deviations, range, and correlations with confidence intervals for Example 1

Variable	*M*	*SD*	min-max	1	2	3	4	5
1. MW-S	4.27	1.42	1-7					
2. MW-D	4.50	1.44	1-7	.40 [.37, .43]				
3. Contamination	3.53	3.15	0-17	.14 [.10, .18]	.05 [.01, .09]			
4. Responsibility for harm and mistakes	3.28	3.19	0-17	.22 [.18, .25]	.10 [.07, .14]	.54 [.52, .57]		
5. Unacceptable thoughts	3.98	3.83	0-19	.36 [.32, .39]	.12 [.08, .16]	.39 [.36, .42]	.48 [.45, .51]	
6. Symmetry / completeness	2.96	3.39	0-19	.20 [.16, .24]	.05 [.01, .09]	.51 [.48, .54]	.48 [.45, .51]	.42 [.39, .46]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. Each of the spontaneous mind wandering (MW-S) and deliberate mind wandering (MW-D) variable scores are averaged across four seven-point Likert scale items for each participant (*N*=2636). Scores on Contamination, Responsibility for harm and mistakes, Unacceptable thoughts, and Symmetry/completeness are summed for each participant across five five-point Likert scale items for a maximum score of 20.

Because Seli et al. (2017) sought to demonstrate that MW-D was negligibly (or not at all) associated with each of the four OCD dimensions, we can use the `neg.reg` function. But, before we do, we must first define what is a (practically) meaningful effect in this context, i.e., the SESOI. Ideally, the SESOI should be derived from substantive knowledge of the effects in the research area, prior to inspecting the data. Numerous approaches to selecting the SESOI are possible; however, these details are beyond the scope of this paper. For more information on justifying the SESOI, with examples, we recommend reading Anvari and Lakens (2021) and Lakens et al. (2018).

Selecting the SESOI

We decided to set the SESOI (i.e., δ) at 5% of the maximum possible score on the Contamination dimension from the OCD scale. We reasoned that a Contamination score equivalent to 5% would be low enough to have no serious practical significance (i.e., negligible). Although OCD research experts might set a slightly different SESOI, we will accept this value for the purpose of this demonstration. The Contamination subscale is measured from 0 to 20 (see Seli et al., 2017), therefore the SESOI was set at $b = 1$ (measured in unstandardized units of the outcome variable) such that our equivalence interval is (-1, 1): if the observed effect for MW-D and its associated uncertainty falls entirely within the SESOI bounds, from -1 to 1, we can conclude a negligible association between MW-D and Contamination.

`neg.reg` Function with Raw Data

The authors reported standardized effects, but here we use raw measurement units to make this example easier to interpret. Because the authors generously shared their original dataset, we successfully replicated the results in Seli et al. (2017) and calculated unstandardized regression coefficients. With access to the dataset, users should first import their data file into the RStudio environment. The functions and packages for importing data vary depending on the data file extension (e.g., .csv, .SAV). For instructions on how to import your data into the RStudio environment, readers are encouraged to follow the tutorial on the RStudio Support website. The object name under which the imported dataset is saved (e.g., `ocd`) should then be inserted as the input for the first argument in the `neg.reg` function, `data`, as shown in Listing 1.

The next argument, `formula`, requires the user to specify the regression model consisting of the outcome (i.e., criterion/dependent variable) to the left of the tilde symbol (~), followed by all the predictor variables in the model with the + sign between each predictor name.

Note that each argument is separated by a comma except for the last argument which precedes the close bracket, indicating the end of the function input. Users should identify the exact variable names in the imported dataset they are interested in modelling and pay close attention to lower or capital case. In our example, the first outcome variable is contamination (one of the OCD symptomatology dimensions), labelled here `ocd_cont`. We have only two predictors in our example model, MW-D and MW-S (`MWD + MWS`), which then go to the left of the ~ (see Listing 1).

The `predictor` argument asks users to specify which of the predictors they would like to test for a negligible effect. Because Seli et al. (2017) hypothesized a negligible or no association of MW-D with the outcomes, we specified `predictor = MWD` in our code example. The next two arguments, equivalence interval upper (`eiu`) and lower (`eil`) refer to the two SESOI bounds. In our example, the SESOI is set to 1, so we specified `eiu = 1` and `eil = -1.` Recall that the SESOI is set in unstandardized units (i.e., $b = 1$), therefore the next argument, `std`, which asks if the units are standardized, is set to `FALSE`. Although `std = FALSE` is a default in the function, we recommend explicitly making this distinction to avoid confusion or incorrect conclusions. The same negligible association testing with standardized effects is presented in the Supplementary Materials.

Finally, there are additional, optional features included in the function such as using bootstrap (and setting the number of iterations and/or seed) to calculate the standard errors, changing test type from AH (default) to TOST or nominal Type I error from .05 to another, custom rate, saving the resulted plots locally (e.g., as .png, or .jpeg) etc. These added features will not be discussed in this example, but readers are encouraged to find more information about the `neg.reg` features and arguments in the `negligible` package documentation.

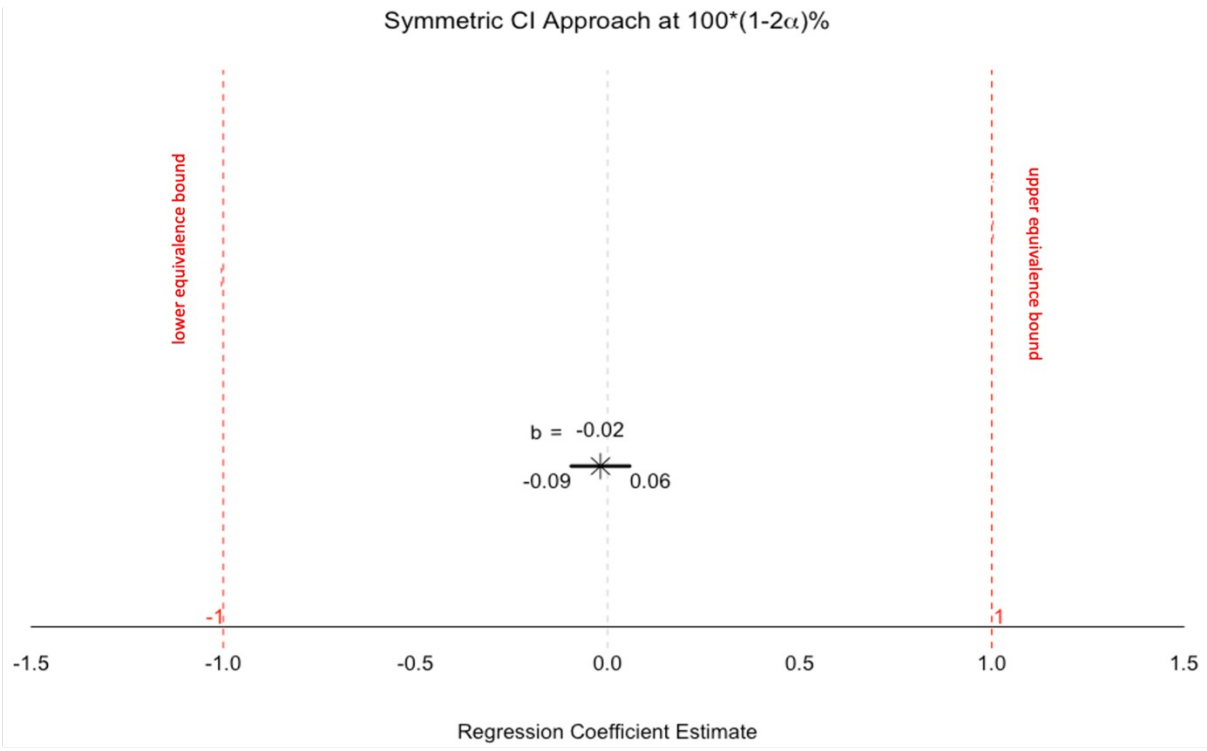
Listing 1. Code Block Input Using `negligible::neg.reg` for Example 1

```
library(negligible)
neg.reg(data = ocd, # name of dataset
        formula = ocd_cont~MWD+MWS, # regression formula
        predictor = MWD, # name of the predictor of interest
        eiu = 1, # upper bound of SESOI (unstandardized)
        eil = -1, # lower bound of SESOI (unstandardized)
        std = FALSE, # using unstandardized units
        bootstrap = FALSE) # not using bootstrap in example
```

At this point, all the necessary inputs are in place and we can now run the function by executing the block of code. The output from the function, containing both results from the negligible association testing and illustrating graphics are presented in Figure 4.

Figure 4. `neg.reg` Function Output: Negligible Association Testing Results for Example 1

```
*** Evaluating Negligible Effects Between Predictor and Outcome in Multiple Regression ***

Unstandardized regression coefficient for MWD:
b = -0.018, 95% CI [-0.109, 0.072]
std. error = 0.046

**********************

Anderson-Hauck (AH) procedure:

Equivalence interval: lower= -1, upper= 1
Anderson-Hauck T statistic = -0.4
p < 0.001
NHST decision: The null hypothesis that the regression coefficient is non-negligible can be
rejected. A negligible effect is concluded. Be sure to interpret the magnitude (and precision) of
the effect size.

*Note that in rare cases, where p is close to α, NHST decisions using the AH procedure may not
match the TOST NHST results or the Symmetric CI Approach at 100*(1-2α)% illustrated in the plots.

**********************
```



Note. Results from the equivalence test are presented below at the top of the figure and a graphical illustration of the Symmetric Confidence Interval (CI) Approach is presented at the bottom. Here, we can reject the null hypothesis that the effect size falls outside of the SESOI bounds and find inferential evidence in support of a statistical and practical negligible association between MW-D and contamination.

Reporting and Interpreting `neg.reg` Output

Negligible association testing results (top of Figure 4) show that MW-D is indeed statistically negligible, $b_{\text{MW-D}} = -0.018$, $SE = .046$, 90% CI [-0.09, 0.06], AH T statistic = -0.40, $p < .001$. To illustrate the magnitude of the effect graphically, the `neg.reg` function also provides a visualization of the regression coefficient's point estimate and its associated 90% CI[1] in relation to the SESOI band (the region within the vertical red dashed lines) and its center (vertical grey dashed line), as demonstrated in the bottom of Figure 4.

We can therefore reject the null hypothesis that the effect size falls outside of the negligible effect bounds and find inferential evidence in support of a statistical and practical negligible association between MW-D and contamination. In other words, given the predefined SESOI of 1 point on OCD symptomatology (scale ranging from 0-20) and Type I error rate of .05, deliberate mind wandering was found to be a negligible predictor of contamination while holding spontaneous mind wandering constant. We can further observe from the bottom of Figure 4 that the effect size estimate ($b = -0.02$) is closely centered around 0 and the CI band is but a small proportion of the entire SESOI area. Finally, both the effect size and CI are completely contained within the SESOI area and are reasonably distant from either bound.

In Example 1, we used the `neg.reg` function with access to the raw data to answer whether deliberate mind wandering truly has a negligible association with contamination, partialling out the effect of spontaneous mind wandering. Results from the negligible effect testing approach using the `neg.reg` function are congruent with Seli et al.'s (2017) conclusion, which was obtained from a statistically nonsignificant difference-based test. It is often the case, however, that results from negligible effect testing contradict negligible association conclusions

---

1 Recall that for the negligible effect test's null hypothesis to be rejected, the entire span of the $100 \cdot (1 - \alpha)$% CI for the associated regression coefficient must be contained between the two SESOI bounds.

made using results from a nonsignificant difference-based test; this scenario will be demonstrated in Example 2 in the following section. Example 2 also shows how users can employ the `neg.reg` function without access to raw data.

Example 2: Personal Control Moderates the Association Between Organizational Stability and Identification

Proudfoot and Kay (2018) predicted that feelings of personal control would moderate the relationship between perceived organizational stability and organization identification. Specifically, they wished to demonstrate that a relationship between organizational stability and identification exists for participants with low personal control ("control threat" condition), but that *no such effect is present* for participants with high personal control ("control affirmation" condition). To test their hypothesis, the authors modelled an interaction between organizational stability and personal control on organization identification in their multiple regression analysis. As reported in Study 3, the control × stability interaction (higher-order effect) was indeed statistically significant. Inspecting the simple slopes, the authors found that participants in the "control threat" condition exhibited a statistically significant relationship between stability and identification, whereas participants in the "control affirmation" condition did not, $b = 0.15$, $SE = .12$, $t(190) = 1.18$, $p = .24$. It was concluded that "for participants who recalled an event wherein they had control, there was no effect of perceived organizational stability on identification" (Proudfoot & Kay, 2018, p. 110).

`neg.reg` Function with No Raw Data

In this example, we do not have access to the dataset. Still, we can formally test whether the association between perceived organizational stability on identification is indeed negligible for participants in the "control affirmation" condition using the `neg.reg` function. All the

information we need can be easily gathered from the reported results in Proudfoot and Kay (2018). Specifically, we will need the following: the regression coefficient point estimate for organizational stability ($b = 0.15$) and its associated standard error ($SE = 0.12$), the sample size used in the analysis ($n = 194$), the number of predictors in the multiple regression model ($k = 3$), and the nominal Type I error rate ($\alpha = .05$). Note that the reported effect and its associated standard error are measured in unstandardized units. As a reminder to readers, if the standard error is not reported, users can simply divide the effect size (i.e., the regression coefficient) by the *t* statistic to retrieve the standard error. This applies to both standardized and unstandardized effects.

Next, we must identify our definition of practical significance in this context (selecting a SESOI). The outcome variable in this example (organization identification) is an average score of six items on a seven-point scale (from "Strongly disagree" to "Strongly agree"). In thinking about the scale, we propose that the *minimum meaningful effect* is a one-point difference (e.g., from "Strongly disagree" to "Disagree") on at least one item from the six that are asked. Here, we reference a one-point difference in the scale's total score as the *anchor* for (the smallest) important difference in identifying with one's organization. Translated to an average score, this difference is about 0.33 points on the organization identification scale. Accordingly, our SESOI will be set at $b = 0.33$ (measured in unstandardized units).

Ideally, anchors for gauging meaningful effects should be planned independently of the study's results (in this example, we were exposed to the observed effect before proposing the SESOI) and be estimated carefully with experiments (for guidelines on how to estimate the SESOI using anchor-based methods, see Anvari & Lakens, 2021). Thus, the SESOI we selected

in this example is justified, but not validated. Still, we will proceed with this value for the purpose of our demonstration.

We can now plug the input into the `neg.reg` function's arguments as demonstrated in Listing 2.

Listing 2: Code Block Input Using `negligible::neg.reg` for Example 2

```
library(negligible)
neg.reg(b = 0.15, # effect size of the predictor of interest
        se = 0.12, # standard error associated with the effect
        n = 194, # sample size used in the analysis
        nop = 3, # number of predictors
        eiu = 0.33, # upper bound of SESOI (unstandardized)
        eil = -0.33, # lower bound of SESOI (unstandardized)
        std = FALSE) # using unstandardized units
```

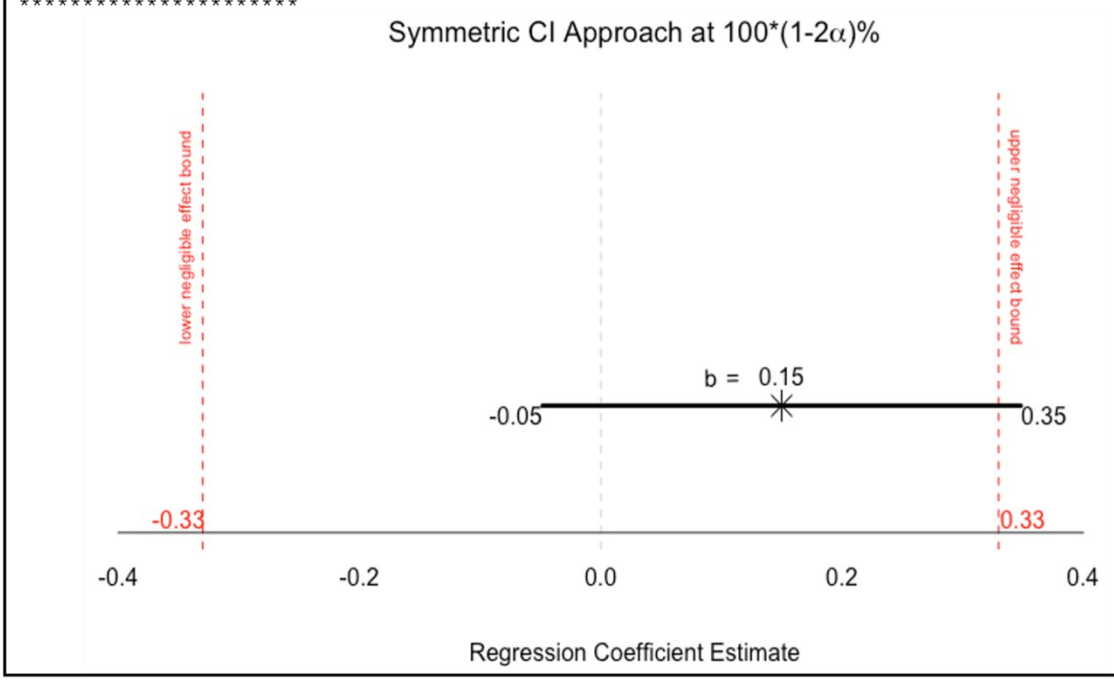
Output from the code in Listing 2 is presented in Figure 5. Negligible effect testing results (top of Figure 5) were not statistically significant, $b = 0.15$, $SE = .12$, 90% CI [-0.05, 0.35], AH T statistic = 1.25, $p = .068$, suggesting that negligible association cannot be concluded. That is, given the predefined SESOI of 0.33 points on organizational identification and $\alpha = .05$, there is insufficient evidence that organizational stability has "no effect" on organizational identification among participants in the control affirmation condition. But, perhaps more important than the significance tests results is the estimated effect and its precision (Amrhein et al., 2019): as seen in Figure 5, the observed effect is relatively distant from zero, its 90% confidence band which ranges from -0.05 to 0.35 is somewhat wide (close to two-thirds of the

entire negligible effect region between the two SESOI bounds), and at least some values within the 90% CI are greater than the upper SESOI bound.

Inferential results from negligible effect testing do not support a lack of relationship between organizational stability and identification. Note that, other than deciding on the SESOI, the input used for the negligible effect testing analysis is exactly the same as the information extracted from the sample in the published article. Yet, using a methodologically appropriate inferential test, the results are incongruent with the conclusions in the original paper.

Figure 5. `neg.reg` Function Output: Negligible Association Testing Results for Example 2

```
*** Evaluating Negligible Effects Between Predictor and Outcome in Multiple Regression ***

Unstandardized regression coefficient for the predictor of interest:
b = 0.15, 95% CI [-0.087, 0.387]
std. error = 0.12

**********************

Anderson-Hauck (AH) procedure:

Equivalence interval: lower= -0.33, upper= 0.33
Anderson-Hauck T statistic = 1.25
p = 0.068
NHST decision: The null hypothesis that the regression coefficient is non-negligible cannot be
rejected. There is insufficient evidence to conclude a negligible effect. Be sure to interpret
the magnitude (and precision) of the effect size.

*Note that in rare cases, where p is close to α, NHST decisions using the AH procedure may not
match TOST NHST results or the Symmetric CI Approach at 100*(1-2α)% illustrated in the plots.

**********************
```



Note. Results from the equivalence test are presented at the top of the figure and a graphical illustration of the Symmetric Confidence Interval (CI) Approach is presented at the bottom. Here, we cannot reject the null hypothesis that the effect size falls outside of the SESOI bounds. Thus, we do not find inferential evidence in support of a statistical and practical negligible association between stability and identification among participants in the control affirmation condition.

Discussion

Behavioural and social researchers often aim to detect a negligible association between a predictor variable and outcome. However, for a lack of better statistical tools and awareness, researchers continue to incorrectly use nonsignificant results from traditional regression analysis to demonstrate negligible effects. In this paper, we sought to provide researchers with an appropriate method for detecting negligible association in multiple regression. The proposed negligible effect testing methods were evaluated and compared to the traditional, difference-based approach using a Monte Carlo simulation. Simulation results support the suitability of the equivalence-based approach and demonstrate its applicability over the difference-based test to detect negligible effects between predictors and outcomes in multiple regression.

This paper also offers a brief tutorial on how negligible effect testing can be implemented in psychological research with examples. We introduce the `neg.reg` function from the `negligible` package and demonstrate how researchers can easily test for negligible associations in multiple regression within the R/RStudio environment. We further provide an accompanying Shiny app (https://udialter.shinyapps.io/negreg-shiny/) for users who prefer a non-syntax-based interface. Finally, results reporting and interpretations are discussed with specific guidelines and recommendations.

Equivalence-Based versus Difference-Based Approaches

The traditional, difference-based approach resulted in substantially higher rates of incorrectly concluding a negligible association between a predictor and outcome variables, and even more so when the relationship between predictors is strong, than those of the equivalence-based approach. This is not the case when using the equivalence-based tests: both equivalence tests (i.e., AH or TOST) reveal acceptable rates (i.e., at or below the nominal Type I error rate)

of incorrectly concluding a negligible association which are *robust* to sample size, correlation, or magnitude of effect fluctuations.

In addition, the rates of correctly concluding a negligible association using the difference-based approach decrease as a function of sample size regardless of the correlation strength between predictors (illustrated in Figure 3). From a theoretical standpoint, this relationship is illogical; the closer a sample is to a population from which it was taken (by increasing sample size), the less chance a difference-based test has to find the true nature of a relationship (or lack thereof). Unlike the traditional approach, equivalence-based tests demonstrate the appropriate relationship between sample size and correct negligible association conclusions; this is the acceptable and excepted relationship between sample size and statistical power.

Although equivalence-based tests represent a better alternative to difference-based tests, one drawback is when the sample size is small ($n \leq 100$), the probability of correctly concluding a negligible association between a predictor and outcome is low. This means that equivalence tests - as with many other statistical tests - are not particularly effective when using small sample sizes. It is only with samples around $n = 500$ (when predictors are weakly correlated and the population effect size is closer to the middle of the equivalence interval) acceptable probabilities of correct conclusions (i.e., $\geq .80$) emerge. But, when predictors are strongly correlated and the true effect size is farther away from the middle of the equivalence interval, acceptable probabilities of correct conclusions will only be attained with sample sizes larger than 1000.

Why Negligible Effect Tests *Should* Require Larger Sample Sizes

Indeed, an equivalence testing approach to determining a negligible association between a predictor and outcome using small samples may be inefficient due to low correct conclusion rates. Although frustrating for researchers who wish to demonstrate negligible effects, this

"inefficiency" might serve as a constructive safeguard. First, acknowledging the difficulty in concluding negligible effects using the appropriate statistical tests with small samples may persuade researchers to increase their sample size. Given rising concerns about low-powered studies and their contribution to the replication crisis (e.g., Anderson & Maxwell, 2017; Crutzen & Peters, 2017; Maxwell et al., 2015), sufficiently large samples could help reduce questionable research practices and false findings.

Second, the difficulty in providing evidence for negligible effects with smaller samples conveys a greater burden of proof. In fact, in formal logic, proving a negative or providing evidence of absence (e.g., non-white swans do not exist) is more challenging than proving the existence of an effect or phenomenon. For example, it takes only one observation of a phenomenon (e.g., a black swan) to claim the existence of something with *certainty*. Though, it would take a very large number of observations of non-existence of the phenomenon (e.g., white swans) only to *infer* or make a *probable* conclusion of non-existence, which would still be without absolute certainty. Returning to psychology, providing evidence of absence (i.e., negligible or no effect) often *should* be more difficult than demonstrating an effect, for example, claiming no adverse effects from a new treatment or practice would be perilous with only a small number of participants. Further, it is potentially dangerous to conclude negligible effects with a high margin of error (due to a small sample size), even if the estimated average effect is itself negligible. Thus, the "inefficiency" of equivalence-based approaches with small sample sizes may actually function as a protective mechanism against such dangers.

Finally, the traditional difference-based test may seem more effective in small sample sizes. However, they do not require a high level of precision surrounding the observed effect; as long as the effect size-to-error ratio is sufficiently large, the null hypothesis is rejected,

regardless of the width of the CI. Although the existence of an effect is declared, our estimation of the effect's magnitude can be extremely imprecise. In contrast, tests of equivalence require a specific amount of precision to declare a negligible association. Namely, the 90% CI around the observed effect must not exceed either the upper or lower bound of the SESOI. This requirement manifests in the aforementioned "inefficiency," but also warrants us to have less uncertainty in our effect size estimates and conclusions. Importantly, focusing on the magnitude of effects, precision, and uncertainty beyond the decision to reject or not to reject the null hypothesis is in line with recommended practices (e.g., Cribbie, 2020; Cumming, 2012, 2014; Farmus et al., 2020; Fidler & Loftus, 2009). In fact, we strongly encourage readers to take stock of the magnitude of the relationship between a predictor and outcome, the CI, and the proportion and location in relation to the SESOI region, regardless of the NHST decision.

Important Considerations for Researchers Testing Negligible Effects

Thinking More About Units of Measurement

In Example 1, we demonstrated how tests of negligible association can be used on either unstandardized or standardized (in Supplemental Materials) effects and noted the inferential statistics for each are identical. But, it is necessary to be consistent with the units of measurement: A raw regression coefficient (*b*) must be used only with its associated error and an unstandardized SESOI, whereas a standardized coefficient (β) should only be used with its associated error and standardized SESOI. Mixing unstandardized and standardized units when using an equivalence (or difference-based) test would yield inaccurate results and likely lead to invalid conclusions. Thus, researchers must be cognizant of the measurement units of the effects and their compatibility. For instructions and details about converting one form to another, see Supplemental Materials.

In Example 2, we demonstrated how to determine a negligible association when probing the simple slopes following a statistically significant interaction in multiple regression. The use of equivalence testing to probe a significant interaction is no different than when used on any other predictor. But, equivalence testing can also be applied to test whether a modelled interaction (the higher-order effect or product term) is statistically and practically negligible. The procedure would be similar to that of an individual predictor (first-order effect) in multiple regression. However, the units of measurement are slightly different because the interpretation of an interaction's effect is different than that of a predictor. Recall that the SESOI must be of the same units as the interaction term. Therefore, an interaction term's SESOI represents the minimum meaningful change in simple slopes for the relationship of interest per one-unit difference on the moderator. Researchers must then consider their SESOI in these terms when testing for a negligible interaction. There is a lot more to be said about testing for a negligible interaction which is beyond the scope of this paper. However, we recommend that interested readers refer to Cribbie et al. (2016) for testing an interaction with categorical predictors and to Jabbari and Cribbie (2021) for continuous predictors.

Selecting Your SESOI

Selecting the right SESOI is perhaps the most important requirement when testing for a negligible effect. This topic has been covered in previous research (see earlier section on Defining Negligible Association), and recommendations for selecting the right SESOI are no different in a multiple regression framework than it is in other forms of statistical analyses. Selecting a SESOI value is independent of the methodology discussed in this study; it is a decision that is field- and context-specific, and it may be different from one researcher to another. In pharmaceutics, standardized methods and guidelines exist for determining SESOI

bounds. For example, the Food and Drug Administration (FDA) specifies that the differences in efficacy between two drugs (e.g., an established and a new, experimental drug) must not exceed 20% (after applying a log transformation) for the two drugs to be declared equivalent (FDA, 2021). In psychology, specifying standardized values of negligible effects for one scale might not be appropriate for another, which makes identifying the right value a difficult task. Rogers et al. (1993) rightfully noted, "as with any statistical analysis, equivalency procedures must involve thoughtful planning by the investigator" (p. 564). However, due to the lack of standardized methods for identifying a SESOI value in psychology, deciding on such values may be subjective and naturally introduce some biases. The SESOI value directly affects equivalence test results whereby larger SESOI values (wider equivalence interval) would make it easier to reject the null hypothesis and conclude negligible effects. Therefore, to avoid researchers' self-serving bias and questionable research practices (see John et al., 2012) it is crucial that researchers select a SESOI a priori, independently of the sample or test results (i.e., statistical significance), and with a strong justification grounded in theory and/or practical implications.

Focusing on Effect Size, Precision, and Practical Implications

Equivalence testing is a method designed within the NHST framework. NHST has been heavily criticized for its overreliance on the dichotomous results of *p* values with little, or no consideration of the effect's magnitude or its implications in practice (e.g., Cumming, 2012; Fidler & Loftus, 2009; Harlow, 1997; Kirk, 2003; Lee, 2016 2014). Researchers must be mindful of the limitations of NHST and disentangle the practical and statistical aspects of the test results. Equivalence testing has the added benefit of comparing an effect size with a value of practical significance (i.e., SESOI). To that extent, the null hypothesis from an equivalence test inherently includes information about the magnitude of meaningful effects. However, equivalence testing

results and conclusions are still tied to *p* values and are, therefore, not immune to criticisms about relying on binary decisions (“null rejected” or “null not rejected”).

To minimize the limitations of *p* values, it is more informative to interpret the observed effect’s magnitude and precision beyond the conclusion of “negligible effects" or “insufficient evidence for negligible effects,” as demonstrated in Examples 1 and 2 (Amrhein et al., 2019). Observed effects should be construed in relation to the SESOI bounds, the extent of their uncertainty (i.e., width and limits of the confidence interval), and their practical implications (or lack thereof). For example, if a researcher finds $p = .05$ from their test of negligible association, they have insufficient evidence in favour of the alternative hypothesis (i.e., negligible effects). In this case, however, the practical implications of the (negligible) effect would probably not be meaningfully different than if $p = .049$, despite the conflicting NHST decisions. Furthermore, because *p* values are directly influenced by sample size and variability, Type I error rate, SESOI value, and observed effect size, the slightest change in one of these factors might lead to a different binary inferential conclusion. These aspects must be taken into account and considered when interpreting the observed effects and test results. For this reason, the `negligible` R package (Cribbie et al., 2022) introduced in this paper also includes a graphical representation of the observed effect and its associated uncertainty in relation to the SESOI. The resulting plots aid in illustrating how close or far and wide or narrow the observed effect and its margins of error are from the SESOI bounds; inferring the proportion and position of the confidence band in relation to the SESOI bounds can help interpret the results over and above *p* values.

Limitations of the Simulation Study

Naturally, the current study has limitations. One limitation is that the simulated data were all normally distributed, with no missing data, or threats to the assumptions underlying multiple

regression. Here, test performance for the two approaches was assessed under ideal conditions whereas data analyzed in psychological studies often demonstrate different degrees of skewness, kurtosis, missing data, assumption violations etc. Therefore, the generalizability of the findings in this study is constrained to the conditions specified above. Another potential limitation is that other simulation conditions could have been tested. For example, different SESOI levels, additional sample sizes, or varying Type I error rates. However, results from any additional conditions are more than likely predictable from the equations and simulation results in this or previous studies.

Conclusion

Traditional difference-based tests are methodologically inappropriate for testing hypotheses about negligible associations. Instead, researchers should use the suitable alternative of negligible effect testing. We demonstrated these claims in a Monte Carlo simulation study, discussed the theoretical underpinning and implications of using negligible effect testing, and provided recommendations for researchers. Using user-friendly tools such as the `negligible` package and `neg.reg` Shiny app, researchers have free and easy access to appropriate methods to test negligible associations in regression. All materials, including R code, results, and slides are available on OSF: https://osf.io/6pmby/.

ACKNOWLEDGEMENT

The authors would like to thank Stephen Want, Andie Noack, Robert Cribbie, Denis Cousineau, Chris Aberson, and two anonymous reviewers for their comments, questions, and suggestions that have improved this paper. We gratefully acknowledge the work of Naomi Martinez Gutierrez who created the Shiny application for the neg.reg function. Lastly, we thank Paul Seli for generously sharing their data which enormously helped to illustrate the use of the new methodology and open-access tools described in this paper.

References

Abramowitz, J. S., Deacon, B. J., Olatunji, B. O., Wheaton, M. G., Berman, N. C., Losardo, D., Timpano, K. R., McGrath, P. B., Riemann, B. C., Adams, T., Björgvinsson, T., Storch, E. A., & Hale, L. R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the Dimensional Obsessive-Compulsive Scale. *Psychological Assessment, 22*(1), 180–198. https://doi.org/10.1037/a0018260

Altman, D., & Bland, J. (1995). Absence of Evidence Is Not Evidence of Absence. *British Medical Journal. 311*. 485. https://doi.org/10.1136/bmj.311.7003.485

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73*(sup1), 262-270. https://doi.org/10.1080/00031305.2018.1543137

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

Anderson, S. F, & Maxwell, S. E. (2017) Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research, 52*(3), 305-324. https://doi.org/10.1080/00273171.2017.1289361

Anvari, & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96*, 104159–. https://doi.org/10.1016/j.jesp.2021.104159

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science, 11* ,283-319. https://doi.org/10.1214/ss/1032280304

Beribisky, N., Mara, C. A., Cribbie, R. A. (2020) An Equivalence Testing Approach for Evaluating Substantial Mediation. *The Quantitative Methods for Psychology, 16*(5), 424-441. https://doi.org/10.20982/tqmp.16.4.p424

Berkson, J. (1938) Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. *Journal of the American Statistical Association, 33*(203), 526-536. https://doi.org/10.1080/01621459.1938.10502329

Brown, L. D., Hwang, J. T. G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics, 25* ,2345-2367. https://doi.org/10.1214/aos/1030741076

Butcher, J.N., Graham, J.R., Ben-Porath, Y.S., Tellegen, A., Dahlstrom, W.G., & Kaemmer, B. (2001). *MMPI-2 (Minnesota Multiphasic Personality Inventory-2): Manual for administration, scoring, and interpretation, revised edition*. Minneapolis: University of Minnesota Press.

Caldwell, A. R. (2022, November 17). Exploring Equivalence Testing with the Updated TOSTER R Package. https://doi.org/10.31234/osf.io/ty8de

Campbell, H. (2020). Equivalence testing for standardized effect sizes in linear regression. *arXiv*. https://doi.org/10.48550/arXiv.2004.01757

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., & Lakens, D. (2021). Can we disregard the whole model? omnibus non-inferiority testing for R^2 in multi-variable linear regression and η^2 in ANOVA. *British Journal of Mathematical & Statistical Psychology, 74*(1), 64-89. https://doi.org/10.1111/bmsp.12201

Carriere, J. S. A., Seli, P., & Smilek, D. (2013). Wandering in Both Mind and Body: Individual Differences in Mind Wandering and Inattention Predict Fidgeting. *Canadian Journal of Experimental Psychology, 67*(1), 19-31. https://doi.org/10.1037/a0031438

Cohen, J. (1990) The Things I Have Learned (So Far). *American Psychologist*. 45, 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Counsell, A. & Cribbie, R. A. (2016). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology, 68*: 292-309. https://doi.org/10.1111/bmsp.12045

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020) Evaluating Equivalence Testing Methods for Measurement Invariance. *Multivariate Behavioral Research, 55*(2), 312-328. https://doi.org/10.1080/10705511.2015.1065414

Cribbie, R. A., Alter, U., Beribisky, N., Chalmers, R. P., Counsell., A., Farmus, L., Martinez Gutierrez, N., Ng, V. (2022). *negligible: A Collection of Functions for Negligible Effect/Equivalence Testing*. R package version 0.1.2. https://cran.r-project.org/web/packages/negligible/index.html

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

Cribbie, R. A., Panzarella, E., Farmus, L., Martinez Gutierrez, N., Beribisky, N., & Alter, U. (2020, February 24). *Effect Size Reporting in Social-Personality Research: The Good, the Bad, and the Ugly* [Presentation]. Quantitative Methods Area Forum, York University, Toronto, ON, Canada.

Cribbie, R.A., Ragoonanan, C. and Counsell, A. (2016), Testing for negligible interaction: A coherent and robust approach. *Br J Math Stat Psychol, 69*: 159-174. https://doi.org/10.1111/bmsp.12066

Crutzen, R., & Peters, G. J. Y. (2017). Targeting next generations to change the common practice of underpowered research. Frontiers in psychology, 8, 1184. https://doi.org/10.3389/fpsyg.2017.01184

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York, NY: Taylor & Francis Group, LLC

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables. *Biometrics, 33*(4), 593-602. https://doi.org/10.2307/2529457

Farmus, L., Beribisky, N., Martinez Gutierrez, N., Alter, U., Panzarella, E., & Cribbie, R. A. (2022). Effect size reporting and interpretation in social personality research. Current Psychology. https://doi.org/10.1007/s12144-021-02621-7

Farruggia, S. P., Chen, C., Greenberger, E., Dmitrieva, J., & Macek, P. (2004). Adolescent self-esteem in cross-cultural perspective: Testing measurement equivalence and a mediation model. *Journal of Cross-Cultural Psychology, 35*(6), 719-733. https://doi.org/10.1177/0022022104270114

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift Für Psychologie/Journal of Psychology, 217*(1), 27-37. https://doi.org/10.1027/0044-3409.217.1.27

Fletcher, T. D. (2012). QuantPsyc: Quantitative Psychology Tools. R package version 1.5. https://CRAN.R-project.org/package=QuantPsyc

Flora, D. B. (2018). *Statistical Methods for the Social and Behavioural Sciences: A Model-Based Approach*. London, UK. SAGE Publications Ltd.

Food and Drug Administration. (2021). Approved drug products with therapeutic equivalence evaluations (41st ed). U.S Department of Health and Human Services. Retrieved from http://www.fda.gov.

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. https://doi.org/10.1348/000711009X475853

Harlow, L.L. (1997). Significance Testing in Introduction and Overview. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.). *What If There Were No Significance Tests?* (pp.1-17). Mahwah, NJ, USA: Lawrence Erlbaum.

Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics, 12*(1), 83-91. https://doi.org/10.1007/BF01063612

Jabbari, Y., & Cribbie, R. (2021). Negligible interaction test for continuous predictors. *Journal of applied statistics, 49*(8), 2001–2015. https://doi.org/10.1080/02664763.2021.1887102

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kim, Y. J., & Cribbie, R. A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology, 71*(1), 1-12. https://doi.org/10.1111/bmsp.12103

Kirk, R. E. (2003). *The importance of effect magnitude*. In S. F. Davis (Ed.), Handbook of research methods in experimental psychology (pp. 83–105). Malden, MA: Blackwell.

Kupats, E., Noviks, I., Vrublevska, J., Kenina, V., Kojalo, U., & Logina, I. (2018). No relationship between generalized anxiety symptoms and cardiovascular autonomic dysfunction. *Neurology, Psychiatry and Brain Research, 30*, 86-90. https://doi.org/10.1016/j.npbr.2018.07.001

Lakens, D. (2022, June 21). *Using p-values to test a hypothesis*. Improving Your Statistical Inferences. https://lakens.github.io/statistical_inferences/pvalue.html

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/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

LeBeau, B. (2019). SIMGLM: Simulate Models Based on the Generalized Linear Model. R package version 0.7.4. http://CRAN.R-project.org/package=simglm

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology, 69*(6), 555-562. https://doi.org/10.4097/kjae.2016.69.6.555

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61*(2), 107-116. https://doi.org/10.1080/00049530802105865

Martinez Gutierrez, N., & Cribbie, R. A. (2022). Effect sizes for equivalence testing: incorporating the equivalence interval [Manuscript submitted for publication]. Department of psychology, York University.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487–498. https://doi.org/10.1037/a0039400

Metzler, C. M. (1974). Bioavailability - A problem in equivalence. *Biometrics, 30*(2), 309-317. https://doi.org/10.2307/2529651

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208-225. https://doi.org/10.1037/met0000126

Proudfoot, D., & Kay, A. C. (2018). How perceptions of one's organization can affect perceptions of the self: Membership in a stable organization can sustain individuals' sense of control. *Journal of Experimental Social Psychology, 76*, 104-115. https://doi.org/10.1016/j.jesp.2018.01.004

Quintana, D. S. (2016). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology, 54*: 344-349. https://doi.org/10.1111/psyp.12798

Quintana, D. S. (2018, July 21). Using summary statistics to determine whether a non-significant result supports the absence of an effect [Blog post]. Retrieved from https://blog.usejournal.com/using-summary-statistics-to-determine-whether-a-non-significant-result-supports-the-absence-of-an-1ff61e97f7cf

R Core Team (2021). R: A Language and Environment for Statistical Computing. R foundation for statistical computing, Vienna, Austria. URL http://CRAN.R-project.org

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Robinson, A. P., Duursma, R. A., & Marshall, J. D. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree Physiology, 25*(7), 903-913. https://doi.org/10.1093/treephys/25.7.903

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565. https://doi.org/10.1037/0033-2909.113.3.553

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 ,657-680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3* ,403-411. https://doi.org/10.1037/1082-989X.3.4.403

Seli, P., Risko, E. F., Purdon, C., & Smilek, D. (2017). Intrusive thoughts: Linking spontaneous mind wandering and OCD symptomatology. *Psychological Research, 81*(2), 392-398. https://doi.org/10.1007/s00426-016-0756-3

Steptoe, A., Kunz-Ebrecht, S. R., & Owen, N. (2003). Lack of association between depressive symptoms and markers of immune and vascular inflammation in middle-aged men and women. *Psychological Medicine, 33*(4), 667-674. https://doi.org/10.1017/S0033291702007250

Thompson, B. (1992), Two and One-Half Decades of Leadership in Measurement and Evaluation. *Journal of Counseling & Development, 70*: 434-438. https://doi.org/10.1002/j.1556-6676.1992.tb01631.x

Tukey, J.W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science, 6*, 100-116. https://doi.org/10.1214/ss/1177011945

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61*(8), 1340-1341. https://doi.org/10.1002/jps.2600610845

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*(4), 741-744. https://doi.org/10.2307/2529259

Westlake, W. J. (1981). Response to T. B. L. Kirkwood: Bioequivalence testing – a need to rethink. *Biometrics, 37*, 589–594. https://doi.org/10.2307/2530573

Wheadon, D. E., Rampey, A. H., Thompson, V. L., Potvin, J. H., Masica, D. N., & Beasley, C. M. (1992). Lack of association between fluoxetine and suicidality in bulimia nervosa. *Journal of Clinical Psychiatry, 53*, 235–241. https://doi.org/10.1016/j.jpsychires.2013.05.025

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405–426. https://doi.org/10.1037/met0000080

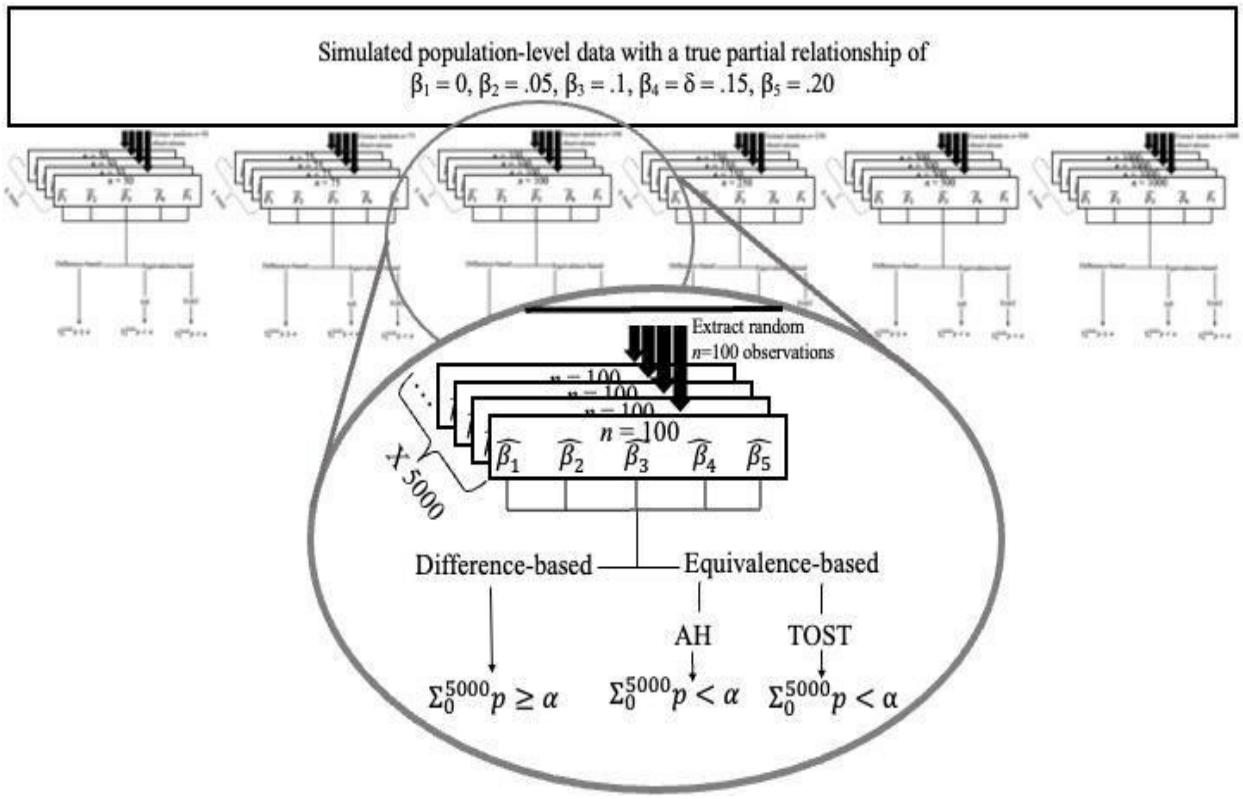
Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indices. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330. https://doi.org/10.1080/10705511.2015.1065414

Supplementary Materials

All materials, including results, code, and slides, are publicly available on the Open Science Framework (OSF): https://osf.io/w96xe/.

Figure S1

Graphical Illustration of the Simulation Procedure



Note. Population-level effects (in the large rectangle) are measured in standardized regression coefficients (β). 5000 multiple regression models of *n* observations were constructed to estimate the true, population-level relationships per sample size level and between-predictor relationship (which are not presented in the Figure). The six models (each of which has a different sample size level, n = 50, 75, 100, 250, 500, 1000) are represented by small rectangular structures. Although not reflected in the Figure, each of the six models is estimated four times, each time with a different association strength between predictor variables (variables were correlated at 0, 0.25, 0.5, and 0.75), for a total of 24 models. Each model was estimated 5000 times and includes five predictors. Each of the five predictor effects ¿) is tested with both difference-based and equivalence-based approaches. The amount of nonsignificant ($p \geq \alpha$) results is counted for difference-based tests, whereas the number of significant results ($p < \alpha$) is counted for equivalence-based tests, two one-sided tests (TOST) and Anderson-Hauck (AH). The number of significant and nonsignificant results are used to compare the statistical performance of both testing approaches.

Simulation Code in R

```
######### PRE-SIMULATION #########

# --------- Loading Packages ---------

library(SimDesign)
library(tidyverse)
library(RColorBrewer)


# --------- Creating local functions (a prerequisite) ---------

# modelstats: estimates regression model + extracts results of difference-
based test

modelstats <- function(dat){

  y <- dat["y"]
  model <- lm(y~., dat)
  modsum <- summary(model)
  b <- model$coefficients[2] #Beta weights estimates extraction
  se <-  modsum$coefficients[2,"Std. Error"] #Standard error extraction per
predictor
  df <- model$df.residual
  t <- modsum$coefficients[2,"t value"]
  p <- modsum$coefficients[2,"Pr(>|t|)"]
  ## traditional difference-based test
  ret <- data.frame(b=b,
                    se=se,
                    df=df,
                    t=t,
                    p=p)
  ret
}


# TOST: performs Schuirmann's Two One-Sided Test on the predictor and provides
the results (the largest p value)
TOST <- function(dat){
  b <- modelstats(dat)$b
  se <- modelstats(dat)$se
  df <- modelstats(dat)$df
  t.value.1 <- (b - l.delta)/se
  t.value.2 <- (b-u.delta)/se
  p.value.1 <-stats::pt(t.value.1, df, lower.tail=FALSE)
  p.value.2 <-stats::pt(t.value.2, df, lower.tail=TRUE)

  ifelse(abs(t.value.1) <= abs(t.value.2), t.value <- t.value.1, t.value <-
t.value.2) # finding the smaller t to present
  ifelse(p.value.1 >= p.value.2, p.value <- p.value.1, p.value <- p.value.2) #
finding the larger p to present
  ret <- data.frame(p=p.value)
  ret
```

```
}

# AH: performs the Anderson-Hauck test on the predictor and provides the p
value
AH <- function(dat){
  b <- modelstats(dat)$b
  se <- modelstats(dat)$se
  df <- modelstats(dat)$df
  t.value <- (b - (l.delta+u.delta)/2)/se
  H.A.del <- ((u.delta-l.delta)/2)/se #this is the delta as defined in Hauck
and Anderson (1986)
  p.value <- stats::pt(abs(t.value)-H.A.del,df) - stats::pt(-abs(t.value)-
H.A.del, df)
  ret <- data.frame(p=p.value)
  ret

}

######### SIMULATION CODE #########

# --------- Fixed parameters ---------

alpha= 0.05
SESOI <- 0.15
l.delta <- -abs(SESOI)
u.delta <- abs(SESOI)
mu <- c(0,0,0,0,0)


# --------- Design stage ---------

Design <- createDesign(N = c(50, 75, 100, 250, 500, 1000),
                       test= c("DB", "TOST", "AH"),
                       beta=c(1, 2, 3, 4, 5),
                       cors= c(0, 0.25, 0.5, 0.75))

#Design

# --------- Generate stage ---------


Generate <- function(condition, fixed_objects = NULL ) {
  Attach(condition)
  sigma <- matrix(data=c(1,cors,cors,cors,cors,
                         cors,1,cors,cors,cors,
                         cors,cors,1,cors,cors,
                         cors, cors, cors, 1, cors,
                         cors, cors, cors, cors, 1),
                  nrow=5,ncol=5)
  xs <- rmvnorm(N, mean = mu, sigma = sigma)
  e <- rnorm(N)
  xs <- as.data.frame(xs)
  y <- 1 + 0*xs$V1+ 0.05*xs$V2 + 0.1*xs$V3 + 0.15*xs$V4 + 0.2*xs$V5 + e
  dat <- data.frame(xs,y)
  dat
}
```

```
# --------- Analyse stage ---------

Analyse <- function(condition, dat, fixed_objects = NULL) {
  Attach(condition)
  if(test=="DB"){
    p<- modelstats(dat, beta)$p
  }
  if(test=="TOST"){
    p <- TOST(dat, beta)$p
  }
  if(test=="AH"){
    p <- AH(dat, beta)$p
  }
  ret <- c(p=p)
  ret
}

# --------- Summarise stage ---------

Summarise <- function(condition, results, fixed_objects = NULL) {
  Attach(condition)
  ifelse(test=="DB",neg <- 1 - EDR(results, alpha=alpha), neg <- EDR(results,
alpha=alpha))
  ret <- c(concluding_negligible=neg)
  ret
}

# --------- Run stage ---------

res <- runSimulation(design=Design, replications=5000, generate=Generate,
                     analyse=Analyse, summarise=Summarise)

######### SAVING RESULTS #########

write.csv(res, "Alter_Counsell_Simulation_Results.csv")


######### VISUALIZING RESULTS #########

simresults <- res
simresults["beta"][simresults["beta"] == 1] <- 0
simresults["beta"][simresults["beta"] == 2] <- 0.05
simresults["beta"][simresults["beta"] == 3] <- 0.1
simresults["beta"][simresults["beta"] == 4] <- 0.15
simresults["beta"][simresults["beta"] == 5] <- 0.2
simresults$cors <- factor(simresults$cors)
simresults$label <- paste("β =", as.character(simresults$beta))

# --------- Figure 2 ---------

simresults |>
  filter(beta == 0.15 | beta==0.2 ) |>
  ggplot( aes(x = factor(N), y = concluding_negligible.p,
              group= interaction(test, cors),
              colour= test, linetype = cors))+
  geom_line(linewidth=1.5, alpha=0.8)+
```

```
  scale_linetype_manual(values = c("solid", "longdash", "dotdash", "dotted"))+
  facet_wrap(~label)+
  theme_minimal()+
  theme(axis.text = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        axis.title.x = element_text(size = 18),
        axis.title.y = element_text(size = 18),
        strip.text = element_text(size = 18),
        text=element_text(family="Times"))+
  labs(color = "Test", linetype="Correlations")+
  scale_colour_brewer(palette = "Dark2")+
  scale_y_continuous(breaks=seq(0,1,0.1))+
  geom_hline(yintercept=0.05, linetype='dotted', col = 'red')+
  labs( y="Rate of Incorrect Negligible Association Conclusions", x = "Sample
Size")+ #title = "Incorrectly Concluding Negligible Association by Test,
Effect, Correlation, and Sample Size",
  annotate("text",x="50" , y = 0.05, label = "\u03B1 = .05",
vjust=-.7,hjust=.55,  family="Times",  size= 5)
#ggsave("Incorrect_in_colour.png", width = 20, height=15, units = "cm")

# --------- Figure 3 ---------

simresults |>
  filter(beta == 0 | beta==0.05 | beta == 0.1) |>
  ggplot( aes(x = factor(N), y = concluding_negligible.p,
              group= interaction(test, cors),
              colour= test,  linetype = cors))+
  geom_line(linewidth=1.5, alpha=0.8)+
  scale_linetype_manual(values = c("solid", "longdash", "dotdash", "dotted"))+
  facet_wrap(~label)+
  theme_minimal()+
  theme(axis.text = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        axis.title.x = element_text(size = 18),
        axis.title.y = element_text(size = 18),
        strip.text = element_text(size = 18),
        text=element_text(family="Times"))+
  labs(color = "Test", linetype="Correlations")+
  scale_colour_brewer(palette = "Dark2")+
  scale_y_continuous(breaks=seq(0,1,0.1))+
  geom_hline(yintercept=0.8, linetype='dotted', col = 'blue')+
  labs(y="Rate of Correct Negligible Association Conclusions", x = "Sample
Size")+ #title = "Correctly Concluding Negligible Association by Test, Effect,
Correlation, and Sample Size"
  annotate("text",x="50" , y = 0.8, label = "1-\u03B2 = .80",
vjust=-.7,hjust=.2, family="Times", size= 5)
#ggsave("Correct_in_colour.png", width = 20, height=15, units = "cm")

######### END #########
```

Example 1 Using Standardized Units

We can conduct the same tests using standardized effects instead. The standardized regression coefficient estimate for MW-D from the same model can be found in Seli at el. (2017), $\beta_{MW\text{-}D}$ = -.008, and the standardized SESOI can be converted from its raw form (0.23) mathematically:

$$\Delta = \frac{\delta * SD_x}{SD_y} \quad \text{(S1)}$$

where Δ is the standardized form of SESOI, δ is the unstandardized SESOI, and SD_x and SD_y are the standard deviations for the predictor and outcome variables, respectively. Using our predefined unstandardized SESOI of 0.23, we obtain a standardized SESOI of .10. To use this standardized value in the `neg.reg` function, we can use the exact same function input, with adjusting only the SESOI to .1 and setting the argument `std = TRUE`. Accordingly, a researcher would enter:

```
library(negligible)
neg.reg(data = ocd, # name of dataset
        formula = ocd_cont~MWD+MWS, # regression formula
        predictor = MWD, # name of the predictor of interest
        eiu = 0.1, # upper bound of SESOI (standardized)
        eil = -0.1, # lower bound of SESOI (standardized)
        std = TRUE, # using unstandardized units
        bootstrap = FALSE) # not using bootstrap in example
```

Importantly, negligible association testing results using standardized units of the regression coefficient and SESOI provide identical inferential statistics and conclusions as the tests conducted on the unstandardized units.

More About Using Standardized or Unstandardized Effects

Using unstandardized regression coefficients as an effect size is both strongly recommended (e.g., Pek and Flora, 2018) and most commonly implemented in psychology (Farmus et al., 2022). Unstandardized regression coefficients and their associated standard errors are also usually the default output from most statistical software and therefore the easiest to retrieve. However, researchers might be more inclined to define the SESOI in standardized units due to the intuitive judgment of the effect's magnitude and uniformity across previous studies or meta-analyses. If researchers are interested in using standardized SESOI units and are inputting a dataset into the `neg.reg` function, they only need to specify `std = TRUE` as one of the arguments in the function. However, if no dataset is fed into the function, users can rearrange Equation S1 to

$$\delta = \frac{\Delta * SD_y}{SD_x} \tag{S2}$$

to convert their desired standardized SESOI to unstandardized units. Then users should plug in the newly calculated SESOI (now in unstandardized units) into the function and specify `std = FALSE`.

It is important to note, however, that the process of converting the units from one form to another arithmetically may yield slightly higher rates of incorrectly concluding negligible effect than expected (Campbell, 2022). Although this difference in error rate is minor, it should be acknowledged. And, more importantly, the SESOI we select should (ideally) be independent of the sample characteristics. By converting the SESOI from unstandardized to standardized or vice versa, we introduce some of the sample's characteristics into the hypotheses (which contains the SESOI) because the conversion equation includes the standard deviations of *X* and *Y*. Instead, if

researchers opt for running the test in standardized units, they should define their SESOI *originally* in standardized units.

References

Abramowitz, J. S., Deacon, B. J., Olatunji, B. O., Wheaton, M. G., Berman, N. C., Losardo, D., Timpano, K. R., McGrath, P. B., Riemann, B. C., Adams, T., Björgvinsson, T., Storch, E. A., & Hale, L. R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the Dimensional Obsessive-Compulsive Scale. *Psychological Assessment, 22*(1), 180–198. https://doi.org/10.1037/a0018260

Altman, D., & Bland, J. (1995). Absence of Evidence Is Not Evidence of Absence. *British Medical Journal. 311*. 485. https://doi.org/10.1136/bmj.311.7003.485

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician, 73*(sup1), 262-270. https://doi.org/10.1080/00031305.2018.1543137

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

Anderson, S. F, & Maxwell, S. E. (2017) Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research, 52*(3), 305-324. https://doi.org/10.1080/00273171.2017.1289361

Anvari, & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96*, 104159–. https://doi.org/10.1016/j.jesp.2021.104159

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory–II (BDI-II). APA PsycTests. https://doi.org/10.1037/t00742-000

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science, 11* ,283-319. https://doi.org/10.1214/ss/1032280304

Beribisky, N., Mara, C. A., Cribbie, R. A. (2020) An Equivalence Testing Approach for Evaluating Substantial Mediation. *The Quantitative Methods for Psychology, 16*(5), 424-441. https://doi.org/10.20982/tqmp.16.4.p424

Berkson, J. (1938) Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. *Journal of the American Statistical Association, 33*(203), 526-536. https://doi.org/10.1080/01621459.1938.10502329

Brown, L. D., Hwang, J. T. G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics, 25* ,2345-2367. https://doi.org/10.1214/aos/1030741076

Caldwell, A. R. (2022, November 17). Exploring Equivalence Testing with the Updated TOSTER R Package. https://doi.org/10.31234/osf.io/ty8de

Campbell, H. (2020). Equivalence testing for standardized effect sizes in linear regression. *arXiv*. https://doi.org/10.48550/arXiv.2004.01757

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., & Lakens, D. (2021). Can we disregard the whole model? omnibus non-inferiority testing for R^2 in multi-variable linear regression and η^2 in ANOVA. *British Journal of Mathematical & Statistical Psychology, 74*(1), 64-89. https://doi.org/10.1111/bmsp.12201

Carriere, J. S. A., Seli, P., & Smilek, D. (2013). Wandering in Both Mind and Body: Individual Differences in Mind Wandering and Inattention Predict Fidgeting. *Canadian Journal of Experimental Psychology, 67*(1), 19-31. https://doi.org/10.1037/a0031438

Cohen, J. (1990) The Things I Have Learned (So Far). *American Psychologist*. 45, 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Counsell, A. & Cribbie, R. A. (2016). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology, 68*: 292-309. https://doi.org/10.1111/bmsp.12045

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020) Evaluating Equivalence Testing Methods for Measurement Invariance. *Multivariate Behavioral Research, 55*(2), 312-328. https://doi.org/10.1080/10705511.2015.1065414

Cribbie, R. A., Alter, U., Beribisky, N., Chalmers, R. P., Counsell., A., Farmus, L., Martinez Gutierrez, N., Ng, V. (2022). *negligible: A Collection of Functions for Negligible Effect/Equivalence Testing*. R package version 0.1.2. https://cran.r-project.org/web/packages/negligible/index.html

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

Cribbie, R. A., Panzarella, E., Farmus, L., Martinez Gutierrez, N., Beribisky, N., & Alter, U. (2020, February 24). *Effect Size Reporting in Social-Personality Research: The Good, the*

Bad, and the Ugly [Presentation]. Quantitative Methods Area Forum, York University, Toronto, ON, Canada.

Cribbie, R.A., Ragoonanan, C. and Counsell, A. (2016), Testing for negligible interaction: A coherent and robust approach. *Br J Math Stat Psychol, 69*: 159-174. https://doi.org/10.1111/bmsp.12066

Crutzen, R., & Peters, G. J. Y. (2017). Targeting next generations to change the common practice of underpowered research. Frontiers in psychology, 8, 1184. https://doi.org/10.3389/fpsyg.2017.01184

Cumming, G. (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York, NY: Taylor & Francis Group, LLC

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 x 2 tables. *Biometrics, 33*(4), 593-602. https://doi.org/10.2307/2529457

Farmus, L., Beribisky, N., Martinez Gutierrez, N., Alter, U., Panzarella, E., & Cribbie, R. A. (2022). Effect size reporting and interpretation in social personality research. Current Psychology. https://doi.org/10.1007/s12144-021-02621-7

Farruggia, S. P., Chen, C., Greenberger, E., Dmitrieva, J., & Macek, P. (2004). Adolescent self-esteem in cross-cultural perspective: Testing measurement equivalence and a mediation model. *Journal of Cross-Cultural Psychology, 35*(6), 719-733. https://doi.org/10.1177/0022022104270114

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Zeitschrift Für Psychologie/Journal of Psychology, 217*(1), 27-37. https://doi.org/10.1027/0044-3409.217.1.27

Fletcher, T. D. (2012). QuantPsyc: Quantitative Psychology Tools. R package version 1.5. https://CRAN.R-project.org/package=QuantPsyc

Flora, D. B. (2018). *Statistical Methods for the Social and Behavioural Sciences: A Model-Based Approach*. London, UK. SAGE Publications Ltd.

Food and Drug Administration. (2021). Approved drug products with therapeutic equivalence evaluations (41st ed). U.S Department of Health and Human Services. Retrieved from http://www.fda.gov.

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. https://doi.org/10.1348/000711009X475853

Harlow, L.L. (1997). Significance Testing in Introduction and Overview. In L.L. Harlow, S.A. Muliak & J.H. Steiger (Eds.). *What If There Were No Significance Tests?* (pp.1-17). Mahwah, NJ, USA: Lawrence Erlbaum.

Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics, 12*(1), 83-91. https://doi.org/10.1007/BF01063612

Jabbari, Y., & Cribbie, R. (2021). Negligible interaction test for continuous predictors. *Journal of applied statistics, 49*(8), 2001–2015. https://doi.org/10.1080/02664763.2021.1887102

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. https://doi.org/10.1177/0956797611430953

Kim, Y. J., & Cribbie, R. A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology, 71*(1), 1-12. https://doi.org/10.1111/bmsp.12103

Kirk, R. E. (2003). *The importance of effect magnitude*. In S. F. Davis (Ed.), Handbook of research methods in experimental psychology (pp. 83–105). Malden, MA: Blackwell.

Kupats, E., Noviks, I., Vrublevska, J., Kenina, V., Kojalo, U., & Logina, I. (2018). No relationship between generalized anxiety symptoms and cardiovascular autonomic dysfunction. *Neurology, Psychiatry and Brain Research, 30*, 86-90. https://doi.org/10.1016/j.npbr.2018.07.001

Lakens, D. (2022, June 21). *Using p-values to test a hypothesis*. Improving Your Statistical Inferences. https://lakens.github.io/statistical_inferences/pvalue.html

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science, 8*(4), 355–362. https://doi.org/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

LeBeau, B. (2019). SIMGLM: Simulate Models Based on the Generalized Linear Model. R package version 0.7.4. http://CRAN.R-project.org/package=simglm

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology, 69*(6), 555-562. https://doi.org/10.4097/kjae.2016.69.6.555

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61*(2), 107-116. https://doi.org/10.1080/00049530802105865

Martinez Gutierrez, N., & Cribbie, R. A. (2022). Effect sizes for equivalence testing: incorporating the equivalence interval [Manuscript submitted for publication]. Department of psychology, York University.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487–498. https://doi.org/10.1037/a0039400

Metzler, C. M. (1974). Bioavailability - A problem in equivalence. *Biometrics, 30*(2), 309-317. https://doi.org/10.2307/2529651

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208-225. https://doi.org/10.1037/met0000126

Proudfoot, D., & Kay, A. C. (2018). How perceptions of one's organization can affect perceptions of the self: Membership in a stable organization can sustain individuals' sense of control. *Journal of Experimental Social Psychology, 76*, 104-115. https://doi.org/10.1016/j.jesp.2018.01.004

Quintana, D. S. (2016). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology, 54*: 344-349. https://doi.org/10.1111/psyp.12798

Quintana, D. S. (2018, July 21). Using summary statistics to determine whether a non-significant result supports the absence of an effect [Blog post]. Retrieved from https://blog.usejournal.com/using-summary-statistics-to-determine-whether-a-non-significant-result-supports-the-absence-of-an-1ff61e97f7cf

R Core Team (2021). R: A Language and Environment for Statistical Computing. R foundation for statistical computing, Vienna, Austria. URL http://CRAN.R-project.org

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Robinson, A. P., Duursma, R. A., & Marshall, J. D. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree Physiology, 25*(7), 903-913. https://doi.org/10.1093/treephys/25.7.903

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565. https://doi.org/10.1037/0033-2909.113.3.553

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* ,657-680. https://doi.org/10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3* ,403-411. https://doi.org/10.1037/1082-989X.3.4.403

Seli, P., Risko, E. F., Purdon, C., & Smilek, D. (2017). Intrusive thoughts: Linking spontaneous mind wandering and OCD symptomatology. *Psychological Research, 81*(2), 392-398. https://doi.org/10.1007/s00426-016-0756-3

Steptoe, A., Kunz-Ebrecht, S. R., & Owen, N. (2003). Lack of association between depressive symptoms and markers of immune and vascular inflammation in middle-aged men and women. *Psychological Medicine, 33*(4), 667-674. https://doi.org/10.1017/S0033291702007250

Thompson, B. (1992), Two and One-Half Decades of Leadership in Measurement and Evaluation. *Journal of Counseling & Development, 70*: 434-438. https://doi.org/10.1002/j.1556-6676.1992.tb01631.x

Tukey, J.W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science, 6*, 100-116. https://doi.org/10.1214/ss/1177011945

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61*(8), 1340-1341. https://doi.org/10.1002/jps.2600610845

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32*(4), 741-744. https://doi.org/10.2307/2529259

Westlake, W. J. (1981). Response to T. B. L. Kirkwood: Bioequivalence testing – a need to rethink. *Biometrics, 37*, 589–594. https://doi.org/10.2307/2530573

Wheadon, D. E., Rampey, A. H., Thompson, V. L., Potvin, J. H., Masica, D. N., & Beasley, C. M. (1992). Lack of association between fluoxetine and suicidality in bulimia nervosa. *Journal of Clinical Psychiatry, 53*, 235–241. https://doi.org/10.1016/j.jpsychires.2013.05.025

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405–426. https://doi.org/10.1037/met0000080

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indices. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 319–330. https://doi.org/10.1080/10705511.2015.1065414