RUNNING HEAD: Equivalence Tests

Equivalence Tests:

A Practical Primer for *t*-Tests, Correlations, and Meta-Analyses.

Daniël Lakens

Eindhoven University of Technology

Words: 4999

Correspondence can be addressed to Daniël Lakens, Human Technology Interaction Group, IPO 1.24, PO Box 513, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl.

Abstract

Scientists should be able to provide support for the absence of a meaningful effect. Currently researchers often incorrectly conclude an effect is absent based a non-significant result. A widely recommended approach within a Frequentist framework is to test for *equivalence*. In equivalence tests, such as the Two One-Sided Tests (TOST) procedure discussed in this article, an upper and lower equivalence bound is specified based on the smallest effect size of interest. The TOST procedure can be used to statistically reject the presence of effects large enough to be considered worthwhile. This practical primer with accompanying spreadsheet and R package enables psychologists to easily perform equivalence tests (and power analyses) by setting equivalence bounds based on standardized effect sizes, and provides recommendations to pre-specify equivalence bounds. Extending your statistical toolkit with equivalence tests might very well be the easiest way for psychologists to improve their statistical and theoretical inferences.

Author Note: The TOSTER spreadsheet and supplementary material is available from https://osf.io/q253c/. The TOSTER R package can be installed in R using: library(devtools); install_github("Lakens/TOSTER") and is available from https://github.com/Lakens/TOSTER.

Equivalence Tests:

A Practical Primer for *t*-Tests, Correlations, and Meta-Analyses.

Scientists should be able to provide support for the null-hypothesis. A limitation of the widespread use of traditional significance tests, where the null hypothesis is that the true effect size is zero, is that the absence of an effect can be rejected, but not statistically supported. When you perform a statistical test, and the outcome is a *p*-value larger than the alpha level α (e.g., $p > 0.05$), the only formally correct conclusion is that the data are not surprising, assuming the null hypothesis is true. It is not possible to conclude there is no effect when $p > \alpha$ – our test might simply have lacked the statistical power to detect a true effect.

It is statistically impossible to support the hypothesis that a true effect size is exactly zero. What *is* possible in a Frequentist hypothesis testing framework is to statistically reject effects large enough to be deemed worthwhile. When researchers want to argue for the absence of an effect that is large enough to be worthwhile to examine, they can test for *equivalence* (Wellek, 2010). By rejecting an effect (indicated in this article by Δ) more extreme than pre-determined lower and upper equivalence bounds ($-\Delta_L$ and Δ_U, for example effect sizes of Cohen's $d = -0.3$ and $d = 0.3$), we can act as if the true effect is close enough to zero for our practical purposes. Equivalence testing originates from the field of pharmacokinetics (Hauck & Anderson, 1984), where researchers sometimes want to show that a new cheaper drug works just as well as an existing drug (for an overview, see Senn, 2007, chapters 15 and 22). A very simple equivalence testing approach is the 'two-one-sided *t*-tests' (TOST) procedure (Schuirmann, 1987). In the TOST procedure an upper (Δ_U) and lower ($-\Delta_L$) equivalence bound is specified based on the smallest effect size of interest (e.g., a positive or negative difference of $d = 0.3$). Two composite null hypotheses are tested: $H0_1$: $\Delta \leq -\Delta_L$ and $H0_2$: $\Delta \geq \Delta_U$. When both these one-sided tests can be statistically rejected, we can conclude that $-\Delta_L < \Delta < \Delta_U$, or that the observed effect falls within the equivalence bounds and is close enough to zero to be practically equivalent (Seaman & Serlin, 1998).

Psychologists often incorrectly conclude there is no effect based on a non-significant test result. For example, the words "no effect" had been used in 108 articles published in SPPS up to August 2016. Manual inspection revealed that in almost all of these articles, the conclusion of 'no

effect' was based on statistical non-significance. Finch, Cumming, and Thomason (2001) reported that in the Journal of Applied Psychology a stable average of around 38% of articles with non-significant results accept the null hypothesis in previous years. This practice is problematic. With small sample sizes, non-significant test results are hardly indicative of the absence of a true effect, and with huge sample sizes, effects can be statistically significant, but practically and theoretically irrelevant. Equivalence tests, which are conceptually straightforward, easy to perform, and highly similar to widely used hypothesis significance tests that aim to reject a null-effect, are a straightforward but underused approach to reject the possibility that an effect more extreme than the smallest effect size of interest exists (Anderson & Maxwell, 2016).

Psychologists would gain a lot by embracing equivalence tests. First, researchers often incorrectly use non-significance to claim the absence of an effect (e.g., "there were no gender effects, $p > .10$"). This incorrect interpretation of *p*-values would be more easily recognized and should become less common in the scientific literature if equivalence tests were better known and more widely used. Second, where traditional significance test only allows researchers to reject the null hypothesis, science needs statistical approaches that allow us to conclude meaningful effects are absent (Dienes, 2016). Finally, the strong reliance on hypothesis significance tests that merely aim to reject a null-effect does not require researchers to think about the effect size under the alternative hypothesis. Exclusively focusing on rejecting a null-effect has been argued to lead to imprecise hypotheses (Gigerenzer, 1998). Equivalence testing invites researchers to make more specific predictions about the effect size they find worthwhile to examine.

There have been previous attempts to introduce equivalence testing to psychology (Quertemont, 2011; Rogers, Howard, & Vessey, 1993; Seaman & Serlin, 1998). I believe there are four reasons why previous attempts have largely failed. First, there is a lack of easily accessible software to perform equivalence tests. To solve this problem, I've created an easy to use spreadsheet and R package to perform equivalence tests for independent and dependent *t*-tests, correlations, and meta-analyses (see https://osf.io/q253c/). These tests can be performed based on summary statistics, which researchers in my experience find convenient (Lakens, 2013). Second, in pharmacokinetics the equivalence bounds are often defined in raw scores, whereas it might be more intuitive for researchers

in psychology to express equivalence bounds in standardized effect sizes. This makes it easier to perform power analyses for equivalence tests (which can also be done with the accompanying spreadsheet and R package), and to compare equivalence bounds across studies in which different measures are used. Third, there is no single article that discusses both power analyses and statistical tests for one-sample, dependent and independent *t*-tests, correlations, and meta-analyses, which are all common in psychology. Finally, guidance on how to set equivalence boundaries has been absent for psychologists, given that there are often no specific theoretical limitations on how small effects are predicted to be (Morey & Lakens, under review), nor cost-benefit boundaries of when effects are too small to be practically meaningful. This is a chicken-egg problem, since using equivalence tests will likely stimulate researchers to specify which effect sizes are predicted by a theory (Weber & Popova, 2012). To bootstrap the specification of equivalence bounds in psychology, I propose that when theoretical or practical boundaries on meaningful effect sizes are absent, researchers set the bounds to the smallest effect size they have sufficient power to detect, which is determined by the resources they have available to study an effect.

Testing for Equivalence

In this article, I will focus on the TOST procedure (Schuirmann, 1987) of testing for equivalence, because of its simplicity and widespread use in other scientific disciplines. The goal in the TOST approach is to specify a lower and upper bound, such that results falling within this range are deemed equivalent to the absence of an effect that is worthwhile to examine (e.g., Δ_L = -0.3 to Δ_U = 0.3, where Δ is a difference that can be defined by either standardized differences such as Cohen's d, or raw differences such as 0.3 scale point on a 5-point scale). In the TOST procedure the null hypothesis is the *presence* of a true effect of Δ_L or Δ_U, and the alternative hypothesis is an effect that falls within the equivalence bounds, or the *absence* of an effect that is worthwhile to examine. The observed data is compared against Δ_L and Δ_U in two one-sided tests. If the *p*-value for both tests indicates the observed data is surprising, assuming $-\Delta_L$ or Δ_U are true, we can follow a Neyman-Pearson approach to statistical inferences and reject effect sizes larger than the equivalence bounds. When making such a statement, we will not be wrong more often, in the long run, than our Type 1 error rate (e.g., 5%). It is also possible to test for inferiority, or the hypothesis that the effect is smaller

than an upper equivalence bound, by setting the lower equivalence bound to ∞.[1] Furthermore, equivalence bounds can be symmetric around zero (Δ_L = -0.3 to Δ_U = 0.3) or asymmetric (Δ_L = -0.2 to Δ_U = 0.4).

When NHST and equivalence tests are both used, there are four possible outcomes of a study: The effect can be significant (statistically different from zero), equivalent (statistically larger than Δ_L and smaller than Δ_U), significant *and* equivalent, or undetermined (neither statistically significant, nor statistically equivalent). In Figure 1, mean differences (black squares) and their 90% (thick lines) and 95% confidence intervals (thin lines) are illustrated for four scenarios. To conclude equivalence (scenario A), the 90% confidence interval around the observed mean difference should exclude the Δ_L and Δ_U values of -0.5 and 0.5 (indicated by black vertical dashed lines)[2].

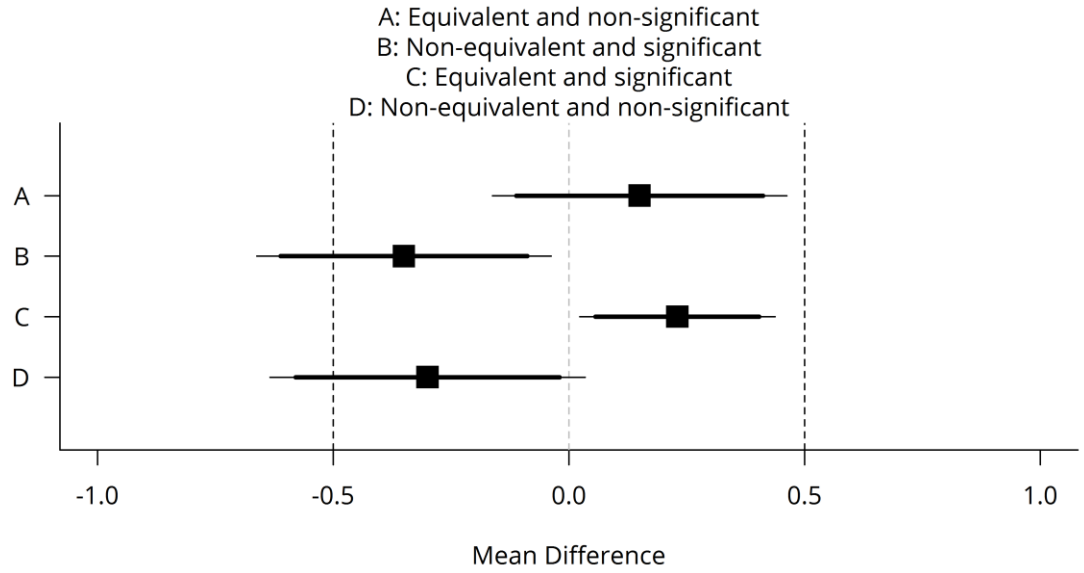

Figure 1. Mean differences (black squares) and 90% confidence intervals (thick horizontal lines) and 95% confidence intervals (thin horizontal lines) with equivalence bounds Δ_L = -0.5 and Δ_U= 0.5 for equivalent, significant, significant and equivalent, and non-significant and non-equivalent test results.

The traditional two-sided null hypothesis significance test is rejected (scenario B) when the confidence interval around the mean difference does not include 0 (the vertical grey dotted line). Effects can be significant *and* equivalent (scenario C) when the 90% confidence interval excluded the

equivalence bounds, and the 95% confidence interval excluded zero. Finally, an effect can be undetermined, or non-significant and non-equivalent (scenario D) when the 90% confidence interval includes one of the equivalence bounds, and the 95% confidence interval includes zero.

In this article, the focus lies on the TOST procedure, where two *p*-values are calculated. Readers are free to replace decisions based on *p*-values by decisions based on 90% confidence intervals if they wish. Formally, hypothesis testing and estimation are distinct approaches (Cumming & Finch, 2001). For example, while sample size planning based on confidence intervals focusses on the width of confidence intervals, sample size planning for hypothesis testing uses power analysis to estimate the probability of observing a significant result (Maxwell, Kelley, & Rausch, 2008). Since the TOST procedure is based on a Neyman-Pearson hypothesis testing approach to statistics, and I'll explain how to calculate the tests, as well as how to perform power analysis, I'll focus on the calculation of *p*-values for conceptual consistency.

Equivalence tests for differences between two independent means

The TOST procedure entails performing two one-sided tests to examine whether the observed data is surprisingly larger than a lower equivalence boundary (Δ_L), or surprisingly smaller than an upper equivalence boundary (Δ_U). The equivalence test assuming equal variances is based on:

$$t_L = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_L}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ and } t_U = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_U}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad (1)$$

where M_1 and M_2 indicate the means of each sample, n_1 and n_2 are the sample size in each group, and σ is the pooled standard deviation:

$$\sigma = \sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}} \quad (2)$$

Even though Student's *t*-test is by far the most popular *t*-test in psychology, there is general agreement that whenever the number of observations are unequal across both conditions Welch's *t*-test (1938), which does not rely on the assumption of equal variances, should be performed by default (Delacre, Lakens, & Leys, 2016; Ruxton, 2006). The equivalence test not assuming equal variances is based on:

$$t_L = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_L}{\sqrt{\frac{SD_1^2}{n_1}+\frac{SD_2^2}{n_2}}} \text{ and } t_U = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_U}{\sqrt{\frac{SD_1^2}{n_1}+\frac{SD_2^2}{n_2}}} \quad (3)$$

where the degrees of freedom for Welch's t-test are based on the Sattherthwaite (1946) correction:

$$df_w = \frac{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)}{\frac{\left(SD_1^2/n_1\right)^2}{n_1-1} + \frac{\left(SD_2^2/n_2\right)^2}{n_2-1}} \quad (4)$$

These formulas are highly similar to the Student's and Welch's *t*-statistic for traditional significance tests. The only difference is that the lower equivalence bound Δ_L and the upper equivalence bound Δ_U are subtracted from the mean difference between groups. These bounds can be defined in raw scores or in a standardized difference, where Δ = Cohen's d×σ, or Cohen's d = Δ/σ. The two one-sided tests are rejected if $t_U \leq -t_{(df,\, \alpha)}$, and $t_L \geq t_{(df,\, \alpha)}$, where $t_{(\alpha,\, df)}$ is the upper 100α percentile of a *t* distribution (Berger & Hsu, 1996). The spreadsheet and R package can be used to perform this test, but some commercial software such as Minitab also includes the option to perform equivalence tests for *t*-tests.

As an example, Eskine (2013) showed that participants who had been exposed to organic food were substantially harsher in their moral judgments relative to those in the control condition (*d* = 0.81, 95% CI [0.19, 1.45]). A replication by Moery and Calin-Jageman, (2016, Study 2) did not observe a significant effect (Control: n = 95, M = 5.25, SD = 0.95, Organic Food: n = 89, M = 5.22, SD = 0.83). The authors followed Simonsohn's (2015) recommendation so set the equivalence bound to the effect size the original study had 33% power to detect. With n = 21 in each condition of the original study, this means the equivalence bound is *d* = 0.48, which equals a difference of 0.384 on a 7-point scale given the sample sizes and a pooled standard deviation of 0.894). We can calculate the TOST equivalence test *t*-values:

$$\frac{5.25-5.22-(-0.384)}{0.894\sqrt{\frac{1}{95}+\frac{1}{89}}} = t_L = 3.14, \text{ and } \frac{5.25-5.22-0.384}{0.894\sqrt{\frac{1}{95}+\frac{1}{89}}} = t_U = -2.69$$

which correspond to *p*-values of 0.001 and 0.004. If alpha = 0.05, and assuming equal variances, the equivalence test is significant, *t*(182) = -2.69, *p* = 0.004. We can reject effects larger than 0.384 scale points. Note that both one-sided tests need to be significant to declare equivalence, but for efficiency only the one-sided test with the highest *p*-value is reported in TOST results (given that if this test is significant, so is the other). Alternatively, because Moery and Calin-Jageman's (2016) main prediction seems to be whether the effect smaller than the upper equivalence bound (a

test for inferiority) only the one-sided *t*-test against the upper equivalence bound could be performed and reported. Note that the spreadsheet and R package allow you to either directly specify the equivalence bounds in Cohen's *d*, or set the equivalence bound in raw units.

An a-priori power analysis for equivalence tests can be performed by calculating the required sample sizes to declare equivalence for two one-sided tests based on the lower equivalence bound and upper equivalence bound. When equivalence bounds are symmetric around zero (e.g., Δ_L = -0.5 and Δ_U = 0.5) the required sample sizes (referred to as n_L and n_U in Formula 5 below) will be identical. Following Chow, Shao, and Wang (2002) the normal approximation of the power formula for equivalence tests (for each independent group of an independent *t*-test) given a specific α level and desired level of statistical power (1-β) is:

$$n_L = \frac{2(z_\alpha + z_{\beta/2})^2}{{\Delta_L}^2}, n_U = \frac{2(z_\alpha + z_{\beta/2})^2}{{\Delta_U}^2} \quad (5)$$

where Δ_L and Δ_U are the standardized mean difference equivalence bounds (in Cohen's *d*). This formula calculates the required sample sizes based on the assumption that the true effect size is zero (see Table 1). If a non-zero true effect size is expected, an iterative procedure must be used. An excellent and highly accessible overview of power analysis for equivalence, superiority, and non-inferiority designs, with power tables for a wide range of standardized mean differences and expected true mean differences that can be used to decide upon the sample size in your study is available from Julious (2004).

Table 1. *Sample sizes (for the number of observations in each group) for equivalence tests for independent means, as a function of the desired power, alpha level, and equivalence bound* Δ (in Cohen's d), *based on exact calculations and the approximation.*

	Approximation				**Exact**			
Bound	**80% power**		**90% power**		**80% power**		**90% power**	
(Δ)	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**
0.1	1713	2604	2165	3155	1713	2604	2165	3155
0.2	429	651	542	789	429	652	542	789
0.3	191	290	241	351	191	291	242	351
0.4	108	163	136	198	108	165	136	199
0.5	69	105	87	127	70	106	88	128
0.6	48	73	61	88	49	74	61	89
0.7	35	54	45	65	36	55	45	66
0.8	27	41	34	50	28	43	35	51

The narrower the equivalence bounds, or the smaller the effect sizes one tries to reject, the larger the sample size that is required. Large sample sizes are required to achieve high power when equivalence bounds are close to zero. This is comparable to the large sample sizes that are required to reject a true but small effect when the null hypothesis is a null-effect. Equivalence tests require slightly larger sample sizes than traditional null hypothesis tests. Because two consecutive one-sided tests are performed in a row and both should be statistically significant, each individual test must have higher power for two tests in a row to have the desired power (Senn, 2007, p. 242). For example, when each test has 0.89 power, two tests in a row have 0.89 × 0.89 = 0.8 power.

Equivalence tests for differences between dependent means

When comparing dependent means, the correlation between the observations has to be taken into account, and the effect size directly related to the statistical significance of the test (and thus used in power analysis) is Cohen's d_z (see Lakens, 2013). The *t*-values for the two one-sided tests statistics are:

$$t_L = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_L}{\frac{\sqrt{SD_1^2 + SD_2^2 - 2 \times r \times SD_1 \times SD_2}}{\sqrt{N}}} \text{ and } t_U = \frac{\bar{M}_1 - \bar{M}_2 - \Delta_U}{\frac{\sqrt{SD_1^2 + SD_2^2 - 2 \times r \times SD_1 \times SD_2}}{\sqrt{N}}} \quad (6)$$

The bounds Δ_L and Δ_U can be defined in raw scores, or in a standardized bound based on Cohen's d_z, where $\Delta = d_z \times SD_{diff}$, or $d_z = \Delta / SD_{diff}$. Formula 3 can be used for a-priori power analyses by inserting Cohen's d_z instead of Cohen's *d*. The number of pairs needed to achieve a desired level of power when using Cohen's d_z is half the number of observations needed in each between subject condition specified in Table 1.

There are no suggested benchmarks of small, medium, and large effects for Cohen's d_z. We can consider two approaches to determining benchmarks. The first is to use the same benchmarks for Cohen's *d* as for Cohen's d_z. This simply ignores the correlation between dependent variables (or assumes $r = 0.5$, when Cohen's *d* and Cohen's d_z are identical)[3]. A second approach is to scale the benchmarks for Cohen's d_z based on the sample size we need reliably detect an effect. For example, in an independent *t*-test, 176 participants are required in each condition to achieve 80% power for d = 0.3 and $\alpha = 0.05$. With 176 pairs of observations and $\alpha = 0.05$, a study has 80% power for a Cohens' d_z of 0.212. The relationship between *d* and d_z is simply a factor of $\sqrt{2}$, which means we can translate the benchmarks for Cohen's *d* for small (0.2), medium (0.5) and large (0.8) into benchmarks for Cohen's d_z of small (0.14), medium (0.35) and large (0.57). There is no objectively correct way to set benchmarks for Cohen's *dz*, and I leave it up to the reader to determine whether either of these approaches is useful.

Equivalence tests for one-sample *t*-tests

The *t*-values for the two one-sided tests for a one-sample *t*-tests are:

$$t_L = \frac{M - \mu - \Delta_L}{\frac{SD}{\sqrt{N}}} \text{ and } t_U = \frac{M - \mu - \Delta_U}{\frac{SD}{\sqrt{N}}} \quad (7)$$

where *M* is the observed mean, *SD* is the observed standard deviation, *N* is the sample size, Δ_L and Δ_U are lower and upper equivalence bounds, and μ is the value that the mean is tested against.

Equivalence tests for correlations

Equivalence tests can also be performed on correlations, where the two one-sided tests aim to reject correlations larger than a lower equivalence bound (r_L) and smaller than an upper equivalence

bound (r_U). I follow Goertzen and Cribbie (2010), who use Fisher's *z* transformation on the correlations, after which critical values are calculated that can be compared against the normal distribution:

$$Z_L = \frac{\frac{LN\left(\frac{1+r}{1-r}\right)}{2} - \frac{LN\left(\frac{1+r_L}{1-r_L}\right)}{2}}{\frac{1}{\sqrt{N-3}}}, Z_U = \frac{\frac{LN\left(\frac{1+r}{1-r}\right)}{2} - \frac{LN\left(\frac{1+r_U}{1-r_U}\right)}{2}}{\frac{1}{\sqrt{N-3}}} \quad (8)$$

The two one-sided tests are rejected if $Z_L \leq -Z_\alpha$, and $Z_U \geq Z_\alpha$. Benchmarks for small, medium, and large effects, which can be used to set equivalence bounds, are $r = 0.1$, $r = 0.3$, and $r = 0.5$. Power analysis for correlations can be performed by converted *r* to Cohen's *d* using:

$$d = \frac{2r}{\sqrt{1-r^2}} \quad (9)$$

after which Formula 5 can be used. This approach is used by for example G*Power (Faul, Erdfelder, Lang, & Buchner, 2007).

Equivalence test for Meta-Analyses

As noted earlier, rejecting small effects in an equivalence test requires large samples. If researchers want to perform an equivalence test with narrow equivalence bounds (e.g., $\Delta_L = -0.1$ and $\Delta_U = 0.1$), in most cases only a meta-analysis will have sufficient statistical power. Rogers and colleagues (1993) explain the straightforward approach to performing equivalence tests for meta-analyses:

$$Z_L = \frac{\Delta + \Delta_L}{SE}, Z_U = \frac{\Delta + \Delta_U}{SE} \quad (10)$$

Where Δ is the meta-analytic effect size (Cohen's *d* or Hedges' *g*), and SE is the meta-analytic standard error (or $\sqrt{var}$). These values can be calculated with meta-analysis software such as metafor (Viechtbauer, 2010). The two one-sided tests are rejected if $Z_L \leq -Z_\alpha$, and $Z_U \leq Z_\alpha$. Alternatively, the 90% confidence interval can be reported. If the 90% confidence interval falls within the equivalence bounds, the observed meta-analytic effect is statistically equivalent.

Setting Equivalence Bounds

In psychology, most theories do not state which effects are too small to be interpreted as support the proposed underlying mechanism. Instead, feasibility considerations are often the strongest

determinant of the effect sizes a researcher can reliably examine. In daily practice, researchers have a maximum sample size they are willing to collect in a single study (e.g., 100 participants in each between subject condition). Given a desired level of statistical power (e.g., 80%) and a specific α (e.g., 0.05) this implies a smallest effect size they find worthwhile to examine, or a smallest effect size of interest (SESOI; Lakens, 2014) they can reliably examine. With 100 participants in each condition, 80% desired power, and an α of 0.05, the SESOI in a null-effect significance test is $\Delta = 0.389$, and for an equivalence test, assuming a true effect size of 0, 80% power is achieved when $\Delta_L = -0.414$ and $\Delta_U = 0.414$. As such, without practical boundaries or theoretical boundaries that indicate which effect size is meaningful, the maximum sample size you are willing to collect implicitly determines your smallest effect size of interest. Therefore, setting equivalence boundaries to your SESOI in an equivalence test allows you to reject effect sizes larger than you find worthwhile to examine, given available resources.

This recommendation differs from practices in drug development, where equivalence bounds are often set by regulations (e.g., differences up to 20% are not considered to be clinically relevant). In psychology, such general regulations about what constitutes a meaningful effect seem unlikely to emerge, and perhaps even undesirable. Using equivalence bounds based on effect sizes a researcher finds worthwhile to examine do not allow psychologists to conclude an effect is too small to be meaningless *for anyone*. When other researchers believe a smaller effect size is plausible and theoretically interesting, they can design a study with a larger sample size to examine the effect. Until theories in psychology predict effects of a specific size, setting equivalence bounds to the effect sizes one finds worthwhile to examine will at least make it explicit which effect sizes a researcher predicts, and allows researchers to statistically falsify their predictions. In randomized controlled trials it is expected that equivalence bounds are pre-specified (e.g., see CONSORT guidelines, Piaggio et al., 2006), and this should also be considered best-practice in psychology.

Simonsohn (2015) proposes to test for inferiority for replication studies (an equivalence test where the lower bound is set to infinity). He suggests to set the upper equivalence bound in a replication study to the effect size that would have given an original study 33% power. For example, an original study with 60 participants divided equally across two independent groups has 33% power

to detect an effect of $d = 0.4$, so Δ_U is set to $d = 0.4$. This approach limits the sample size required to test for equivalence to 2.5 times the sample size of the original study. The goal is not to show the effect is too small to be feasible to study, but too small to have been reliably detected by the original experiment, thus casting doubt on the original observation.

If feasibility constraints are practically absent (e.g., in online studies), another starting point to set equivalence bounds is by setting bounds based on benchmarks for small, medium, and large effects. Although using these benchmarks to interpret effect sizes is typically recommended as a last resort (e.g., Lakens, 2013), their use in setting equivalence bounds seems warranted by the lack of other clear-cut recommendations. By far the best solution would be for researchers to specify their smallest effect size of interest when they publish an original result, or describe a theoretical idea (Morey & Lakens, under review). The use of equivalence testing will no doubt lead to a discussion about which effect sizes are too small to be worthwhile to examine in specific research lines in psychology, which in itself is progress.

Discussion

Equivalence tests are a simple adaptation of traditional significance tests that allow researchers to design studies that reject effects larger than pre-specified equivalence bounds. It allows researchers to reject effects large enough to be considered worthwhile. Adopting equivalence tests will prevent the common misinterpretations of non-significant *p*-values as the absence of an effect, and nudge researchers towards specifying which effects they find worthwhile. By providing a simple spreadsheet and R package to perform power calculations and equivalence tests for common statistical tests in psychology, researchers should be able to easily improve their research practices.

Rejecting effects more extreme than the equivalence bounds implies that we can conclude equivalence for a specific operationalization of a hypothesis. It is possible that a meaningful effect would be observed with a different manipulation or measure. Confounds can underlie observed equivalent effects. An additional non-statistical challenge in interpreting equivalence concerns the issue of whether an experiment was performed competently (Senn, 1993). Complete transparency (sharing all materials) is a partial solution since it allows peers to evaluate whether the experiment was well-designed (Morey et al., 2016), but this issue is not easily resolved when the actions of an

experimenter might influence the data (e.g., when a study relies on a confederate). In such experiments, even blinding the experimenter to conditions is no solution since an experimenter can interfere with the data quality of all conditions. This is an inherent asymmetry between demonstrating an effect, and demonstrating the absence of a worthwhile effect. The only solution for anyone skeptical about studies demonstrating equivalence is to perform an independent replication.

Equivalence testing is based on a Neyman-Pearson hypothesis testing approach that allows researchers to control error rates in the long run, and design studies based on a desired level of statistical power. Error rates in equivalence tests are controlled at the alpha level when the true effect equals the equivalence bound. When the true effect is more extreme than the equivalence bounds, error rates are smaller than the alpha level. It is important to take statistical power into account when determining the equivalence bounds, because in small samples (where confidence intervals are wide) a study might have no statistical power (i.e., the confidence interval will always be so wide that it is necessarily wider than the equivalence bounds).

There are alternative approaches to the TOST procedure. Updated versions of equivalence tests exist, but their added complexity does not seem to be justified by the small gain in power (for a discussion, see Meyners, 2012). There are also alternative approaches to providing statistical support for a small or null effect, such as estimation (calculating effect sizes and confidence intervals), specifying a region of practical equivalence (Kruschke, 2010), or calculating Bayes factors (Dienes, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Researchers should report effect size estimates in addition to hypothesis tests, and since Bayesian and Frequentist tests answer complementary questions, these tests can be reported side by side.

Other fields are able to use raw measures due to the widespread use of identical measurements (e.g., the number of deaths, the amount of money spent), but in some subfields in psychology the variability in the measures that are collected require standardized effect sizes to make comparisons across studies (Cumming & Fidler, 2009). A consideration of using standardized effect sizes as equivalence bounds is that in two studies with the same mean difference and confidence intervals in raw scale units (e.g., a difference of 0.2 on a 7-point scale with 90% CI[-0.13;0.17]) the same standardized equivalence bounds can lead to different significance levels in a equivalence test.

The reason for this is that the pooled standard deviation can differ across the studies, and as a consequence, the same equivalence bounds in standardized scores imply different equivalence bounds in raw scores. If this is undesirable, researchers should specify equivalence bounds in raw scores instead.

Ideally, psychologists could specify equivalence bounds in raw mean differences based on theoretical predictions or cost-benefit analyses, instead of setting equivalence bounds based on standardized benchmarks. My hope is that as equivalence tests become more common in psychology, researchers will start to discuss which effect sizes are theoretically expected while setting equivalence bounds. When theories do not specify which effect sizes are too small to be meaningless, theories can't be falsified. Whenever a study yields no significant effect, one can always argue that there is a true effect that is smaller than the study could reliably detect (Morey & Lakens, under review). Maxwell, Lau, and Howard (2015) suggest that replication studies demonstrate the absence of an effect by using equivalence bounds of Δ_L = -0.1 and Δ_U = 0.1, or even Δ_L = -0.05 and Δ_U = 0.05. I believe this creates an imbalance where we condone original studies that fail to make specific predictions, while replication studies are expected to test extremely specific predictions that can only be confirmed by collecting huge numbers of observations. Even though the substantial effort required to collect such large sample sizes can be shared by performing prospective meta-analyses based on large scale collaborations (Simons, Holcombe, & Spellman, 2014), we should expect theories proposed in original studies specify a smallest effect size of interest.

Extending your statistical toolkit with equivalence tests might very well be the easiest way for psychologists to improve their statistical and theoretical inferences. The TOST procedure provides a straightforward approach to reject effect sizes that one considers large enough to be worthwhile to examine.

Footnotes

[1] As Wellek (2010, p. 30) notes, for all practical purposes (such as the use of the accompanying spreadsheet), one can simply specify a very large value for the infinite equivalence bound.

[2] A 90% confidence interval (1-2α) is used instead of a 95% confidence interval (1-α) because two one-sided tests (each with an alpha of 5%) are performed.

[3] I'd like to thank Jake Westfall for this suggestion.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets. *Statistical Science*, *11*(4), 283–302.

Chow, S.-C., Shao, J., & Wang, H. (2002). A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics*, *12*(4), 441–456.

Cumming, G., & Fidler, F. (2009). Confidence Intervals: Better Answers to Better Questions. *Zeitschrift Für Psychologie / Journal of Psychology*, *217*(1), 15–26. https://doi.org/10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

Delacre, M., Lakens, D., & Leys, C. (2016). Why psychologists should by default use Welch's t-test instead of Student's t-test with unequal group sizes. *Under Review*.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Quantitative Psychology and Measurement*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2015.10.003

Eskine, K. J. (2013). Wholesome Foods and Wholesome Morals? Organic Foods Reduce Prosocial Behavior and Harshen Moral Judgments. *Social Psychological and Personality Science*, *4*(2), 251–254. https://doi.org/10.1177/1948550612447114

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Finch, S., Cumming, G., & Thomason, N. (2001). Colloquium on Effect Sizes: the Roles of Editors, Textbook Authors, and the Publication Manual Reporting of Statistical Inference in the

Journal of Applied Psychology: Little Evidence of Reform. *Educational and Psychological Measurement*, *61*(2), 181–210. https://doi.org/10.1177/0013164401612001

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology*, *8*(2), 195–204.

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

Hauck, D. 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

Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine*, *23*(12), 1921–1986. https://doi.org/10.1002/sim.1783

Kruschke, J. (2010). *Doing Bayesian data analysis: A tutorial introduction with R*. Academic Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Moery, E., & Calin-Jageman, R. J. (2016). Direct and Conceptual Replications of Eskine (2013): Organic Food Exposure Has Little to No Effect on Moral Judgments and Prosocial Behavior.

Social Psychological and Personality Science, *7*(4), 312–319. https://doi.org/10.1177/1948550616639649

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., … Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547.

Morey, R. D., & Lakens, D. (under review). Why most of psychology is statistically unfalsifiable.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J., Group, C., & others. (2006). Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *Jama*, *295*(10), 1152–1160.

Quertemont, E. (2011). How to Statistically Show the Absence of an Effect. *Psychologica Belgica*, *51*(2), 109. https://doi.org/10.5334/pb-51-2-109

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

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

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*(4), 403–411. https://doi.org/http://dx.doi.org.dianus.libr.tue.nl/10.1037/1082-989X.3.4.403

Senn, S. (2007). *Statistical issues in drug development* (2nd ed). Chichester, England ; Hoboken, NJ: John Wiley & Sons.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555.

Simonsohn, U. (2015). Small Telescopes Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1–48.

Weber, R., & Popova, L. (2012). Testing Equivalence in Communication Research: Theory and Application. *Communication Methods and Measures*, *6*(3), 190–213. https://doi.org/10.1080/19312458.2012.703834

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed). Boca Raton: CRC Press.