Why “Not Different” Is Not the Same as “Equivalent”: An Introduction to Equivalence Testing in Psychological Research

Sunthud Pornprasertmanit

Faculty of Psychology, Chulalongkorn University

Abstract

In psychological research, researchers may wish to examine whether two groups have similar means on a target variable, such as whether online and in-class learning are equally effective. However, researchers often rely on traditional null hypothesis testing (TNHST) to claim equivalence based on non-significant results, an approach that may lead to misleading conclusions. This article explains the limitations of this practice and introduces the concepts and methods of equivalence testing, also referred to as the negligible effect significance test. In brief, researchers specify a smallest effect size of interest (SESOI), below which an effect is considered practically negligible. They then test whether the observed statistic is significantly smaller than this threshold. If so, it can be concluded that the population effect is sufficiently small to be considered negligible. The article also provides examples of applications in psychological research and demonstrates how equivalence can be evaluated using outputs from commonly used statistical software or online tools.

Keywords: equivalence testing, negligible effect significance test, effect size, confidence interval, null hypothesis significance testing

Author Note

Sunthud Pornprasertmanit, Faculty of Psychology, Chulalongkorn University

The author thanks Narun Pat (also known as Narun Pornpatthanangkul) for his valuable comments and suggestions on this article.

Participation in this academic presentation was supported by funding from the Faculty of Psychology, Chulalongkorn University.

Correspondence concerning this article should be addressed to Sunthud Pornprasertmanit, Faculty of Psychology, Chulalongkorn University, Bangkok, Thailand, or via email at sunthud.p@chula.ac.th

This article was originally published in Thai in the proceedings of the 11th Thailand National Conference on Psychology 2026, held on June 4–5, 2026, in Chiang Mai, Thailand. The original article is available at https://tncp2026.app-mice.net/

Introduction

In psychological research, most researchers aim to test correlations or differences between groups, such as examining whether intelligence quotient (IQ) is related to job performance (Schmidt & Hunter, 1998), or comparing differences between treatment and control groups in experiments (Maxwell et al., 2018). However, many psychological hypotheses concern "equivalence" between groups or "the absence of a relationship" between variables. For example, in quasi-experimental studies, researchers may wish to examine whether two sample groups do not differ on important baseline variables (Reichardt, 2019). In decision-making research, investigators may seek to test whether highly trained experts can make decisions equally well in quiet environments and in noisy environments (Szalma & Hancock, 2011; Zhou et al., 2024), because experts are expected to be capable of managing distractions. In addition, in measurement and evaluation research, researchers may wish to demonstrate that an aptitude test is unrelated to English language ability in order to confirm that the test is not contaminated by extraneous variables and truly measures the intended construct (Alderman, 1982; American Educational Research Association et al., 2014).

Statistically, it is often assumed that if researchers wish to claim "no difference" or "no relationship," they can use non-significant hypothesis test results as evidence. For example, in research examining the effect of religious priming on prosocial behavior (Shariff & Norenzayan, 2007), researchers may compare experimental and control groups on certain baseline characteristics, such as belief in God. If the independent *t*-test result is not statistically significant, they may conclude that the two groups are "equivalent" on this variable. However, drawing conclusions of equivalence from non-significant results has important limitations, especially when the sample size is small, which leads to low statistical power. Under such conditions, the effect size may still be large enough that it cannot reasonably be ignored.

For example, suppose that the mean score of belief in God in the experimental group is 50 points (SD = 10), whereas the control group has a mean of 56 points (SD = 10), with 20 participants in each group. When the data are analyzed using an independent *t*-test, the result is $t(38) = -1.897$, $p = .074$, which is not statistically significant. However, when considering the standardized effect size using Cohen's *d*, the value equals -0.6. According to Cohen's (1988) recommendations, values of 0.2, 0.5, and 0.8 represent small, medium, and large effects, respectively. Therefore, a value of 0.6 reflects a moderately large difference between groups. In other words, even though the test result is not statistically significant, the difference between groups is large enough that it should not be ignored. Concluding that the two groups are "equivalent" may therefore lead to misleading conclusions.

Conversely, suppose that the two groups have identical means (M = 50, SD = 10, n = 20). An independent *t*-test would yield $t(38) = 0$, $p = 1.00$, and the standardized mean difference, or Cohen's *d*, would equal 0. Such results may increase researchers' confidence in concluding equivalence. However, when examining the 95% confidence interval (CI) of the mean difference—which corresponds to a two-tailed traditional hypothesis test at $\alpha = .05$—the interval ranges from (-5.33, 5.33). Likewise, the CI for Cohen's *d* ranges from (-0.53, 0.53). This means that even though the observed difference in the sample

equals zero, the true population parameter could still reflect a moderate difference. Therefore, failing to detect a difference in a sample does not mean that the two groups have equal means. It only indicates that there is "not yet sufficient evidence" to conclude that a difference exists.

Theoretically, a true difference of exactly zero between groups, or a correlation of exactly zero between variables, may occur—for example, in situations involving perfect random assignment in experiments. However, in practice, differences are rarely exactly zero. Instead, they may simply be very small and practically unimportant. For example, in studies of gender differences in mathematical ability (Steele & Ambady, 2006), differences may truly exist but be so small that they lack practical importance. Therefore, the more important question is not "whether a difference exists," but rather "whether the difference is small enough to be ignored." For instance, an effect size smaller than Cohen's d = 0.2 may be considered negligible.

Equivalence Testing

Equivalence testing originated in biostatistics (Schuirmann, 1987) and commonly appears in the form of non-inferiority trials (Snapinn, 2000), which are used to examine whether a new treatment has effectiveness or risk comparable to a standard treatment. In psychology, the concept of equivalence testing has been discussed for a long time (Seaman & Serlin, 1998; Steiger, 2004), but it became much more widely recognized following the work of Lakens (2017) and Lakens et al. (2018). More recently, Psychological Methods published a review article by Beribisky and Cribbie (2026), which referred to this approach as the "negligible effect significance test" in order to emphasize that the goal is not to examine whether two statistics are exactly "equal," but whether the statistic is small enough to be ignored. This terminology is especially suitable for statistics assessing relationships, such as correlations or regression coefficients, because the phrase "small enough to be ignored" sounds more natural than asking whether the statistics are "equivalent." Nevertheless, the term "equivalence testing" remains the more widely used term at present. Therefore, this article continues to use the term equivalence testing.

In equivalence testing, researchers must specify the smallest effect size of interest (SESOI; Lakens, 2017), or what Beribisky and Cribbie (2026) refer to as the minimally meaningful effect size. In other words, researchers must define what magnitude of effect should begin to be considered important. For example, suppose two teachers teach English courses at the same grade level. A difference in test scores exceeding 5 points may be considered meaningfully important and deserving of intervention. In this case, 5 points would be specified as the SESOI. In another context, such as evaluating fairness in employment interviews (Sacco et al., 2003), researchers may define differences greater than Cohen's d = 0.5 as practically meaningful, whereas values between -0.5 and 0.5 are considered negligible (SESOI of d = 0.5). As another example, when examining whether chronological age is unrelated to job performance among employees who already possess sufficient background knowledge (Guzzo et al., 2022), researchers may specify that correlations of .25 or greater are too large to ignore (SESOI of r = .25).

The principle of equivalence testing is to determine whether the observed statistic is smaller than the SESOI in the population. For example, if the SESOI for a correlation is set at .25 and the sample correlation equals .07, researchers must test whether .07 is significantly smaller than .25. Because correlations may also take negative values, researchers must additionally test whether .07 is significantly greater than -.25. If both tests are statistically significant, researchers may conclude that the correlation lies within the range that can be ignored, or is "practically equivalent to zero" (that is, between -.25 and .25).

Statistically, this idea can be expressed through null hypotheses (H_0). For two-sided statistics, researchers formulate the null hypotheses as H_0: $\theta \geq$ SESOI and H_0: $\theta \leq$ -SESOI, where θ represents the target statistic. The alternative hypotheses (H_A) are H_A: $\theta <$ SESOI and H_A: $\theta >$ -SESOI. Only if both null hypotheses can be rejected ($p < .05$ for both tests) can equivalence be claimed. In other words, both alternative hypotheses must hold simultaneously, meaning that -SESOI < θ < SESOI. This procedure is called the two one-sided tests procedure (TOST), which differs from the conventional two-tailed testing approach commonly used in psychological research (see Figure 1a).

For statistics that involve one-sided hypothesis testing, such as the coefficient of determination (R^2) or eta-squared (η^2), which range from 0 upward, equivalence testing requires only a single hypothesis. For example, if the SESOI for η^2 is set at .01, researchers formulate the null hypothesis as H_0: $\eta^2 \geq .01$ and the alternative hypothesis as H_A: $\eta^2 < .01$. If the null hypothesis can be rejected, researchers may conclude that the effect of the variable is small enough to be ignored (see Figure 1b). Table 1 summarizes the steps involved in equivalence testing to facilitate practical application.

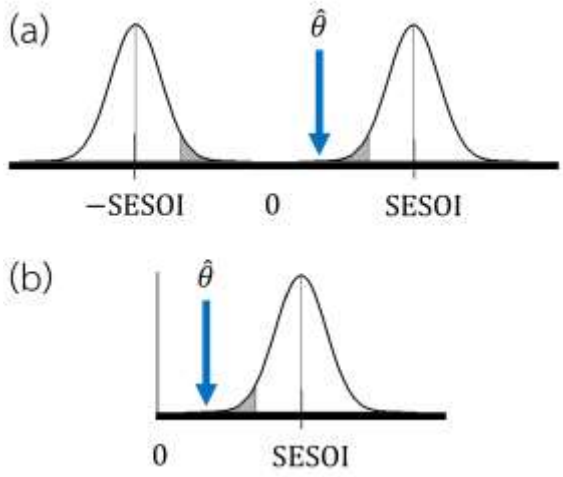

Figure 1 illustrates the principle of equivalence testing. Panel (a) shows the case of two-sided statistics, where researchers must demonstrate that the statistic is significantly smaller than the SESOI and significantly larger than -SESOI in order to claim equivalence. Panel (b) shows the case of one-sided statistics, where researchers only need to demonstrate that the statistic is significantly smaller than the SESOI.

Table 1

Summary of Steps in Equivalence Testing

Type of Statistic	Two One-Sided Tests (TOST)	Confidence Interval (CI) Approach
Two-sided	1. Specify the SESOI. 2. Test whether the statistic is significantly smaller than the SESOI. 3. Test whether the statistic is significantly greater than -SESOI. 4. If both (2) and (3) are statistically significant, **conclude equivalence**. 5. If at least one condition is not met, **equivalence cannot be concluded**.	1. Specify the SESOI and construct the NEI. 2. Construct the CI and compare it with the NEI. 3. If the CI lies entirely within the NEI, **conclude equivalence**. 4. If part or all of the CI lies outside the NEI, **equivalence cannot be concluded**.
One-sided	1. Specify the SESOI. 2. Test whether the statistic is significantly smaller than the SESOI. 3. If the result is statistically significant, **conclude equivalence**. 4. If the result is not statistically significant, **equivalence cannot be concluded**.	1. Specify the SESOI and construct the NEI. 2. Construct the CI and compare it with the NEI. 3. If the CI lies entirely within the NEI, **conclude equivalence**. 4. If part or all of the CI lies outside the NEI, **equivalence cannot be concluded**.

Interpreting Equivalence Testing Through Confidence Intervals

Equivalence testing, whether involving one-sided or two-sided statistics, can conveniently be interpreted through comparisons between CIs and negligible effect intervals (NEIs). Specifically, analysts may define the NEI as [-SESOI, SESOI] for two-sided statistics, or [0, SESOI] for one-sided statistics. The analyst then compares the location of the CI relative to the NEI. When the significance level (α) is set at .05, the CI corresponding to the two one-sided tests procedure (TOST) is the 90% CI [$(1 - 2\alpha) \times 100\%$]. Interpretation therefore becomes straightforward: Researchers simply examine whether the CI falls entirely within the negligible effect interval.

In Figure 2a (the two-sided case), the CI lies entirely within the NEI. Specifically, the lower bound of the CI is greater than -SESOI, and the upper bound is smaller than SESOI. Researchers may therefore conclude that the population parameter is small enough to be ignored and may claim equivalence. Conversely, if part or all of the CI falls outside the NEI (rows 2–4), equivalence cannot be concluded because it remains possible that the population parameter lies outside the negligible range. For Figure 2b (the one-sided case), the NEI ranges from [0, SESOI], and the same interpretive principle applies. That is, if the CI lies entirely within the NEI, equivalence may be concluded. However, if any portion of the CI lies outside the NEI, equivalence cannot be concluded.

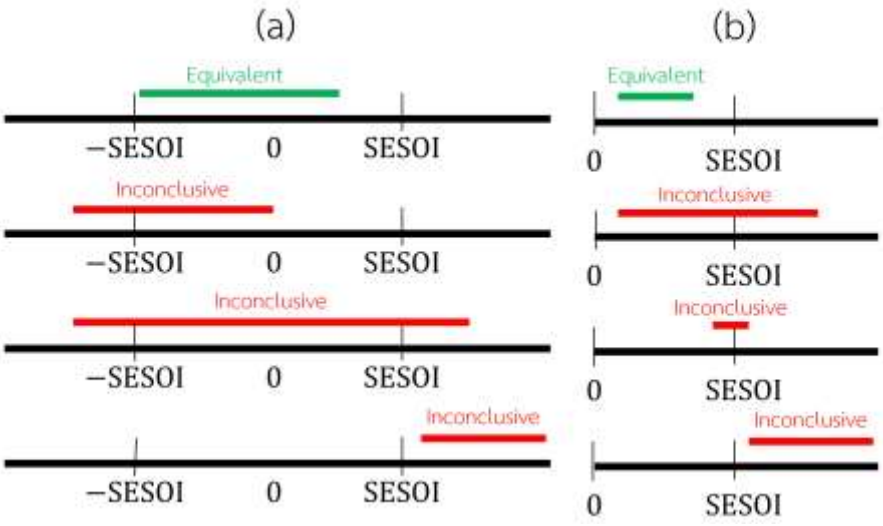

Figure 2 illustrates the principles of equivalence testing through CIs (represented by the red or green lines in the figure). Panel (a) presents the case of two-sided statistics, where the NEI ranges from [-SESOI, SESOI], whereas panel (b) presents the case of one-sided statistics, where the NEI ranges from [0, SESOI]. In both cases, the first row depicts a situation in which the CI lies entirely within the NEI, allowing equivalence to be concluded. Rows 2 and 3 show cases where part of the CI lies outside the NEI, and Row 4 shows a case where the entire CI lies outside the NEI. In the situations illustrated in Rows 2–4, equivalence cannot be concluded.

In general, TOST results at the .05 significance level and comparisons between 90% CIs and NEIs will lead to consistent conclusions. However, in some cases—such as statistics whose standard errors depend on the estimated value itself (e.g., correlations or proportions)—the conclusions derived from the two approaches may differ. Such issues are beyond the scope of this article.

Combining Conclusions with Traditional Null Hypothesis Testing

Researchers may wonder how the results of equivalence testing can be interpreted together with results from traditional null hypothesis testing (TNHST). Table 2 presents the possible conclusions that may arise when both approaches are used jointly. In general, TNHST provides information about the "existence of an effect" (that is, whether the population parameter differs from zero), whereas equivalence testing provides information about the "magnitude of the effect" (that is, whether the effect is small enough to be ignored). Therefore, using both methods together allows for more complete interpretation of research findings.

For two-sided statistics, if the TNHST result is statistically significant, researchers may infer that the population parameter is unlikely to equal zero and that the effect follows the direction indicated by the estimate (Jones & Tukey, 2000). For example, if the sample estimate is positive, the population parameter is likely positive (and unlikely negative). Likewise, if the sample estimate is negative, the population parameter is likely negative (and unlikely positive). In contrast, if the result is not statistically

significant, the population parameter may be zero, positive, or negative; therefore, no directional conclusion can yet be drawn. For one-sided statistics, a statistically significant TNHST result indicates that the population parameter differs from zero and that the effect exists in the specified direction (consistent with the direction of the statistic). A non-significant result, however, suggests that the parameter may equal zero. In contrast, equivalence testing does not provide information about the direction of the effect or whether the interval includes zero. Instead, it addresses whether the magnitude of the effect is small enough to be considered negligible. Consequently, considering the results from both methods together helps distinguish between "absence of evidence for an effect" and "evidence that the effect is small."

Table 2

Possible Conclusions from Equivalence Testing and Traditional Null Hypothesis Testing (TNHST)

Equivalence	TNHST statistically significant	TNHST not statistically significant
	One-Sided Statistics	
Equivalent	- Different from zero in the population	- The parameter may be 0
	- Clear direction	- Direction is unclear
	- Small effect size (within the negligible range)	- Small effect size (within the negligible range)
Inconclusive	- Different from zero in the population	- The parameter may be 0
	- Clear direction	- Direction is unclear
	- Effect size may be small or large	- Effect size may be small or large
	Two-Sided Statistics	
Equivalent	- Different from zero in the population	- The parameter may be 0
	- Small effect size (within the negligible range)	- Small effect size (within the negligible range)
Inconclusive	- Different from zero in the population	- The parameter may be 0
	- Effect size may be small or large	- Effect size may be small or large

Examples of Data Analysis

If readers are familiar with the R language, they may directly conduct equivalence testing using packages such as TOSTER (Caldwell, 2022), marginaleffects (Arel-Bundock et al., 2024), or negligible (Cribbie et al., 2026). In addition, Lakens (2017) has made available an Excel file for equivalence testing on OSF (https://osf.io/q253c/files/qzjaj), which can be conveniently used. However, this article focuses on using outputs from commonly used statistical software packages in psychology, such as SPSS or Jamovi, and then applying equivalence testing together with results from traditional null hypothesis significance testing (TNHST).

Independent Mean Differences (Independent *t*-test). An example from Bloom et al. (2024), which examined the effects of hybrid work compared with exclusively in-office work on job performance, found a performance difference of 0.056 (*SE* = 0.043), 95% *CI* = (-0.029, 0.140), and 90% *CI* = (-0.015, 0.127) during the period from July to December 2021. These results can readily be obtained from common statistical software packages such as SPSS or Jamovi. In this study, the SESOI was defined as Cohen's *d* = 0.5. Dividing the 90% CI above by the standard deviation of job performance (0.989) yields a

90% CI for Cohen's *d* of (-0.015, 0.128). When this CI is compared with the NEI = (-0.5, 0.5), the entire interval falls within the NEI. Researchers may therefore conclude that hybrid work and exclusively in-office work are practically equivalent in terms of job performance. At the same time, the TNHST result was not statistically significant, indicating that the direction of the effect cannot be determined. In other words, it cannot be concluded whether hybrid work produces better or worse outcomes. However, it can be concluded that the observed effect is small enough to be ignored in practical terms. Therefore, if researchers are able to obtain a 90% CI and know the standard deviation used to compute Cohen's *d*, they can conduct equivalence testing without requiring additional specialized software.

Correlation. An example from Orben and Przybylski (2019), which examined the relationship between digital screen time and psychological well-being among 11,884 adolescents[1], reported a correlation of $r = -.02$ ($p < .05$). The researchers defined the SESOI such that correlations with magnitudes no greater than .10 would be considered negligible. However, commonly used statistical software packages such as SPSS or Jamovi generally do not directly report CIs for correlations. Researchers may therefore use online tools such as the Psychometrica website developed by Alexandra Lenhard (https://www.psychometrica.de/correlation.html), which computes CIs based on Fisher's (1925) method[2]. The calculation yields a 90% CI for the correlation of (-.035, -.005), as shown in Figure 3. When compared with the NEI = (-.10, .10), the entire CI falls within the negligible effect interval. Therefore, the relationship between screen time and psychological well-being may be concluded to be small enough to be ignored. Although the relationship is negative—and statistically significant—the effect size is so small that it lacks practical importance. Readers may alternatively use Excel or R packages to calculate CIs and conduct equivalence testing with greater precision than is possible through online tools.

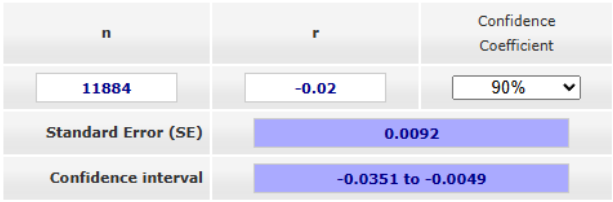

Figure 3. Results from calculating the 90% CI for a correlation using Alexandra Lenhard's Psychometrica website.

1 This analysis was adapted from the original report by Orben and Przybylski (2019), which used the coefficient of determination in equivalence testing. The presentation in this article was modified into the form of a correlation coefficient to align with the article's instructional purpose.

2 The website states that the calculation formula is based on Bonett and Wright (2000), which is itself based on Fisher's (1925) method.

Determining the Smallest Effect Size of Interest (SESOI)

The most important question in equivalence testing concerns the determination of the SESOI. If the SESOI is set too wide, effects that are actually meaningful may incorrectly be classified as negligible, potentially leading to misleading conclusions. Conversely, if the SESOI is set too narrowly, researchers will require very large sample sizes in order for the CI to become sufficiently narrow to fall within the NEI. In practice, this is often difficult to achieve and results in low power for concluding equivalence. Therefore, selecting an appropriate SESOI is an extremely important issue.

Nevertheless, determining what magnitude of effect should be considered "meaningful" necessarily depends on the research context, and no universal criterion applies to all situations. In some contexts, researchers may rely on the concept of the just noticeable difference (Lakens et al., 2018), such as the level of association in a scatterplot at which observers begin to perceive a relationship, or the level of IQ difference at which two individuals become meaningfully distinguishable. The advantage of this approach is that it is intuitively understandable, although in practice it may be difficult to implement.

Another widely accepted approach is to define the SESOI based on expert consensus within the field (Beribisky & Cribbie, 2026; Riesthuis et al., 2022; Riesthuis & Woods, 2024). For example, in evaluating test fairness, researchers may define Cohen's d = 0.1 as the maximum acceptable group difference because such a difference reflects approximately 4% of one group scoring above the mean of the other group[3], and differences larger than this may have practical or policy implications. As another example, when studying whether age is unrelated to personality, researchers may specify that if a variable explains no more than 1% of variance, it will be considered unimportant. In such a case, r^2 = .01 or $|r|$ = .1 may be used as the SESOI, which is similar to the criterion used by Orben and Przybylski (2019).

In some fields, there may be no clear expert consensus, but researchers may instead rely on analyses of effect sizes commonly observed in the literature. Examples include Brydges (2019) in gerontology and Lovakov and Agadullina (2021) in social psychology, both of which used meta-analytic evidence. Such analyses may provide another basis for determining the SESOI. Lakens et al. (2018) suggested that researchers identify meta-analytic findings most similar to their own context and use the average or medium-sized effect as the SESOI, or alternatively use the lower bound of the CI (e.g., the 5th percentile) to establish a more stringent criterion.

In practice, however, many researchers still rely on conventional effect size benchmarks (e.g., Cohen's d = 0.2, 0.5, 0.8), despite recommendations to avoid using these values directly (Lakens et al., 2018; Beribisky & Cribbie, 2026). For example, setting the SESOI at Cohen's d = 0.5 (as in Bloom et al., 2024) may be considered relatively large, but it is feasible with realistic sample sizes (e.g., approximately

[3] Calculated using the pnorm() function in R. For example, pnorm(q = -0.1) yields 0.46, indicating that approximately 4% of one group scores above the mean of the other group under the normal distribution.

50 participants per group).[4] In contrast, specifying a smaller SESOI, such as d = 0.2, would require substantially larger sample sizes, potentially hundreds of participants per group.[5] Smaller effect sizes may be more theoretically appropriate because even small effects can have theoretical importance (see Funder & Ozer, 2019). However, practical constraints on sample size often lead researchers to adopt medium-sized effects as their SESOI.

Given these limitations, selecting the SESOI necessarily involves balancing conceptual rigor against practical feasibility. Therefore, when SESOI values are determined under data collection constraints, researchers should interpret results cautiously and report CIs for effect sizes alongside equivalence testing results. Doing so allows readers to evaluate conclusions under alternative SESOI specifications.

Conclusion

If researchers fail to find statistically significant differences or relationships using traditional null hypothesis testing, this does not mean that the substantive hypothesis is true or that the effect size is equivalent to zero. A non-significant result merely indicates that the null hypothesis remains one plausible explanation. A more appropriate approach is to examine whether the population parameter is small enough to be ignored through equivalence testing. This approach allows researchers to claim the "absence of a practically meaningful effect" with supporting evidence.

Equivalence testing can be applied to a wide range of psychological research contexts, such as examining baseline equivalence between groups prior to experiments, evaluating the fairness or bias of psychological tests, or testing theories that predict no meaningful practical differences—for example, comparisons of treatment effectiveness or comparisons between different tests intended to measure the same underlying construct (e.g., intelligence). Therefore, equivalence testing should be considered an additional standard tool alongside traditional null hypothesis significance testing, especially in situations where researchers wish to claim that there is "no effect" with empirical support.

References

Alderman, D. L. (1982). Language proficiency as a moderator variable in testing academic aptitude. *Journal of Educational Psychology, 74*(4), 580–587. https://doi.org/10.1037/0022-0663.74.4.580

[4] Calculated using the ss.aipe.smd() function in the MBESS package (Kelley, 2007), assuming that the true population difference is Cohen's d = 0.1 and that the desired confidence interval has a width of 0.8 (-0.3, 0.5), which yields a sample size of approximately 49 participants (see Kelley & Rausch, 2006).

[5] Calculated using the same function, assuming Cohen's d = 0.1 and a desired confidence interval width of 0.2 (0, 0.2), which yields a sample size of approximately 770 participants.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Arel-Bundock, V., Greifer, N., & Heiss, A. (2024). How to interpret statistical models using marginaleffects for R and Python. *Journal of Statistical Software, 111*, 1-32. https://doi.org/10.18637/jss.v111.i09

Beribisky, N., & Cribbie, R. A. (2026). A primer on equivalence (negligible effect) testing. *Psychological Methods*. https://doi.org/10.1037/met0000800

Bloom, N., Han, R., & Liang, J. (2024). Hybrid working from home improves retention without damaging performance. *Nature*, *630*(8018), 920-925. https://doi.org/10.1038/s41586-024-07500-2

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika, 65*(1), 23-28. https://doi.org/10.1007/BF02294183

Brydges, C. R. (2019). Effect size guidelines, sample size calculations, and statistical power in gerontology. *Innovation in Aging, 3*(4), igz036. https://doi.org/10.1093/geroni/igz036

Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R package. *PsyArXiv*. https://doi.org/10.31234/osf.io/ty8de

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hilldale, NJ: Erlbaum.

Cribbie, R. A., Alter, U., Beribisky, N., Chalmers, R. P., Counsell, A., Farmus, L., Martinez Gutierrez, N., & Ng, V. (2026). *Negligible: A collection of functions for negligible effect/equivalence testing*. R Package Version 0.1.11. https://cran.r-project.org/web/packages/negligible/index.html

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Hafner Press.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Guzzo, R. A., Nalbantian, H. R., & Anderson, N. L. (2022). Age, experience, and business performance: A meta-analysis of work unit-level effects. *Work, Aging and Retirement*, *8*(2), 208-223. https://doi.org/10.1093/workar/waab039

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, *5*(4), 411-414. https://doi.org/10.1037/1082-989X.5.4.411

Kelley, K. (2007). Methods for the Behavioral, Educational, and Social Sciences: An R Package. *Behavior Research Methods, 39*, 979–984. https://doi.org/10.3758/BF03192993

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363-383. https://doi.org/10.1037/1082-989X.11.4.363

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

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology, 51*(3), 485-504. https://doi.org/10.1002/ejsp.2752

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed). New York, NY: Routledge.

Orben, A., & Przybylski, A. K. (2019). Screens, teens, and psychological well-being: Evidence from three time-use-diary studies. *Psychological Science*, *30*(5), 682-696. https://doi.org/10.1177/0956797619830329

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford Publications.

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology, 36*(1), 203–215. https://doi.org/10.1002/acp.3911

Riesthuis, P., & Woods, J. (2024). "That's just like, your opinion, man": The illusory truth effect on opinions. *Psychological Research, 88*(1), 284–306. https://doi.org/10.1007/s00426-023-01845-5

Sacco, J. M., Scheu, C. R., Ryan, A. M., & Schmitt, N. (2003). An investigation of race and sex similarity effects in interviews: A multilevel approach to relational demography. *Journal of Applied Psychology*, *88*(5), 852-865. https://doi.org/10.1037/0021-9010.88.5.852

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262-274. https://doi.org/10.1037/0033-2909.124.2.262

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. 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(4), 403–411. https://doi.org/10.1037/1082-989X.3.4.403

Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, *18*(9), 803-809. https://doi.org/10.1111/j.1467-9280.2007.01983.x

Snapinn, S. M. (2000). Noninferiority trials. *Trials*, *1*(1), 19. https://doi.org/10.1186/cvm-1-1-019

Steele, J. R., & Ambady, N. (2006). "Math is Hard!" The effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology*, *42*(4), 428-436. https://doi.org/10.1016/j.jesp.2005.06.003

Steiger, J. H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychological Bulletin*, *137*(4), 682-707. https://doi.org/10.1037/a0023987

Zhou, H., Molesworth, B. R., Burgess, M., & Hatfield, J. (2024). The effect of moderate broadband noise on cognitive performance: a systematic review. *Cognition, Technology & Work*, *26*(1), 1-36. https://doi.org/10.1007/s10111-023-00746-2