*Title*: Correspondence Measures for Assessing Replication Success

*Authors*: Peter M. Steiner, University of Maryland – College Park, psteiner@umd.edu (corresponding author)

Patrick Sheehan, University of Maryland – College Park, psheehan@umd.edu

Vivian C. Wong, University of Virginia – Charlottesville, vcw2n@virginia.edu

*Author note*:

This work has been supported by NSF grant #2015-0285-00 and
IES grants #R305D190043 and #R305D220034.

We thank Dennis Kondžić and Jerome Hoffmann for helpful discussions and comments, and Steffen Erickson for providing the Stata script.

Part of this work has been presented at the Annual Meeting of the American Educational Research Association (AERA) in San Diego (2022), the Conference of the Society for Research on Educational Effectiveness (SREE) in Arlington (2022), and the Annual Meeting of the American Educational Research Association (AERA) in Chicago (2023). A version of the manuscript is also available at PsyArXiv Preprints: https://psyarxiv.com/6vy75/ .

# Correspondence Measures for Assessing Replication Success

Peter M. Steiner[1], Patrick Sheehan[1], and Vivian C. Wong[2]

[1]University of Maryland - College Park

[2]University of Virginia - Charlottesville

May, 2023

## Author Note

Correspondence concerning this article should be addressed to Peter M. Steiner, Department of Human Development and Quantitative Methodology, University of Maryland–College Park; email: psteiner@umd.edu

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grants R305D190043 and R305D220034.

We thank Dennis Kondžić and Jerome Hoffmann for helpful discussions and comments, and Steffen Erickson for providing the Stata script.

**Abstract.** Given recent evidence challenging the replicability of results in the social and behavioral sciences, critical questions have been raised about appropriate measures for determining replication success in comparing effect estimates across studies. At issue is the fact that conclusions about replication success often depend on the measure used for evaluating correspondence in results. Despite the importance of choosing an appropriate measure, there is still no wide-spread agreement about which measures should be used. This paper addresses these questions by describing formally the most commonly used measures for assessing replication success, and by comparing their performance in different contexts according to their replication probabilities – that is, the probability of obtaining replication success given study-specific settings. The measures may be characterized broadly as conclusion-based approaches, which assess the congruence of two independent studies' conclusions about the presence of an effect, and distance-based approaches, which test for a significant difference or equivalence of two effect estimates. We also introduce a new measure for assessing replication success called the correspondence test, which combines a difference and equivalence test in the same framework. To help researchers plan prospective replication efforts, we provide closed formulas for power calculations that can be used to determine the minimum detectable effect size (and thus, sample sizes) for each study so that a predetermined minimum replication probability can be achieved. Finally, we use a replication dataset from the Open Science Collaboration (2015) to demonstrate the extent to which conclusions about replication success depend on the correspondence measure selected.

**Keywords**: Causal replication, equivalence test, correspondence test, power analysis

# Introduction

The advancement of science presumes that research results are governed by a set of stable, potentially context-specific laws that can be uncovered through repeated experimentation or observation, resulting in reliable knowledge (Popper, 1959; Schmidt, 2009). Replication failure, however, challenges the accumulation of trusted knowledge and the credibility of scientific results. Over the last two decades, low replication rates of published research findings have led many to conclude that research results are fragile and hard to replicate, prompting a "replication crisis" across the social and behavioral sciences (e.g., Duvendack et al., 2017; Ioannidis, 2005, 2008; Klein et al., 2014; Madden et al., 1995; Makel and Plucker, 2014; Maxwell et al., 2015; Open Science Collaboration, 2015; Valentine et al., 2011).

But when results are compared across replication studies, when are effect estimates sufficiently similar (or different) to conclude replication success (or failure)? Over the years, this question has been addressed differently by researchers, with some looking at correspondence in the direction, size, and statistical significance pattern of effects (Open Science Collaboration, 2015; Wilde & Hollister, 2007), while others have examined results from statistical tests of difference (Open Science Collaboration, 2015) or equivalence (Dong & Lipsey, 2018; Wellek, 2010) of study effects. The challenge with such myriad approaches for assessing replication success, however, is three-fold. First, the assessment of replication success or failure from replication efforts may be ambiguous if multiple criteria for assessing correspondence in results yield contradictory conclusions. Table 1 summarizes results from two hypothetical studies in a replication effort. Study 1 has an effect estimate of 12 points and is statistically significant, and Study 2 has an effect estimate of 11 points but is not statistically significant. A researcher assessing correspondence in results would arrive at different conclusions depending on the measure used for determining replication success. A test of statistical difference of the two effects yields a non-significant result, leading the researcher to conclude the replication success was achieved; however, a comparison of statistical significance pattern or an equivalence test of effects leads the researcher to a different conclusion—that the replication failed.

Second, without clear consensus about appropriate correspondence measures for determining replication success, reporting results from replication efforts may be subject to its own form of biases. When multiple criteria yield different conclusions about replication success—such as the example in Table 1—researchers may be tempted to report and weigh more heavily results from criteria that demonstrate the robustness of results. Such ambiguity for drawing conclusions about replication success provides opportunities for researchers to engage in questionable research practices, like p-hacking, results-driven choice of assessment criteria, or selective outcome reporting, when analyzing the data of individual studies, assessing replication success, and reporting of the results of the replication effort.

Finally, without well-defined and *a priori* chosen criteria for determining replication success, the researcher cannot ensure that a prospectively planned replication effort has sufficient statistical power to guarantee a minimum replication probability of, say, .8 if the unknown true effects of the two studies were identical. This is especially important because study-specific power requirements for determining replication success across two studies are generally larger than what is needed for detecting an effect in a single study (Anderson & Kelley, 2022; Bonett, 2021; Hedges & Schauer, 2018). To see the logic of why this is true, consider two studies that have a power of .8 to detect the unknown true effect. Then, the probability of concluding replication success from both studies, that is, obtaining either two significant or two non-significant effect estimates, is not very high—it is $.8 \times .8 + .2 \times .2 = .68$. Moreover, statistical power for detecting replication success and thus sample size requirements will differ depending on which criterion is used for assessing replication success. In the absence of prospective planning it is also unclear whether researchers should expect high or only low probabilities of replication success, even if the underlying true effects were identical.

These methodological challenges in replication are especially important now given recent federal funding initiatives from the Institute for Education Sciences (Institute of Education Sciences, 2022), National Institutes of Health (Collins & Tabak, 2014), and National Science Foundation (Bollen et al., 2015) to support and promote studies that evaluate the replicability of findings. For example, a recent Request for Applications for Systematic Replication Efforts from the Institute for Education Sciences asked researchers to specify a plan for comparing effects to determine replication success, but did not provide guidance on what criteria should be used, and was silent on power analyses for demonstrating that replication success could be detected in the proposed study (Institute of Education Sciences, 2022). Given recent developments in the replication literature that demonstrate the lack of statistical power for comparing effects in most replication efforts (Hedges & Schauer, 2018; Maxwell et al., 2015; McShane & Böckenholt, 2014), replications funded by these initiatives are likely to yield ambiguous or even incorrect conclusions unless both studies were adequately powered to yield at least a replication probability of .8 (if effects where identical) with respect to a predetermined correspondence measure.

This article addresses these concerns. We begin by formalizing common approaches for evaluating replication success, which we describe as *conclusion-based* and *distance-based* measures for determining correspondence in results. Conclusion-based measures examine whether results from two studies yield the same substantive conclusion about an intervention or policy's effectiveness; they include examining the direction, magnitude, and statistical significance pattern of effects from individual studies. Distance-based measures examine how close or different two estimates are, usually using a difference estimate of two study effects. To address sampling error, researchers may conduct tests of statistical difference between two study-level effect estimates. But if individual studies in the replication effort are not well-powered to detect a statistical difference in results, the conclusion from the difference test may be ambiguous—where a null result may mean that there is no difference in the effect estimates or it may mean that the test is underpowered for detecting a difference. Instead, researchers may opt to examine the statistical equivalence between two effect estimates. But equivalence tests require careful determination of a tolerance threshold for defining how close two estimates must be to be considered equivalent and generally require large sample sizes for establishing equivalence (Anderson & Kelley, 2022; Anderson & Maxwell, 2016; Bonett, 2021; Maxwell et al., 2015; Wellek, 2010). Many more approaches for assessing replication success have been suggested, like prediction intervals (Patil et al., 2016) or Bayesian methods (Maxwell et al., 2015; Rindskopf et al., 2018), but in this article we restrict our discussion to frequentist approaches that rely on null-hypothesis significance testing (NHST) in comparing the results of two studies. We consider results from two studies only, as pairwise comparisons of replication results have been common practice. And, we focus on correspondence measures that involve NHST because they provide a clear decision criterion of whether the results of two studies have been successfully replicated or failed to replicate. This does not imply that other approaches for assessing replication success are inferior to NHST-based criteria (they actually might be more informative). However, covering additional approaches is beyond the scope of this article, except to note that they face similar issues with regard to the interpretation of replication success and the determination of replication probabilities and sample size requirements.

To address the ambiguous interpretations resulting from traditional NHST-based criteria for determining replication success, we propose a new metric for assessing replication success, called the correspondence test, which combines the test of difference and equivalence in the same framework (Steiner & Wong, 2018; Tryon & Lewis, 2008). The test has the advantage of yielding a result of statistical difference or equivalence when there is sufficient power from each study in the replication effort but it will indicate statistical indeterminacy when the replication effort of two independent studies is underpowered. In this sense, the proposed correspondence test is a more severe test (Mayo, 2018) than the difference and equivalence tests on their own because they cannot distinguish between the truth of the null hypothesis and lack of power.

In this article, we demonstrate the statistical properties of each correspondence measure, highlighting the contexts and conditions under which these measures will yield correct conclusions about replication success. To facilitate the prospective planning of replications, we derive formulas for researchers to conduct power analyses for each correspondence measure, demonstrating the critical design parameters that researchers must consider in planning sufficiently powered replication efforts. Finally, through an applied example, we consider how different criteria for replication success affect the interpretation of results in a replication effort. To this end, we reanalyze data from the Open Science Collaboration (2015) to demonstrate the differing conclusions that researchers would arrive at with different choices of correspondence measures.

In our evaluation of correspondence measures, we adopt the causal perspective of the Causal Replication Framework (Steiner et al., 2019; Wong & Steiner, 2018b) and assume that both studies estimate the *causal effect* of the same treatment-control contrast and that all causal replication assumptions are met. In particular, this implies the absence of any confounding bias, publication bias, or any questionable research practices because they would result not only in biased estimates of effects and standard errors, but also in actual Type I error rates that are presumably much larger than the nominal $\alpha$-level (Ioannidis, 2005; Kerr, 1998; McShane & Böckenholt, 2014; Simmons et al., 2011).

We also restrict our discussion to pairwise comparisons of results of two or multiple prospectively planned studies rather than the comparison of post-hoc replications with results from an already published study because the original studies are regularly insufficiently powered for assessing replications success (as will become clear later). Prospectively planned replication efforts allow for replication designs that achieve a predefined replication probability (say, of .8) and can also better address issues related to publication bias and questionable research practices because studies are conducted with regard to replication as the primary research question. Taking a prospective perspective has also implications on what counts as replication success. Unlike in post-hoc replications, a repeated demonstration of the absence of an effect is also considered a successful replication in a prospective approach. Similarly, if a test shows that two effects are equivalent (within a certain equivalence threshold), it is irrelevant whether the effects are significant.

Thus, we evaluate the properties of different correspondence measures under ideal conditions and do not consider adjustments to power calculations for post-hoc replications (Anderson and Kelley, 2022; Anderson et al., 2017; Anderson and Maxwell, 2017) or adjustments for publication bias (e.g., Andrews and Kasy, 2019; Carroll et al., 2017). This does not imply that correspondence measures cannot be used for post-hoc replication efforts, but one should keep in mind that we only discuss the measures' pure statistical properties. In the presence of confounding bias, publication bias, and questionable research practices, the measure's performance and the meaning of replication success or failure is unclear, that is, the nominal replication probabilities discussed in this article will not hold and the outcome of a single replication effort might not be meaningfully interpretable with regard to confirming scientific theories or previous evaluation results.

This article will be most useful for researchers who are prospectively planning replication studies (Wong et al., 2021), because two individual ad hoc replication studies will often be underpowered for making definitive conclusions about replication success. In our Discussion, we describe a path forward for replication designs that plan a series of replication efforts to test systematic sources of variation across studies.

## Correspondence Measures for Assessing Replication Success

Due to sampling uncertainty, it is unlikely that two studies produce identical effect estimates even if all causal and statistical assumptions needed for direct replication were met (Wong & Steiner, 2018b). Thus, researchers need to rely on decision rules for assessing the correspondence in effect estimates. A wide range of conclusion- and distance-based correspondence measures has been proposed and used in the replication literature to assess replication efforts.

The correspondence measures presented here are based on a scenario comparing two independent, perfectly implemented randomized controlled trials (RCTs) estimating two true but unknown effects, $\tau_1$ and $\tau_2$, for study 1 and 2, respectively. We assume that each RCT focuses on the same single treatment-control contrast. However, the two RCTs are allowed to vary with respect to populations, settings, and time which may result in different true effects, $\tau_1 \neq \tau_2$. The unbiased effect estimators of the true effects are denoted by $\hat{\tau}_1$ and $\hat{\tau}_2$, respectively, with sampling variances $\sigma^2_{\hat{\tau}_1}$ and $\sigma^2_{\hat{\tau}_2}$. The estimators may refer, for example, to differences in treatment and control means or covariate-adjusted regression estimators. Throughout the article we assume that the variance estimators used by researchers accurately reflect the sampling (or randomization) uncertainty. The two studies may use different methods for estimating and testing the effects. However, we assume that the effect estimators' sampling distributions are well-approximated by a normal distribution provided sample sizes are sufficiently large. Since all correspondence measures require large sample sizes to ensure adequate power, normal approximations are well justified and sufficient for planning replications in practice.[1]

Though we discuss our correspondence measures with regard to two perfectly implemented RCTs, these results directly apply to any type of independent studies, including quasi-experiments relying on regression discontinuity, difference-in-differences, matching, or instrumental variable estimators (Angrist and Pischke, 2009; Shadish et al., 2002). However, if researchers are interested in claiming equivalence or difference in *causal* effect estimates, they must assure that both studies successfully identify a causal effect and that the effect estimators and their variance estimators are both unbiased or at least consistent (for details on the Causal Replication Framework see Steiner et al., 2019; Wong and Steiner, 2018b). While these results can apply to properly identified causal effects in observational research, their application to non-causal findings, that is, conditional

[1] This might not necessarily hold for estimators relying on more complex models like random or mixed-effect models.

associations between an outcome and an independent variable of interest, is less straightforward. Replicating conditional associations requires the strong assumption that the extent of confounding between the two variables under consideration is the same in both studies (conditional on the covariates controlled for in the analyses). If this assumption is violated, replication failure is difficult to interpret because any differences in the confounding associations could have led to replication failure. Even if replication success is concluded, the causal effect and confounding structure might have varied simultaneously across studies so that the change in the causal effect is approximately compensated by the corresponding change in the confounding bias. In either case it is unclear what one can learn from successful or failed replications of conditional associations.

## Conclusion-based Correspondence Measures

Conclusion-based measures compare the conclusions drawn from two studies, that is, whether the evidence obtained allows researchers to draw the same conclusions about the presence or absence of an effect. For each study, the conclusion about an effect's presence is derived from a decision rule that may involve the sign, magnitude, or statistical significance of the effect estimate. Whether the two studies arrive at the same conclusion depends on the unknown true effect of each study, $\tau_1$ and $\tau_2$, and their respective power to detect the effect. As the true effects are never known in practice, it will become clear that the conclusion-based measures' dependence on the true effects' magnitude presents obstacles to successfully planning replication studies. Here, we consider only a single conclusion-based correspondence measure: correspondence in the statistical significance pattern. Other measures assess the correspondence in the effect estimates' sign and magnitude without considering sampling uncertainty. Since these measures are rarely used we provide a full discussion of these measures in Appendix A. For all correspondence measures, functions written in R (R Core Team, 2022) and Stata (StataCorp., 2021) are provided in the supplement.[2]

[2] We thank Steffen Erickson for providing the Stata script file.

### *Correspondence in Significance Pattern*

Comparing the pattern of statistical significance in two studies has been a common approach for evaluating replication success (e.g., Camerer et al., 2016; Gleason et al., 2018; Open Science Collaboration, 2015). This measure assesses correspondence based on the two studies' NHST outcome, that is, whether both studies consistently suggest a significant or non-significant effect (we also consider non-significant effects because we focus on prospectively planned replication studies). Correspondence in significance pattern is consistent with a replication goal of showing the *existence* of an effect as compared to demonstrating that the effects are consistent with respect to their magnitude (Anderson and Maxwell, 2016; see also Bonett, 2021). In each study, NHST is typically used to test the two-sided null hypothesis of no effect, $H_0 : \tau_k = 0$, for $k = 1, 2$. Though it is also possible to test one-sided null hypotheses of whether the estimated effect provides reliable evidence to reject the null hypothesis of an effect size less than or equal to a magnitude $\delta_M$, $H_0 : \tau_k \leq \delta_M$, we only focus on the two-sided default null hypothesis because that is what researchers regularly use when assessing correspondence in significance—despite the fact that research questions are often one-sided and that effect sizes of less than certain magnitude $|\delta_M|$ might neither be of practical nor theoretical relevance. Thus, we formalize correspondence in significance with regard to the two-sided null hypothesis $H_0 : \tau_k = 0$ with a common Type I error rate $\alpha$ for both studies. The observed effects are converted to $z$-scores, $z_k$, and compared to the critical $z$-value that corresponds to the selected $\alpha$, $z^*_{1-\alpha/2}$.

Correspondence in significance pattern, $S(\alpha)$, is indicated if either both effects have the same sign *and* both null hypotheses are rejected, $|z_k| \geq z^*_{1-\alpha/2}$, or both studies fail to reject the null hypothesis regardless of either effect's sign, $|z_k| < z^*_{1-\alpha/2}$:

$$\begin{aligned} S(\alpha) = 1[\{\text{sgn}(\hat{\tau}_1) = \text{sgn}(\hat{\tau}_2) \,\&\, (|z_1| \geq z^*_{1-\alpha/2} \,\&\, |z_2| \geq z^*_{1-\alpha/2})\} \vee \\ (|z_1| < z^*_{1-\alpha/2} \,\&\, |z_2| < z^*_{1-\alpha/2})], \end{aligned} \tag{1}$$

where sgn(.) is the sign function and 1[.] the indicator function that returns replication success if the logical expression is *true* and replication failure if it is *false* (the ampersand,

&, is the logical *and* and the symbol $\vee$ the logical *or*). Thus, the probability of a successful replication is a function of each effect estimate's probability of being significant. Assuming that the effect estimators' sampling distributions are asymptotically normal with expectations $\tau_1$ and $\tau_2$ and variances $\sigma^2_{\hat{\tau}_1}$ and $\sigma^2_{\hat{\tau}_2}$, the replication probability is given by

$$\begin{aligned} &P(S(\alpha) = 1|\tau_1, \tau_2, \sigma_{\hat{\tau}_1}, \sigma_{\hat{\tau}_2}, \alpha) = \\ &\qquad [1 - \Phi(\zeta_1^+)][1 - \Phi(\zeta_2^+)] + [\Phi(\zeta_1^-)][\Phi(\zeta_2^-)] + \\ &\qquad [\Phi(\zeta_1^+) - \Phi(\zeta_1^-)][\Phi(\zeta_2^+) - \Phi(\zeta_2^-)], \end{aligned} \tag{2}$$

with $\zeta_k^+ = z^*_{1-\alpha/2} - \frac{\tau_k}{\sigma_{\hat{\tau}_k}}$, $\zeta_k^- = z^*_{\alpha/2} - \frac{\tau_k}{\sigma_{\hat{\tau}_k}}$ and $\Phi$ the standard normal distribution function (see also Schauer & Hedges, 2021). The proofs of this and all further results are provided in Appendix B.

The formula for the replication probability indicates that the unknown true effects $\tau_k$ and the estimable variances $\sigma^2_{\hat{\tau}_k}$ directly affect the replication probabilities. Note that, for correspondence in significance, the replication probability depends on the *true* effect because it is the true effect size that impacts the probability of a test being significant. In situations where both studies have large true effects, correspondence in significance pattern is more likely than in situations with smaller effects unless both studies are sufficiently powered. When the true effects are medium-sized or small, the study-specific test outcomes have a higher chance of contradicting each other. Also, each study's statistical power to detect the true effect impacts whether correspondence in significance pattern can be established. Studies with sufficient power to detect the true effects may both find significant effect estimates and thus indicate correspondence. However, studies that lack power may also suggest correspondence even if the two effects are dissimilar, as both results are likely to be non-significant. This is one drawback of correspondence in significance pattern, but can be resolved by defining correspondence with respect to significant outcomes only. Since we take a prospective point of view in planning replications, we do not further consider this option.[3]

[3] In practice, however, replication efforts using correspondence in significance pattern often begin with an

The main drawback of conclusion-based measures like correspondence in significance pattern is the replication probability's dependence on the true but unknown effects. As we will see, this makes it difficult to plan and sufficiently power replication efforts so that a predetermined replication probability will be achieved (even if the true effects of both studies can be assumed to be identical). A possibility to overcome these issues is to define correspondence with regard to the distance (or difference) in true effects.

### Distance-based Correspondence Measures

Distance-based measures use NHST to evaluate whether the distance between two effect estimates, $\hat{\tau}_1 - \hat{\tau}_2$, differs from 0. This provides evidence for a significant difference or equivalence of the true effects $\tau_1$ and $\tau_2$, respectively. Thus, unlike the conclusion-based measures, distance-based measures apply formal NHST to the distance in effect estimates of the two replication studies. We consider three tests: (1) the standard NHST for the effect difference, to which we refer as the *difference test*; (2) the *equivalence test*; and (3) the *correspondence test*, which combines the outcomes of the difference and equivalence test.

A key difference compared to the conclusion-based measures is that the replication probabilities of the three tests do *not* depend on the magnitude of the true but unknown effects $\tau_1$ and $\tau_2$. Rather, these measures depend only on the difference between the two effects and the estimator variances. For example, distance-based measures have the same probability to detect correspondence when the true effects are $\tau_1 = 0.1$ and $\tau_2 = 0.3$ SD or when they are $\tau_1 = 0.9$ and $\tau_2 = 1.1$ SD because the differences and thus test statistics are identical (provided the estimator variances are the same for both situations). Whether one effect is positive, $\tau_1 = 0.1$, and the other negative, $\tau_2 = -0.1$, does not make a difference either—only the distance matters. Thus, replication success is possible even if effects are of

existing, significant effect and attempt to directly replicate this effect in a new study. In a post-hoc replication study, given that a significant positive effect has already been found, the probability of showing correspondence is simply the probability that the replication effort will show a positive, significant effect: $1 - \Phi(z^*_{1-\alpha/2} - \frac{\tau_2}{\sigma_{\hat{\tau}_2}})$.

opposite sign or different significance pattern. If researchers find such a situation unsatisfactory, one can combine conclusion-based measures about the sign or significance of effects with distance-based measures into a new correspondence measure and then determine the corresponding replication probabilities. The properties of such a composite measure will be a combination of its constituent correspondence measures, but the replication probability would again depend on the magnitude of the unknown true effects. However, we do not discuss such combinations of measures further.

### ***Difference test***

The difference test compares the estimates of two independent studies, $\hat{\tau}_1$ and $\hat{\tau}_2$. The test is consistent with a replication goal of showing whether two effects are inconsistent with each other (Anderson and Maxwell, 2016; see also Bonett, 2021) and is implemented as a $z$-test based on a normal approximation (unknown variances $\sigma^2_{\hat{\tau}_1}$ and $\sigma^2_{\hat{\tau}_2}$ are estimable from the observed data of the two studies). The null hypothesis of the test claims that the difference between the true effects from the studies is zero versus the alternative that the difference is non-zero: $H_0 : \tau_1 - \tau_2 = 0$ vs $H_1 : \tau_1 - \tau_2 \neq 0$. The null hypothesis is tested using the $z$-test statistic, $z_{DT} = (\hat{\tau}_1 - \hat{\tau}_2)/\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}$. The difference test, $DT(\alpha_R)$, suggests correspondence of effects if a non-significant result is obtained, which occurs when the absolute value of the observed test statistic, $|z_{DT}|$, is smaller than the critical value $z^*_{1-\alpha_R/2}$ for a given Type I error rate $\alpha_R$ (the subscript $R$ in $\alpha_R$ is used to distinguish the Type I error rates of the replication tests from the Type I error $\alpha$ of effect tests conducted for each single study):

$$DT(\alpha_R) = 1[|z_{DT}| < z^*_{1-\alpha_R/2}] \tag{3}$$

Alternatively, correspondence can be defined using the test's $p$-value and the pre-determined $\alpha_R$ level. The outcome of the difference test, $DT(\alpha_R)$, depends both on the difference between the two estimated effects and their corresponding variance.

The replication probability of the difference test, that is, the probability that the null hypothesis of no difference in effects cannot be rejected, is given by

$$P(DT = 1|\tau_1, \tau_2, \sigma_{\hat{\tau}_1}, \sigma_{\hat{\tau}_2}, \alpha_R) = \Phi\left(z^*_{1-\alpha_R/2} - \frac{\tau_1 - \tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}}\right) - \Phi\left(-z^*_{1-\alpha_R/2} - \frac{\tau_1 - \tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}}\right). \tag{4}$$

This expression is equivalent to the Type II error of the $z$-test for two independent samples, with a probability of $P(DT = 1) = 1 - \alpha_R$ if the effect difference is zero. The probability of a successful replication decreases as the difference in true effects increases or the estimator variances decrease.

However, one cannot infer correspondence between two estimated effects based solely on a non-significant difference test because it is impossible to distinguish between the scenario in which the null hypothesis is true and that in which the alternative is true but the test has insufficient power to detect a difference. The difference test is able to provide evidence regarding the lack of correspondence between effects, which sometimes might be the goal of a replication effort. But a non-significant difference test provides only inconclusive evidence (Anderson & Maxwell, 2016; Bonett, 2021). On its own, the difference test is thus not able to demonstrate the equivalence of effects. Equivalence tests, which we discuss next, are designed to test for the equivalence rather than the difference of two effects.

### ***Equivalence test***

While the difference test examines the null hypothesis that the two true effects are identical, the equivalence test examines the null hypothesis that the absolute difference in the two effects from the two studies is equal to or exceeds a predefined equivalence threshold $\delta_E$, $H_0 : |\tau_1 - \tau_2| \geq \delta_E$ (Rogers et al., 1993; Tryon, 2001; Tryon & Lewis, 2008). To establish equivalence of two effects, the null hypothesis must be rejected; when the null hypothesis is rejected, the test provides evidence that the effect of interest has been replicated (Bonett, 2021). Therefore, the equivalence test avoids the ambiguity of non-significant difference tests. It is aligned with the replication goal of showing whether two effects are consistent with each other (Anderson and Maxwell, 2016; see also Bonett, 2021). Equivalence is only established if the replication effort has sufficient power to reject the null hypothesis of an effect difference.

Equivalence testing requires the determination of a pre-specified equivalence threshold, $\delta_E > 0$, that is, the maximum difference in effects one considers negligibly small or inconsequential (Bonett, 2021; Steiner & Wong, 2018; Tryon, 2001; Tryon & Lewis, 2008). For instance, if researchers believe that an effect size of about 0.5 SD is needed for both studies to be meaningful (from a theoretical or practical perspective), then an effect difference of 0.1 SD or even 0.2 SD might be considered negligibly small. If researchers believe that an effect size of about 0.2 SD is needed to be meaningful, a threshold of 0.2 SD seems too large but 0.1 SD might be acceptable. In any case, it is advisable to avoid too large equivalence thresholds as equivalence could be established for non-negligible effects of opposite sign.[4] As will become clear later, the choice of a specific equivalence threshold implies that both studies must be powered with respect to a minimum detectable effect size that is smaller than the threshold.

The equivalence test, $ET(\delta_E, \alpha_R)$, is implemented as two one-sided tests (Anderson & Maxwell, 2016; Schuirmann, 1987; Tryon, 2001) each with a nominal Type I error rate $\alpha_R$. The first test examines the hypothesis pair with the positive threshold: $H_0^+ : \tau_1 - \tau_2 \geq \delta_E$ vs $H_1^+ : \tau_1 - \tau_2 < \delta_E$. The second test probes the corresponding hypothesis pair with the negative threshold: $H_0^- : \tau_1 - \tau_2 \leq -\delta_E$ vs $H_1^- : \tau_1 - \tau_2 > -\delta_E$.

[4] To avoid an equivalence outcome when the two effects are of opposite sign or show different significance patterns one could combine correspondence in sign or significance pattern with the equivalence test into a new correspondence metric with its own formula for the replication probability. Then, replication success would be achieved if both effect estimates have the same sign or significance pattern and if the null hypothesis of a difference greater than the equivalence threshold is rejected. For such a measure, the replication probabilities depend again on the magnitude of the unknown true effects.

Assuming asymptotic normality, the corresponding test statistics for the two tests are

$$z_{ET^+} = \frac{(\hat{\tau}_1 - \hat{\tau}_2) - \delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}} \quad \text{and} \quad z_{ET^-} = \frac{(\hat{\tau}_1 - \hat{\tau}_2) + \delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}}. \tag{5}$$

Correspondence in equivalence is then given if both one-sided null hypotheses are rejected. Thus, for the significance level $\alpha_R$, the two observed test statistics, $z_{ET^+}$ and $z_{ET^-}$, must be more extreme than the corresponding critical values $z^*_{1-\alpha_R}$ and $z^*_{\alpha_R}$ :

$$ET(\delta_E, \alpha_R) = 1[z_{ET^+} \leq z^*_{\alpha_R} \;\&\; z_{ET^-} \geq z^*_{1-\alpha_R}]. \tag{6}$$

The replication probability of establishing correspondence in equivalence between the two effect estimates is then obtained as the probability that both one-sided hypotheses are rejected (see also Chow et al., 2008),

$$\begin{aligned} &P(ET = 1 | \tau_1, \tau_2, \sigma_{\hat{\tau}_1}, \sigma_{\hat{\tau}_2}, \alpha_R, \delta_E) = \\ &\Phi\left(-z^*_{1-\alpha_R} - \frac{(\tau_1 - \tau_2) - \delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}}\right) - \Phi\left(z^*_{1-\alpha_R} - \frac{(\tau_1 - \tau_2) + \delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1} + \sigma^2_{\hat{\tau}_2}}}\right). \end{aligned} \tag{7}$$

Unlike the difference test, the equivalence test provides direct evidence as to whether the effects of two studies are equivalent with respect to the threshold $\delta_E$. As with the difference test, the magnitude of the difference between the true effects and the estimator variances impact the probability of detecting an effect. However, choosing an adequate threshold $\delta_E$ for the equivalence test is crucial, as it directly impacts the replication probability. The smaller the threshold, the harder it is to establish equivalence because the two true effects must be rather similar and the variances of both effect estimators must be small to reject the two one-sided null hypotheses (i.e., large sample sizes are required). Further, if a single study's effect estimate has a large variance (e.g., due to a small sample size), the replication probability may be close to 0 regardless of how small the other study's estimator variance is. Also note that a non-significant equivalence test does not necessarily indicate that the two true effects differ by more than $\delta_E$. It is also possible that the null hypothesis of a difference could not be rejected due to a lack of power. Thus, to overcome this limitation of the equivalence test, it makes sense to combine the difference and equivalence test together into a *correspondence test*.

### ***Correspondence test***

Both the difference and equivalence test can provide information as to whether two studies succeeded in replicating an effect. However, each test provides limited information on its own. A significant difference test suggests that the two effects are different, while a significant equivalence test suggests that the two effects do not differ by more than $\delta_E$. However, if both tests are non-significant, no reliable conclusions about a difference or equivalence can be drawn because both tests lacked sufficient power to provide evidence for a difference or equivalence.

To capture the possibility of both tests being inconclusive, the two tests can be combined into a correspondence test (Steiner & Wong, 2018; Tryon & Lewis, 2008). Such a correspondence test allows researchers to make a more nuanced inference regarding replication success or failure based on whether the null hypothesis of each test can or cannot be rejected. As a combination of a difference and equivalence test, the correspondence test simultaneously addresses two replication goals: the demonstration either of the consistency or inconsistency between two effects (Anderson & Maxwell, 2016).

The correspondence test, $CT(\delta_E, \alpha_R)$, has four possible outcomes:

$$CT(\delta_E, \alpha_R) = \begin{cases} \textit{Equivalence } (EQU) & \text{if } DT(\alpha_R) = 1 \,\&\, ET(\delta_E, \alpha_R) = 1 \\ \textit{Difference } (DIF) & \text{if } DT(\alpha_R) = 0 \,\&\, ET(\delta_E, \alpha_R) = 0 \\ \textit{Trivial Difference } (TRI) & \text{if } DT(\alpha_R) = 0 \,\&\, ET(\delta_E, \alpha_R) = 1 \\ \textit{Indeterminacy } (IND) & \text{if } DT(\alpha_R) = 1 \,\&\, ET(\delta_E, \alpha_R) = 0 \end{cases} \tag{8}$$

The test returns *Equivalence* with respect to the equivalence threshold $\delta_E$ if the difference test is non-significant (suggesting correspondence), $DT(\alpha_R) = 1$, and the equivalence test is significant, $ET(\delta_E, \alpha_R) = 1$. In this case, the equivalence test has sufficient power to establish *Equivalence* while the difference test does not indicate a significant effect difference.

Conversely, to establish a *Difference* between the two effects, a significant difference test, $DT(\alpha_R) = 0$, and a non-significant equivalence test, $ET(\delta_E, \alpha_R) = 0$, are required. That is, both tests suggest a lack of correspondence between the two effects. A *Difference* can only be established if the difference test has sufficient power to reject the null hypothesis of equivalence.

The third possible outcome of the correspondence test is a *Trivial Difference* in effects, that is, the difference test is significant and indicates non-correspondence, $DT(\alpha_R) = 0$, while the equivalence test is also significant and suggests correspondence with regard to the given threshold $\delta_E$, $ET(\delta_E, \alpha_R) = 1$. This means that while there is evidence that the true effects differ, they differ by an amount smaller than the equivalence threshold. Such a result most likely occurs when both the difference and equivalence tests are highly powered. But note that the power of the equivalence test might be due to the choice of a large threshold $\delta_E$.

The final possible result is *Indeterminacy*, which occurs when neither the equivalence nor difference test produces a significant result, that is, $DT(\alpha_R) = 1$ and $ET(\delta_E, \alpha_R) = 0$. Such a test outcome is fairly uninformative because inadequate power resulted in a failure to reject the null hypothesis of both the difference and equivalence test. In this circumstance, neither a reliable difference nor equivalence (within the threshold $\delta_E$) is established. A result of indeterminacy means that additional replications are necessary to determine whether the true effects correspond or not.

By differentiating between these four possible outcomes, the correspondence test improves upon both of its constituent tests. Specifically, it does not confound insufficient power with a result of substantive interest, that is, the two effects being equivalent or different. Thus, the correspondence test provides stronger evidence where the equivalence or difference tests alone could not differentiate between their respective null hypotheses and insufficient power. In this sense, it is a more *severe* test than its constituent tests because the correspondence test clearly indicates when claims regarding replication success or failure are warranted or unwarrented due to lack of power (Mayo, 2018).

Because $CT$ is based on the difference and equivalence test, the probabilities for the four possible outcomes of the correspondence test are directly linked to the replication probability of $DT$ and $ET$. To determine the replication probabilities we need to distinguish between two cases. First, where correspondence from the equivalence test is less likely than correspondence from the difference test, $P(ET = 1) \leq P(DT = 1)$, and second, where it is more likely, $P(ET = 1) > P(DT = 1)$. The first condition can also be written as $P(DT = 0) + P(ET = 1) \leq 1$, which implies that the critical regions of the difference and equivalence test do not overlap and thus leave space for *Indeterminacy*, while the outcome of a *Trivial Difference* becomes impossible because it requires that both tests are simultaneously significant. The second condition is equivalent to $P(DT = 0) + P(ET = 1) > 1$, which indicates a situation where the critical regions of the two tests overlap such that a simultaneous significance of both the difference and equivalence test becomes possible (*Trivial Difference*), while *Indeterminacy* is no longer possible and thus has a probability of zero. In practice, the second condition is unlikely to occur unless large equivalence thresholds are used or sample sizes are huge. This second condition which allows for a *Trivial Difference* becomes possible only because the difference test does not us an equivalence threshold. If the difference test were conducted with regard to the null hypothesis $H_0 : |\tau_1 - \tau_2| \leq \delta_E$, then $P(ET = 1) \leq P(DT = 1)$ of the first condition would always hold (but the indeterminacy region would become much larger because small effect differences would no longer be detected). For more detailed explanations and a visualization of the outcomes of the correspondence test, see the proof in Appendix B.

Table 2 lists the replication probabilities for all four outcomes of the two possible cases. Note that for both cases, the four probabilities sum to one. Though the correspondence test has four possible outcomes, only *Equivalence* is a clear replication success, though one may also consider a *Trivial Difference* as a replication success provided that the threshold $\delta_E$ is sufficiently small.

## Replication Probabilities of Correspondence Measures

As indicated by the correspondence measures' probability formulas, the replication probabilities depend on several parameters and thresholds ($\tau_1$, $\tau_2$, $\sigma_{\hat{\tau}_1}$, $\sigma_{\hat{\tau}_2}$, $\alpha$, $\alpha_R$, $\delta_M$, $\delta_E$). To produce generally valid probability plots that also facilitate the comparison of all correspondence measures discussed in this article, we assume that the true effects, the effect estimators' variances, and the thresholds refer to standardized effect size measures where the effect sizes of both studies have been standardized by the same standard deviation (SD) of the outcome variable (otherwise, the effects would not be comparable). Moreover, in each probability formula, we express the standard errors $\sigma_{\hat{\tau}_k}$ in terms of their study-specific minimum detectable effect size (*MDES*). The *MDES* is the smallest effect for which the null hypothesis of no effect is rejected with probability $1-\beta$ for given Type I and II error rates and the standard error of the effect estimate:

$$MDES_k = (\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta))\sigma_{\hat{\tau}_k} \tag{9}$$

for study $k = \{1, 2\}$, where $\Phi^{-1}(1-\alpha/2)$ is the normal quantile corresponding to the selected Type I error rate $\alpha$ and $\Phi^{-1}(1-\beta)$ is the normal quantile corresponding to the selected Type II error rate $\beta$. Thus, the standard error $\sigma_{\hat{\tau}_k}$ can be expressed in terms of the *MDES*:

$$\sigma_{\hat{\tau}_k} = \frac{MDES_k}{(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta))}. \tag{10}$$

For given Type I and II error rates, smaller *MDES*s imply smaller standard errors and thus greater power for significance tests. Since we use a default Type I error rate of $\alpha = .05$ and power of $1-\beta = .8$ for both studies, the standard errors are given by $\sigma_{\hat{\tau}_k} \approx MDES_k/2.8$. These are deliberate though common choices for $\alpha$ and $1-\beta$, but different choices only result in scale shifts while the shape of the plots remain the same. Using standardized measures and *MDES*s allows for a description of the correspondence measures' replication probabilities that does not depend on the scale of the outcome variables and directly relates the unknown true effects—the target of the replication effort—to the *MDES*s. Moreover, using *MDES*s also helps in planning sufficiently powered replication studies (see next section). It is important to note that when we discuss *MDES* in this context, we refer to the *MDES* as a rescaled standard error that is necessary for both studies to attain a specific replication probability for a given correspondence measure. This should not be confused with the target *MDES* when determining sample size requirements in a power analysis to detect the smallest effect of substantive interest in a single study.

The following probability plots are all presented in terms of ratios between the unknown true effects and the studies' *MDES*, $\tau_k/MDES_k$, or the ratios of the effect differences and the average $MDES^*$ across the two studies $|\tau_1 - \tau_2|/MDES^*$ (defined below). The replication probabilities are fully determined by these ratios. For instance, the replication probability of the difference test is the same for an effect difference of 0.1 SD and an *MDES* of 0.2 SD for both studies, and a difference of 0.2 SD and *MDES*s of 0.4 SD—in both scenarios we obtain the same difference ratio of $|\tau_1 - \tau_2|/MDES = 0.5$. Similarly, the equivalence threshold is also expressed in relation to *MDES*. Using such ratios allows us the represent all possible scenarios (with fixed Type I and II error rates) in a single plot rather than having separate plots for different choices of true effects, *MDES*s, and thresholds. These ratios also make intuitive sense. For instance, if one is conducting an RCT in which the true effect of interest is $\tau = 0.5$ SD, one would need to power the study with an *MDES* of 0.5 SD or lower to detect the effect with a probability of at least $1 - \beta$. This implies an effect ratio of $\tau/MDES \leq 1$. If the *MDES* is much lower than the true effect, say 0.1 SD, the effect ratio of $\tau/MDES = 0.5/0.1 = 5$ indicates more than sufficient power to detect the effect. However, with an *MDES* of 0.75 SD, the effect ratio of $\tau/MDES = 0.5/0.75 = 0.67$ suggests a poorly powered study. Thus, effect ratios of greater than 1 indicate properly powered individual studies. But as we will see, two properly powered individual studies might still be poorly powered for achieving a sufficiently high replication probability. R functions for computing the replication probabilities are provided in the supplement.

### Conclusion-based Measures: Correspondence in Significance Pattern

Figure 1 shows the replication probabilities for correspondence in significance pattern, that is, $P(S(\alpha) = 1)$. Probabilities are shown as functions of $\tau_k/MDES_k$, for $k = 1, 2$, on the $x$- and $y$-axis, respectively. The scales for these ratios range from -2 to 2, accounting for the possibility that the true effects may have different signs. Note that $MDES_k$ is always assumed to be positive because it is used as an alternative expression for the standard error. Thus, negative values indicate negative true effects ($\tau_k$). The contour lines within the plot show different levels of replication probabilities, with darker shading indicating a higher probability of replication success. The gray scale becomes significantly darker once a replication probability of .8 is exceeded. We consider a probability of .8 to be the minimum for serious replication efforts.

The plot in Figure 1 shows three areas with high replication probabilities. The first two refer to situations where both studies' effect ratio exceeds at least 1 (for positive effects) or falls below -1 (for negative effects). That is, both studies are sufficiently powered with regard to the unknown true effects. But note that two studies with an effect ratio of $\tau_k/MDES_k = 1$, which is sufficient to detect the true effect with a probability of .8, only achieve a replication probability of .68. The third area with high replication probabilities is located at the center of the plot where $\tau_k/MDES_k = 0$. In this case, both studies lack power to detect an effect and likely fail to reject the null hypothesis of no effect.

Replication failure is most likely when the true effects of the two studies have opposite signs or when the effect ratio of study 1 falls below 1, $\tau_1/MDES_1 < 1$, while the other study has a ratio greater than 1, $\tau_2/MDES_2 > 1$ (or vice versa; analogous for negative effects). That is, while study 1 most likely fails to reject the null hypothesis of no effect, study 2 will indicate a significant effect with high probability.

It is clear that researchers must have a reasonable guess about the true effect for both studies in order to prospectively plan for a specific replication probability. Otherwise, it is possible that one of the studies will not be sufficiently powered, making replication success unlikely. While knowing the effect estimate of study 1 before planning study 2 might be helpful, the assumption that the true effects are identical across the two studies, even when treatments, populations, and settings are held constant, may be incorrect and result in a replication failure. Moreover, the effect estimate of the first study should never be directly used as the target *MDES* for the second study even if the underlying true effects were identical (Anderson & Maxwell, 2017).

While correspondence in significance pattern is not an ideal measure of replication success, there are some circumstances where it is suitable. For instance, it can be a suitable measure if a replication question primarily concerns whether the same decision would be made to implement a policy. A major drawback of correspondence in significance pattern is that replication success can be achieved even if the true effects of the two studies are extremely different. This is so because the replication probability is determined by the effect ratios rather than the magnitude of the effects. For instance, assume that study 1 has an underlying true effect of $\tau_1 = 0.25$ SD and study 2 an effect of $\tau_2 = 1$ SD, and both are sufficiently powered with an effect ratio of $\tau_k/MDES_k = 1.5$. We then obtain a replication probability of .98 despite the heterogeneity in effect sizes.

### Distance-based Correspondence Measures

Just as the correspondence in significance pattern depends on ratios involving the true effects, all of the distance-based measures depend on the ratio of the difference in effects to the *MDES*, $(\tau_1 - \tau_2)/MDES$. The equivalence and correspondence tests also depend on the corresponding ratio of the equivalence threshold, $\delta_E/MDES$. For the probability plots of distance-based measures it is sufficient to cover only positive effect differences (or the absolute value of the difference) and positive equivalence thresholds because they are symmetric with respect to both the difference and threshold. The scales of the $x$- and $y$-axis range from 0 to 3. While the replication probabilities of the conclusion-based measures depend on each studies' effect ratio, the probabilities of distance-based measures can be shown to depend on the *average MDES* of the two studies, that is,

$$MDES^* = \sqrt{\frac{MDES_1^2 + MDES_2^2}{2}}. \tag{11}$$

The average of the squared *MDES*s has to be considered because the *MDES* is used as an alternative expression for the standard error (i.e., this is akin to averaging variances). Using the average $MDES^*$ simplifies our plots without any loss of generality, because once researchers determine the *MDES* for both studies, the average $MDES^*$ and associated replication probability can be determined accordingly. The same applies when powering individual studies to achieve a certain replication probability. All *MDES* combinations of two studies have identical replication probabilities as long as their average $MDES^*$ is the same.

### ***Difference Test***

Figure 2 shows the replication probability for the difference test, $P(DT = 1)$. The replication probability is determined solely by the single ratio $|\tau_1 - \tau_2|/MDES^*$. Thus, its replication probability can be described by a single curve. The replication probability peaks at $1 - \alpha = .95$ when there is no difference in effects (implying a ratio of 0), and converges to zero for large ratios. The shape of the plot is identical to the Type II error plot of the two-sample $z$-test.

The plot indicates that a sufficiently high replication probability, $P(DT = 1) > .8$, is achieved only with ratios $|\tau_1 - \tau_2|/MDES^* < 0.563$. Thus, whenever the effect difference, $|\tau_1 - \tau_2|$, is clearly smaller than the average $MDES^*$ replication success is very likely. This is achieved when the effect difference is negligibly small or when at least one of the two studies is poorly powered—a large *MDES* of one study strongly affects the average $MDES^*$ even when the other study's *MDES* were close to zero. For this reason, the difference test is a poor choice for assessing correspondence in two effect estimates.

### ***Equivalence Test***

The replication probability for the equivalence test, $P(ET = 1)$, is shown in Figure 3 and depends on two ratios: one for the effect difference $|\tau_1 - \tau_2|/MDES^*$ ($x$-axis) and the other for the equivalence threshold $\delta_E/MDES^*$ ($y$-axis). The plot shows that the equivalence test has a high probability ($> .8$) of indicating correspondence in effect estimates if the effect difference is close to zero and the equivalence threshold is at least 1.48 times larger than the $MDES^*$ (i.e., $\delta_E/MDES^* > 1.48$).

Under the ideal equivalence scenario where $\tau_1 - \tau_2 = 0$, the threshold ratio of $\delta_E/MDES^* = 1.48$ implies that both studies' *MDES*s must be very small to achieve a replication probability of .8. The average $MDES^*$ must be less than $\delta_E/1.48$. For instance, an equivalence threshold of $\delta_E = 0.3$ SD demands an average $MDES^*$ that is less than $0.3/1.48 = 0.203$ SD—an *MDES* not attained by many studies in practice. Moreover, if one study has a large *MDES*, for instance due to a small sample size or unexpectedly large error variance, then the second study will be unable to bring the average $MDES^*$ down to a sufficiently small size (see also Bonett, 2021). Consider for example a threshold of $\delta_E = 0.5$ and an $MDES_1 = 1$ (which may be sufficient to demonstrate an effect size of 1 SD or greater in this single study), then even with an *MDES* close to zero for the second study, $MDES_2 \approx 0$, the smallest possible average $MDES^*$ is 0.71. The corresponding threshold ratio of $\delta_E/MDES^* = 0.5/0.71 = 0.7$ then immediately suggests that the replication probability is essentially zero. Thus, establishing correspondence in equivalence requires that both studies have a small *MDES*, implying that samples sizes generally must be large.

However, if the true effects differ by more than the chosen equivalence threshold, $|\tau_1 - \tau_2| > \delta_E$, the replication probability is always less than $\alpha = .05$. Thus, the equivalence test has a high probability of indicating replication failure if the effects actually differ. But replication failure might also be due to insufficient power. The correspondence test is able to distinguish between these two scenarios as the next section demonstrates.

### *Correspondence Test*

Figure 4 shows the probability distribution across the four outcomes of the correspondence test based on the effect difference ratio $|\tau_1 - \tau_2|/MDES^*$ ($x$-axes) and threshold ratio $\delta_E/MDES^*$ ($y$-axes). The probabilities across the four outcomes sum to 1 (i.e., if we overlay the four plots, the probabilities add up to 1). Adding the probabilities of the *Equivalence* and *(Trivial) Difference* outcomes in the top two plots in Figure 4 results in the probability plot for the equivalence test.

For a more compact representation in a single plot, Figure 5 shows for each outcome the area where its probability exceeds .5 and thus is more likely than any other of the three outcomes. According to this plot, *Equivalence* is the most probable outcome of the correspondence test only if the effect difference ratio is less than 0.99 and the threshold ratio exceeds at least 1.17.

Whenever $|\tau_1 - \tau_2|/MDES^* > 0.99$, the correspondence test most likely indicates a *Difference* or *Trivial Difference*, the latter being more likely for large equivalence thresholds and smaller effect differences. Finally, *Indeterminacy* is the predominant outcome of the correspondence test whenever the effect difference ratio is less than 0.99 and the threshold ratio below about 1.2 (Figure 4). *Indeterminacy* represents a situation that most likely occurs in current replication practices because most studies' *MDES* is not small enough to either demonstrate *Equivalence* or a *(Trivial) Difference*.

Though the threshold ratio can be controlled by researchers, thresholds greater than the actual studies' *MDES*s are rarely desirable. Consider a replication effort with an average $MDES^*$ of 0.5 SD and the presumption that the effect difference is less than 0.5 SD, implying an effect difference ratio of less than one, $|\tau_1 - \tau_2|/MDES^* < 1$, and thus, a rather low probability ($< .5$) for a *(Trivial) Difference.* Then, to avoid the undesirable *Indeterminacy* outcome and to demonstrate *Equivalence*, researchers would need to choose an equivalence threshold greater than 0.6 or 0.7 SD to obtain a threshold ratio greater than 1.2 or 1.4, respectively. Such high equivalence thresholds generally do not make sense regardless of the *MDES*$^*$, especially so when studies are powered to demonstrate effect sizes smaller than the threshold (such as 0.5 SD). Thus, the *MDES* of each study is ideally smaller than the maximum tolerable threshold $\delta_E$.

The plot in Figure 6 builds on Figure 5 by showing the area where most current replication efforts are presumably located (bounded by the solid line)—in or just outside of the uninformative *Indeterminacy* zone, provided the underlying true effects are not too different. Increasing the chances for establishing *Equivalence* or a *(Trivial) Difference* requires smaller *MDES*s with respect to a power of $1 - \beta = .8$ or, equivalently, powering the desired study-specific *MDES*s with $1 - \beta = .9$ or even .95. In practice, this can only be achieved by reducing sampling variances (e.g, by increasing sample sizes or using better experimental control of extraneous noise factors, blocking, covariate adjustments, or a careful control of study settings and populations). But even if replication efforts would power their *MDES*s with $1 - \beta = .9$ or .95 for a replication probability of .8, the area outside the *Indeterminacy* zone does not become much larger—as shown in Figure 6 by the dashed and dotted boundaries. This suggests that replication efforts with two, not prospectively planned studies will often fail to successfully establish *Equivalence* or a *(Trivial) Difference*, even if the true effects are almost identical.

## Determining Minimum Detectable Effect and Sample Sizes for Replication

Achieving a replication probability of at least .8 requires two well-planned, highly powered studies. Otherwise, it is unlikely that any tests conducted to determine replication success or failure will help in establishing stable knowledge and advancing subject matter theory. Thus, it is vital to prospectively plan replication studies to ensure a reasonable chance of demonstrating replication success.

The formulas for the replication probabilities we presented above can be used to determine the required sample sizes such that a desired minimum replication probability $p_R$ for the chosen correspondence measure is achieved. In this section, we demonstrate how the *MDES* and the corresponding minimum required sample size for each of the two studies can be determined. Compared to determining the required samples sizes for a single study, where the *MDES* is chosen based on researchers' intent to demonstrate a minimum effect size of interest, the choice of a specific correspondence measure and the desired replication probability directly dictates each study's *MDES* and thus sample size. To highlight *MDES*'s dependence on the desired replication probability $p_R$, we use $MDES_R$ to represent the necessary minimum detectable effect size to achieve the pre-specified replication probability. Note again, we use the *MDES*s as rescaled standard errors for the effect estimators. Thus, if replication success should be established with a certain probability, researchers cannot choose an *MDES* they consider as meaningful or important—the *MDES*s are fixed once the desired replication probability has been determined. The obtained $MDES_R$ might seem small for a single study to demonstrate an effect of a certain magnitude but this is irrelevant when the research question concerns demonstrating replicability of results across two studies. Thus, it is necessary to power both studies such that they *jointly* provide conclusive evidence about the presence or equivalence of effects. Since the presence of an effect has to be demonstrated twice, sample size requirements are rather demanding, particularly if one wants to establish equivalence within a small equivalence threshold.

Determining the minimum required sample size for each study involves the following steps:

1. For each study, determine the effect estimator and its variance estimator to be used in the analysis of the data.

2. Determine the study-specific Type I and II error rates $\alpha$ and $\beta$ and their corresponding $z$-values ($z_{1-\alpha/2}$, $z_{1-\beta}$).

3. Choose a correspondence measure and determine

   - the desired replication probability ($p_R$),

- depending on the chosen measure, the magnitude of true effects ($\tau_k$) or threshold $\delta_E$, and
- the Type I and II error rates for distance-based replication measures ($\alpha_R$, $\beta_R$).

4. Compute the required $MDES_R$ for the chosen correspondence measure according to the formula presented below.
5. Translate the $MDES_R$ into standard error units, $\sigma_{\hat{\tau}_k} = \frac{MDES_R}{z_{1-\alpha/2}+z_{1-\beta}}$ (for the standardized effect size), and derive the necessary sample sizes using the variance estimator from Step 1.

Before we use an example to demonstrate the computation of the required sample sizes, we present the $MDES_R$ formulas for each correspondence measure. We discuss the determination of $MDES_R$ and sample sizes for two prospectively planned studies, that is, they are planned jointly in advance but might still be implemented by different research teams.[5] The $MDES_R$ calculations are done under the hypothesis that the true effects are identical across both studies, $\tau_1 = \tau_2$, and that both studies are planned to have the same $MDES_R$. These restrictions are required to ensure that all $MDES_R$ formulas are analytically solvable. In practice, the study-level true effects will likely be different, which may cause the resulting studies to be underpowered depending on the method used to assess correspondence. Nonetheless, powering replication efforts to demonstrate a presumed equivalence of the underlying true effects is the goal of most replication efforts, thus it also makes sense use the same $MDES_R$. However, this does not imply that the two studies must have the same sample sizes because they might differ with regard to the expected sampling variance (for instance, due to different sampling or blocking strategies, or because different

[5] Post-hoc $MDES_R$ computations for the second study, after the effect estimates of the first study are available, pose additional difficulties such as imprecise effect estimates and publication bias. These difficulties arise even if one were to assume that the true effects are identical across the two studies (Anderson & Kelley, 2022; Anderson & Maxwell, 2016; Maxwell et al., 2015)

estimators are used). For the computation of the needed $\mathit{MDES}_R$, functions written in R (R Core Team, 2022) and Stata (StataCorp., 2021) are provided in the supplement. The script files also contain a completely worked-out example for determining the minimum required sample sizes (the example follows the setup discussed below).

### Conclusion-based Measures: Correspondence in Significance Pattern

The determination of a study-specific $\mathit{MDES}_R$ and the required sample sizes for correspondence in significance pattern requires researchers to first specify the magnitude of the presumed true effects ($\tau_1 = \tau_2 = \tau$). The formula given in Equation (2) is not analytically solvable for the $\mathit{MDES}_R$, but if the significance test being conducted is a one-sided test, it becomes possible. If the true effects are assumed to be positive, and a corresponding one-sided test is performed, the replication probability $p_R$ is

$$p_R = p_1^+ p_2^+ + (1 - p_1^+)(1 - p_2^+), \tag{12}$$

where $p_k^+$ is the probability for an individual significance test being significant. To recover the $\mathit{MDES}_R$ needed for a two-sided test with Type I error rate $\alpha$, we use $\alpha/2$ for the Type I error rate of the one-sided test. Then, if both true effects are assumed to be equal, the required $\mathit{MDES}_R$ under the null hypothesis is

$$\mathit{MDES}_R = (z_{1-\alpha/2} + z_{1-\beta}) \frac{\tau}{z_{1-\alpha/2} - \Phi^{-1}\left(\frac{1-\sqrt{1-2(1-p_R)}}{2}\right)}. \tag{13}$$

The formula highlights that the $\mathit{MDES}_R$ needed to obtain a prespecified replication probability $p_R$ depends only on the true effect $\tau$ once the Type I and II error rates have been fixed. Thus, researchers cannot choose the $\mathit{MDES}_R$ based on the smallest effect size of interest as is commonly done for a standard power analysis of a single study. $\mathit{MDES}_R$ is fixed once the true effect $\tau$ and desired replication probability have been determined. Also note that researchers' smallest effect size of interest cannot be used as a substitute for the unknown true effect $\tau$. Doing so would result in an $\mathit{MDES}_R$ with unknown actual replication probability—one would not even know whether to expect high or low replication probabilities. If the equality of effects is not tenable then the two studies' $\mathit{MDES}_R$ or sample size needs to be determined in a simulation that allows for different true effects across the two studies.

## Distance-based Correspondence Measures

Determining the $\mathit{MDES}_R$ needed to achieve a predefined replication probability $p_R$ for distance-based correspondence measures requires researchers to specify the presumed equivalence threshold $\delta_E$. Unlike the conclusion-based measures, researchers do not need to make any assumptions about the magnitude of effects. Since the $\mathit{MDES}_R$ computations are performed under the hypothesis that the true effects are identical across studies, no explicit specification of the effect difference is needed either because $\tau_1 - \tau_2 = 0$. Replication probabilities can again be expressed in terms of the average $\mathit{MDES}_R^*$.

### ***Difference Test***

For $\tau_1 - \tau_2 = 0$, the replication probability of the difference test depends only on the Type I error rate $\alpha_R$ and is given by $p_R = 1 - \alpha_R$. Thus, under the null hypothesis, any $\mathit{MDES}_R$ would be adequate for the difference test to indicate correspondence with a high replication probability.

### ***Equivalence Test***

The necessary $\mathit{MDES}_R$ to demonstrate equivalence depends on the equivalence threshold, in addition to the Type I and II error rates. The average $\mathit{MDES}_R^*$ needed to ensure a replication probability of $p_R$ for the equivalence test is

$$\mathit{MDES}_R^* = (z_{1-\alpha_R/2} + z_{1-\beta_R})\sqrt{\frac{1}{2}\left(\frac{\delta_E}{\Phi^{-1}(\frac{p_R}{2} + .5) + z_{1-\alpha_R}}\right)^2}. \tag{14}$$

The formula highlights that the $\mathit{MDES}_R^*$ needed to obtain a prespecified replication probability $p_R$ depends only on the equivalence threshold $\delta_E$ once the Type I and II error rates have been fixed. As before, the $\mathit{MDES}_R^*$ cannot be chosen by researchers with respect to the smallest effect size of interest; it is fixed once the desired replication probability and equivalence threshold have been determined.

### ***Correspondence Test***

To determine the $MDES_R^*$ and sample size necessary for the correspondence test, one can use the formula for the equivalence test to obtain a predefined replication probability. This is sufficient if the goal of the replication effort is to demonstrate equivalence under the presumption that the effects are stable across studies. If this presumption is false, the correspondence test has the advantage to signal significant effect differences. Importantly, if a replication effort has sufficient power to demonstrate equivalence with a reasonably small equivalence threshold, then an *Indeterminacy* outcome will be very unlikely but either a *Trivial Difference* or *Difference* will be indicated if the effect difference is greater than the studies' $MDES_R^*$. In discussing the outcomes of the correspondence test, we need to distinguish between two scenarios. First, where the two studies are powered for an $MDES_R^*$ that results in a replication probability $p_R \leq 1 - \alpha_R = .95$ for the equivalence tests, and second, where $p_R > 1 - \alpha_R = .95$ (for the standard choice of $\alpha_R = .05$).

If the equivalence test has a replication probability $p_R \leq .95$, the correspondence test can result in *Equivalence*, *Difference*, or *Indeterminacy* but not in a *Trivial Difference* provided $\tau_1 = \tau_2$ is true. Thus, the probability of the correspondence test showing *Equivalence* is equal to the replication probability used for the equivalence test, and the probability of indicating a *Difference* is equal to $\alpha_R = .05$ (because the difference test will have a replication probability of $.95 = 1 - \alpha_R$). Finally, the probability of showing *Indeterminacy* is $1 - .05 - P(ET = 1)$.

If the studies are powered such that the replication probability of the equivalence test is larger than .95, then the outcome of a *Trivial Difference* becomes a possibility even if $\tau_1 = \tau_2$ holds. The probability of the correspondence test showing *Equivalence* is less than the probability found using the formula for the equivalence test (because some significant equivalence results will also suggest a significant difference). However, a replication probability $p_R > .95$ for an equivalence test is unlikely in practice. Even when $\tau_1 = \tau_2$, the sample size requirements to achieve the necessary study-level power are very high, and thus

*Trivial Difference* is unlikely unless an unreasonably large equivalence threshold is used.

### Example: $\boldsymbol{MDES_R}$ and Sample Size Calculations

To demonstrate the $MDES_R$ and sample size calculations, we consider the implementation of two independent RCTs to replicate a well-defined treatment-control contrast for a specific population and setting of interest. The causal effects of the two RCTs, with $i = 1, 2, ..., n_k$ participants in study $k = 1, 2$, will be independently estimated and tested with a simple regression model without any additional control variables: $\hat{Y}_{ik} = \hat{\gamma}_k + \hat{\tau}_k Z_{ik}$, where $\hat{\gamma}_k$ and $\hat{\tau}_k$ are the estimators of the intercept and causal effect, respectively, and $Z_{ik}$ is the dummy-coded treatment indicator. Assuming homoskedasticity, equal treatment and control group sizes, and a standardized outcome variable $Y$, the standard error $\sigma_{\hat{\tau}_k}$ of each regression estimator $\hat{\tau}_k$ is given by

$$\sigma_{\hat{\tau}_k} = \frac{\sqrt{1 - R_k^2}}{.5\sqrt{n_k}}, \tag{15}$$

where $R_k^2$ is the coefficient of determination of the regression model (Hanley, 2016). The explicit specification of the estimators for the causal effect and its standard error concludes Step 1 of the procedure for determining the studies' required sample sizes (outlined above at the beginning of this section).

Step 2 demands the determination of the study-specific Type I and II error rates which are set to the conventional levels of $\alpha = .05$ and $\beta = .2$, the latter implying a power of .8 for each study. The resulting $z$-values are then given by $z_{1-\alpha/2} \approx 1.96$ and $z_{1-\beta} \approx .84$.

Step 3 first requires the choice a correspondence measure. For demonstration purposes we compute the $MDES_R$ and required sample sizes for all correspondence measures. For each measure, we aim for a replication probability of at least $p_R = .8$. For correspondence in significance pattern, we calculate sample size needs for three different values for the unknown true effects: $\tau_1 = \tau_2 = \tau = 0.1$, 0.5, and 1 SD. For the equivalence test we investigate four different thresholds: $\delta_E = 0.1$, 0.2, 0.3, and 0.5 SD. Finally, we use Type I and II error rates of $\alpha_R = .05$ and $\beta_R = .2$ together with $z_{1-\alpha_R/2} \approx 1.96$ and

$z_{1-\beta_R} \approx .84$ for the difference and equivalence test.

In Step 4, we plug the numbers from Step 3 into the $MDES_R$ formulas for each correspondence measure to obtain the minimum required $MDES_R$ for each RCT to guarantee a replication probability of $p_R = .8$. Step 5 first transforms the $MDES_R$ into the study-specific standard error, $\sigma_{\hat{\tau}} = \frac{MDES_R}{z_{1-\alpha/2}+z_{1-\beta}} = \frac{MDES_R}{1.96+0.84}$. Then, we solve the formula of our regression-based standard error Equation (15) for the sample size, $n_k = \left(\frac{\sqrt{1-R_k^2}}{0.5\sigma_{\hat{\tau}_k}}\right)^2$, and after plugging in the standard error expressed in terms of $MDES_R$ we obtain the minimum required sample size for each study. Note that the two RCTs will need different sample sizes if the error variances $(1 - R_k^2)$ are assumed to be different. For effect estimators beyond Ordinary Least Squares (OLS) regression, like from random effects models, sample sizes $n_k$ need to be computed using the appropriate formulas for the estimators of the standard error.

Table 3 shows the results for correspondence in significance pattern. The table highlights that the required $MDES_R$ must be smaller than the true effect to achieve a replication probability of .8. For instance, if the true effect is $\tau = 0.3$ SD for both studies, the $MDES_R$ must be 0.26 SD or smaller. It is also important to note that sample size requirements increase drastically if the true effects become small. The main practical issue here is that the true effect sizes are unknown and most likely vary across studies if populations and settings differ. Thus, planning for sufficiently powered replication studies remains a challenge.

Table 4 shows the results for the equivalence test with four different thresholds: $\delta_E = 0.1$, 0.2, 0.3, and 0.5 SD. No results for the difference and correspondence test are contained in the table because the difference test always has a replication probability of $1 - \alpha_R = .95$ under the assumption of identical effects, while the correspondence test uses the same $MDES_R^*$ as the equivalence test. In comparison to correspondence in significance pattern, the equivalence test requires a smaller $MDES_R^*$ if small equivalence thresholds, $\delta_E$, are used. The results in Table 4 indicate that the studies' $MDES_R^*$ must be considerably lower than the chosen equivalence threshold $\delta_E$. For instance, if a threshold of $\delta_E = 0.2$ SD is used, then both studies' $MDES_R$ must be 0.14 SD to achieve a replication probability of .8. The corresponding sample size requirements then depend on the magnitude of the error variances in the two regression models, here expressed in terms of the study-specific $R^2$ (assumed to be identical for the two studies). Then, for $\delta_E = 0.2$ and $R^2 = .1$, we would need 1542 units for each study in our example; With $R^2 = .5$ we still would need 858 units per study. From these results, it is clear that achieving a high replication probability requires high precision for both studies.

## Re-analysis of the Open Science Collaboration Data

Using the data set from the study of the Open Science Collaboration (2015), we compare the performance of the discussed correspondence measures with real data and assess whether the choice of a specific correspondence measure makes a difference in practice. The re-analysis also serves as a demonstration for how the correspondence test can be used to analyze results of pairwise replication studies and highlights that post-hoc replication efforts tend to be underpowered. The OSC data consist of 100 replications of statistically significant results of original studies in psychology. Since each replication effort consists of an original study and a single replication study we can apply our correspondence measures to each replication pair. Since the effect sizes in the OSC data set were reported in terms of Pearson's correlation $r$, with no original effect sizes or standard errors reported, we applied Fisher's $z$-transformation to convert the correlations to $z$-scores ($z = \text{arctanh}(r)$) and computed corresponding standard errors for each estimate of an effect size ($\sigma = \frac{1}{\sqrt{n-3}}$). Though the effect sizes and standard errors on the raw scales would be preferable for assessing replication success, the $z$-scores provide a good approximation and are sufficient for our purpose of comparing the correspondence measures' performance. Of the original 100 studies, the OSC team reported that standard errors are only calculable for 73 of them—specifically, the studies that reported their results with a $t$-test, $F$-test with 1 numerator degree of freedom, or a Pearson's $r$. For the other studies that reported their results with a $\chi^2$-statistic or an $F$-statistic with more than 1 numerator degree of freedom, standard errors are not calculable as these test statistics correspond to analyses with more than two variables (i.e., treatment conditions).

With these 73 studies, correspondence between the original and replication effects is assessed based on their conclusion-based correspondence in significance pattern, and their distance-based difference, equivalence and correspondence test. We conducted the equivalence and correspondence test with three different equivalence thresholds in terms of Pearson's $r$ ($r = 0.1, 0.3, 0.5$, which corresponds to thresholds of $\delta_E = 0.1, 0.31, 0.55$ in SD units). For each correspondence measure, the portion of successful replications is shown in Table 5.

For the correspondence in significance measure we obtain a replication rate of 37% which is well aligned with the 36% that OSC found when judging replication by correspondence in significance. Switching to the distance-based correspondence measures, the difference test suggests that 71% of the studies successfully replicated. To probe whether this high replication success is due to a lack of power or whether the obtained effect estimates are truly very similar, we conduct the equivalence and correspondence tests. Since the replication rates for the equivalence test are rather low, ranging from 0% to 8% and 40% for thresholds $\delta_E = 0.1, 0.3,$ and 0.5, respectively, there is strong indication that both the equivalence and difference tests lacked sufficient power to establish equivalence or a difference in effects, respectively. This is confirmed by the very high *Indeterminacy* rate of the correspondence test, which is 71% for $\delta_E = 0.1$ and 64% for $\delta_E = 0.3$. Only when a threshold of $\delta_E = 0.5$ is used the *Indeterminacy* percentage drops down to 40%. For thresholds $\delta_E = 0.1$ and 0.3, *Equivalence* is established for none (0%) or only 7% of the replications. To increase the *Equivalence* percentage to 32% we would need to use an unreasonably large equivalence threshold of $\delta_E = 0.5$. However, a significant *Difference* is indicated for 29% of the studies, with 1% and 8% suggesting a *Trivial Difference* for $\delta_E = 0.3$ and 0.5, respectively.

As expected, the *Equivalence* and *Indeterminacy* rates of the correspondence test strongly depend on the magnitude of the chosen equivalence threshold. As the threshold increases, the likelihood of replication success increases while the inconclusive outcome of *Indeterminacy* become less likely. The overall conclusion about a difference in effects does not depend on the chosen threshold. As the threshold increases, the split between a *Difference* and a *Trivial Difference* changes, with more *Trivial Differences* observed when the threshold is higher. The choice of an appropriate equivalence threshold should be informed by the largest acceptable difference in effects that researchers consider as being in agreement with subject matter theory or meaningful with regard to the population of interest. The threshold should be smaller than the largest difference considered as inconsequential or insufficiently meaningful. Lacking this information for each single study, we examined the differences between the original and replicated effect estimates, and transformed those differences back to the scale of Pearson's $r$. The resulting average absolute difference in observed effects across all studies is 0.29 (with a median of 0.24). This explains the low *Equivalence* rates for thresholds of $\delta_E = 0.1$ and 0.3.

Overall, the results in Table 5 indicate that the choice of an appropriate correspondence measure matters. The equivalence and correspondence test apparently suffer from insufficiently powered studies as indicated by the high *Indeterminacy* percentage. As the OSC dataset is comprised of post-hoc replications, it is unsurprising that *Indeterminacy* was so common. This underscores the importance of prospectively planned replication studies, as this makes it more likely that sufficient power is attained to show replication success. Insufficient power is most likely also an issue for the correspondence in significance pattern, but since the true effect sizes are not known, insufficient power and actual effect differences contribute to the low replication percentage. Finally, note that the low replication rates of the equivalence test and the correspondence in significance pattern may also be due to publication biases or questionable research practices in the original studies.

## Discussion

In this article we investigated the statistical properties of commonly used and some new correspondence measures under ideal conditions (absent of any publication bias and questionable research practices) and provided formulas for determining individual study $MDES_R$s that are needed to achieve a predefined replication probability.

Although conclusion-based measures like correspondence in significance pattern have the disadvantage that the replication probabilities depend on the unknown true effects, they are, in general, less demanding with regard to sample size needs unless the true effects are small. Despite their dependence on the magnitude of the true effects, conclusion-based measures are of interest whenever the research question of the replication effort takes a policy perspective. That is, would researchers draw the same conclusion about the effectiveness of an intervention or treatment from two independent but comparable studies? Whether the effect magnitudes differ is not of concern here.

Distance-based measures assess the similarity of the two effect estimates rather than the congruence of conclusions derived from the two studies. The main advantage of distance-based measures is that the replication probabilities do not depend on the magnitude of the unknown true effects but on their difference only. Thus, determining sample size requirements to ensure a predefined replication probability is easier for distance- than conclusion-based measures. However, equivalence and correspondence tests require the specification of an equivalence threshold that is considered negligibly small or inconsequential for a given replication effort.

The assessment of replication success or failure depends on the choice of the correspondence metric. Measures may contradict each other. For instance, the correspondence test might reveal *Equivalence* of the two effect estimates when considered jointly, while the estimate may be significant in one study but non-significant in the other (which may even occur for identical effect estimates). Conversely, both studies may result in a significant positive effect, but the difference in effect sizes might result in a significant

*Difference* outcome when the correspondence test is applied. Thus, it is important that researchers choose their preferred correspondence measure before analyzing the data, ideally when planning the replication effort.

Whenever the main goal of a replication effort is about demonstrating stable effects or learning about effect heterogeneity, the correspondence test should be the preferred measure because it is able to explicitly distinguish between *Equivalence*, *(Trivial) Difference*, and *Indeterminacy*. Sometimes, the two studies might be sufficiently powered to establish *Equivalence* if the underlying true effects are (almost) identical, or to establish a *Difference* if the the true effects are sufficiently different across the two studies. Moreover, an *Indeterminacy* outcome prevents researchers from over-interpreting the result of a replication effort and points towards the need for more evidence. Thus, the correspondence test is a severe test (Mayo, 2018) in the sense that it has a high probability of indicating *Equivalence* if the effects are equivalent, a high probability of indicating a *(Trivial) Difference* if the effects actually differ, and a high probability of *Indeterminacy* if the evidence is insufficient to establish *Equivalence* or a *(Trivial) Difference*. None of the other correspondence measures achieves this level of severity. A key requirement for making effective use of the correspondence test is the prospective planning of replication studies. The sample size requirements to consistently show results other than *Indeterminacy* are quite high, and are unlikely to be attained from a post-hoc replication effort. Thus, a move away from post-hoc replication studies towards more prospectively planned ones is necessary to make use of methods most suited for assessing replication success and failure like the correspondence test.

However, the main practical challenge with all correspondence measures is the potential lack of power to demonstrate equivalence even if the underlying true effects are identical. To guarantee replication probabilities of at least .8, the $MDES_R$ must be quite small for both studies. Thus, without prospectively planned and sufficiently powered replication efforts, researchers should not be surprised to regularly see failed replications even if the underlying effects are (almost) identical. After all, replication efforts address a research question that is different from demonstrating an effect only once in a single study. Thus, power considerations must appropriately reflect the replication question. One way to increase the replication probabilities (provided the true effects are not too different) is to conduct prospectively planned replications with sufficiently powered individual studies. In general, the required $MDES_R$s will be smaller than what is needed for the demonstration of an effect in a single study. Alternatively, one may use the same *MDES*s but increase the power to at least .9 or .95 instead of using the conventional power of .8.

Although we have assumed throughout this article that the effect estimates were obtained from an RCT, these results generalize to effect estimates from quasi-experimental and observational studies, provided the assumptions of the Causal Replication Framework hold. Observational studies are likely to have larger sample sizes than RCTs, leading to effect estimates with more precision. Thus, well-designed and planned observational studies could make use of the correspondence test and be more likely to have sufficient power to obtain a result other than *Indeterminacy*.

Another way out of the replication crisis is to acknowledge that single studies are unlikely to be sufficiently powered for pairwise replication assessments, and to conduct an entire series of prospectively and systematically planned replications. While two studies often will not have sufficient power to show conclusive evidence either of replication success or failure, evidence from multiple replication efforts can provide more power (Hedges & Schauer, 2019). Meta-analytic techniques such as the $Q$-statistic (Hedges & Schauer, 2018) can be used to assess if the effect heterogeneity among any number of replications exceeds a certain threshold, and response surface modeling can be used to assess and model effect heterogeneity if present (Box & Draper, 2007; Cooper et al., 2011; Rubin, 1992). Then, even in the presence of effect heterogeneity, more stable and reliable knowledge about interventions or treatments can be derived from multiple, potentially under-powered replications than from two highly powered studies. However, whether the overall sample size requirements across all studies entering a meta-analysis are less demanding than for two larger studies, particularly if the multiple studies vary considerably with respect to populations and settings, still needs to be investigated.

But even with a series of replication studies, pairwise assessments of replication success or failure should be conducted. First, researchers might want to assess replication success or failure after the results from the first two or three studies are in. Second, not all replication studies may be directly comparable by design. In such situations, researchers should assess pairwise correspondence for comparable studies (in addition to a meta-analysis or response surface model). Third, if effect heterogeneities are suspected, then it makes sense to investigate which effects significantly differ from each other. This helps identify effect moderators to establish generalizability boundaries or constraints on generality (COG; Simons et al., 2017).

Pairwise replication assessments can be particularly useful when researchers have a goal of falsifying a causal claim. For instance, if a study makes a claim about the effectiveness of a treatment based on an observational study, it is desirable to examine if this finding holds under an RCT. Correspondence metrics are useful in assessing whether the results of an RCT have falsified the claims of an observational study. If paired with a research design such as a within-study comparison, the researcher has some degree of control over sources of effect heterogeneity such as differences in settings or populations (Wong & Steiner, 2018a). This allows for *a priori* planning of the replication study to ensure a suitable replication probability is achieved. The correspondence test is the best-suited measure for this purpose as it enables researchers to understand whether their replication effort falsified the original finding, supported it, or if there is insufficient information to make a claim either way.

For pairwise replication assessments, we advocate using the correspondence test with its four possible outcomes of *Equivalence*, *Trivial Difference*, *Difference*, and *Indeterminacy*, because it allows researchers to clearly distinguish between replication success and failure with regard to the magnitude of the underlying true effects. The application of the correspondence test also highlights that the replication crisis is a crisis of indeterminacy rather than an explicit failure. Replication failure due to a *(Trivial) Difference* in effects does not need to be regarded as failure as long as the source for the difference is causally identifiable (Steiner et al., 2019; Wong et al., 2021). In such situations, researchers actually learn from failed replications—they provide information that is at least as meaningful as establishing the comparability of effect sizes in direct replication efforts.

## References

Anderson, S., & Kelley, K. (2022). Sample size planning for replication studies: The devil is in the design. *Psychological Methods*. https://doi.org/10.1037/met0000520

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

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.

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

Andrews, I., & Kasy, M. (2019). Identification of and Correction for Publication Bias. *American Economic Review*, *109*(8), 2766–2794. https://doi.org/10.1257/aer.20180310

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press. https://doi.org/10.1515/9781400829828/HTML

Bollen, K., Cacioppo, J., Kaplan, R., Krosnick, J. A., & Olds, J. L. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (tech. rep.).

Bonett, D. G. (2021). Design and analysis of replication studies. *Organizational Research Methods*, *24*(3), 513–529. https://doi.org/10.1177/1094428120911088

Box, G. E. P., & Draper, N. R. (2007). *Response surfaces, mixtures, and ridge analyses* (2nd ed.). Wiley Interscience.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Carroll, H. A., Toumpakari, Z., Johnson, L., & Betts, J. A. (2017). The perceived feasibility of methods to reduce publication bias. *PLOS ONE*, *12*(10), e0186472. https://doi.org/10.1371/JOURNAL.PONE.0186472

Chow, S.-C., Shao, J., & Wang, H. (Eds.). (2008). *Sample size calculations in clinical research* (2nd ed). Chapman & Hall/CRC.

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*(7485), 612–613. https://doi.org/10.1038/505612a

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2011). *The Handbook of Research Syntheses and Meta-Analysis* (2nd). Russell Sage Foundation.

Dong, N., & Lipsey, M. W. (2018). Can Propensity Score Analysis Approximate Randomized Experiments Using Pretest and Demographic Information in Pre-K Intervention Research? *Evaluation Review*, *42*(1), 34–70. https://doi.org/10.1177/0193841X17749824

Duvendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by "replication" and why does it encounter resistance in economics? *American Economic Review*, *107*(5), 46–51. https://doi.org/10.1257/AER.P20171031

Gleason, P., Resch, A., & Berk, J. (2018). RD or not RD: Using experimental studies to assess the performance of the regression discontinuity approach. *Evaluation Review*, *42*(1), 3–33. https://doi.org/10.1177/0193841X18787267

Hanley, J. A. (2016). Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology*, *79*, 112–119. https://doi.org/10.1016/J.JCLINEPI.2016.05.014

Hedges, L. V., & Schauer, J. M. (2018). Statistical Analyses for Studying Replication: Meta-Analytic Perspectives. *Psychological Methods*. https://doi.org/10.1037/MET0000189

Hedges, L. V., & Schauer, J. M. (2019). More Than One Replication Study Is Needed for Unambiguous Tests of Replication. *Journal of Educational and Behavioral Statistics*, *44*(5), 543–570. https://doi.org/10.3102/1076998619852953

Institute of Education Sciences. (2022). Research grants focused on systematic replication: Request for applications. https://ies.ed.gov/funding/ncer_rfas/systematic_replications.asp

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/JOURNAL.PMED.0020124

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0B013E31818131E7

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How Journal Editors View Replication Research. *Journal of Advertising*, *24*(4), 77–87. https://doi.org/10.1080/00913367.1995.10673490

Makel, M. C., & Plucker, J. A. (2014). Facts Are More Important Than Novelty: Replication in the Education Sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

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

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press. https://doi.org/10.1017/9781107286184

McShane, B. B., & Böckenholt, U. (2014). You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Popper, K. (1959). *The logic of scientific discovery*. Basic Books.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rindskopf, D. M., Shadish, W. R., & Clark, M. H. (2018). Using Bayesian Correspondence Criteria to Compare Results From a Randomized Experiment and a Quasi-Experiment Allowing Self-Selection. *Evaluation Review*, *42*(2), 248–280. https://doi.org/10.1177/0193841X18789532

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. https://doi.org/10.1037/0033-2909.113.3.553

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, *17*(4), 363–374. https://doi.org/10.3102/10769986017004363

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. http://dx.doi.org/10.1037/met0000302

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/A0015108

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.

Shadish, W. R., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

StataCorp. (2021). *Stata statistical software: Release 17*. College Station, TX: StataCorp LLC.

Steiner, P. M., & Wong, V. C. (2018). Assessing Correspondence Between Experimental and Nonexperimental Estimates in Within-Study Comparisons. *Evaluation Review*. https://doi.org/10.1177/0193841X18773807

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A Causal Replication Framework for Designing and Assessing Replication Efforts. *Zeitschrift fur Psychologie / Journal of Psychology*. https://doi.org/10.1027/2151-2604/a000385

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371–386.

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, *13*(3), 272.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*(2), 103–117.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority* (Second). Chapman and Hall/CRC. https://doi.org/10.1201/EBK1439808184

Wilde, E. T., & Hollister, R. (2007). How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. *Journal of Policy Analysis and Management*, *26*(3), 455–477. https://doi.org/https://doi.org/10.1002/pam.20262

Wong, V. C., Anglin, K., & Steiner, P. M. (2021). Design-Based Approaches to Causal Replication Studies. *Prevention Science 2021*, 1–16. https://doi.org/10.1007/S11121-021-01234-7

Wong, V. C., & Steiner, P. M. (2018a). Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings. *Evaluation Review*, *42*(2), 176–213. https://doi.org/10.1177/0193841X18778918

Wong, V. C., & Steiner, P. M. (2018b). Replication Designs for Causal Inference. *EdPolicyWorks Working Paper Series*.

**Table 1**

*Assessing Replication Success: Example*

| Study | Effect estimate (standard error) | Conclusion |
|---|---|---|
| Study 1 | 12 pts (4 pts) | significant |
| Study 2 | 11 pts (7 pts) | non-significant |
| Difference Study 1 vs 2 | 1 pt (8 pts) | non-significant |
| **Correspondence Measure** | **Outcome** | **Replication conclusion** |
| Significance pattern | different | failure |
| Difference test | non-significant | success |
| Equivalence test (3 pts threshold) | non-significant | failure |

**Table 2**

*Probabilities for each outcome of the correspondence test (CT)*

| | Case 1 | Case 2 |
|---|---|---|
| $CT$ Outcome | $P(ET = 1) \leq P(DT = 1)$ | $P(ET = 1) > P(DT = 1)$ |
| $P(CT = EQU)$ | $P(ET = 1)$ | $P(DT = 1)$ |
| $P(CT = DIF)$ | $1 - P(DT = 1)$ | $1 - P(ET = 1)$ |
| $P(CT = TRI)$ | 0 | $P(ET = 1) - P(DT = 1)$ |
| $P(CT = IND)$ | $P(DT = 1) - P(ET = 1)$ | 0 |

The table shows the probability of the *Equivalence*, *Difference*, *Trivial Difference*, and *Indeterminacy* outcomes in terms of replication probabilities of the Equivalence Test (ET) and Difference Test (DT). The CT probabilities depend on the sum of the replication probabilities for the ET and DT, that is, whether the sum is less than or equal to one (Case 1) or greater than one (Case 2). Each of the two columns sums to one.

**Table 3**

*Required replication minimum detectable effect size ($MDES_R$) and study-specific sample sizes for the correspondence in significance pattern under different values of the true effects* $\tau_1 = \tau_2 = \tau$

| | $\tau = 0.1$ | $\tau = 0.3$ | $\tau = 0.5$ | $\tau = 1$ |
|---|---|---|---|---|
| $MDES_R$ | 0.09 | 0.26 | 0.44 | 0.88 |
| Sample size ($n$) | 4016 | 440 | 152 | 32 |
| $R^2$ (i.e., variance explained by treatment indicator $Z_{ik}$) | .003 | .02 | .06 | .25 |

**Table 4**

*Required replication minimum detectable effect size ($MDES_R^*$) and study-specific sample sizes for the equivalence test with different equivalence thresholds ($\delta_E$, in SD units) and model $R^2$ ($= R_1^2 = R_2^2$), assuming no difference between true effects ($\tau_1 - \tau_2 = 0$)*

| | $\delta_E = 0.1$ | $\delta_E = 0.2$ | $\delta_E = 0.3$ | $\delta_E = 0.5$ |
|---|---|---|---|---|
| $MDES_R^*$ | 0.07 | 0.14 | 0.20 | 0.33 |
| Sample size ($n$) | | | | |
| $R^2 = .1$ | 6166 | 1542 | 686 | 248 |
| $R^2 = .3$ | 4796 | 1200 | 534 | 192 |
| $R^2 = .5$ | 3426 | 858 | 382 | 138 |

**Table 5**

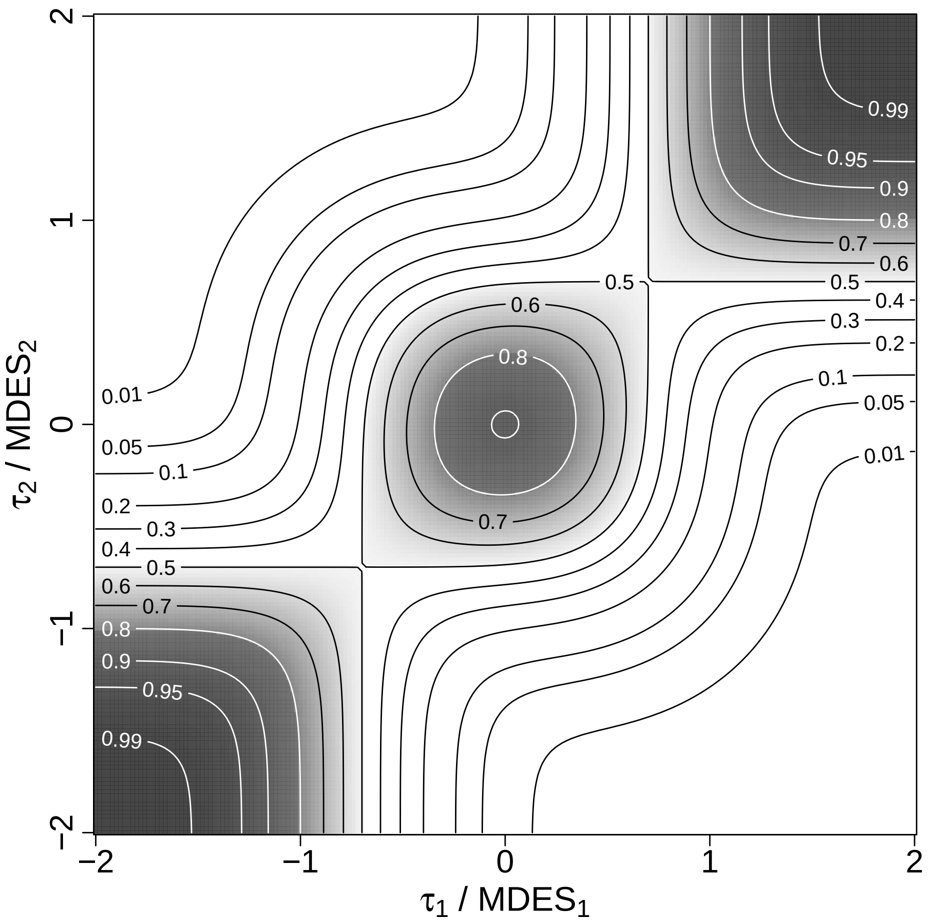
*Proportion of successful replications of 73 replication efforts of the Open Science Collaboration (2015)*

| Correspondence Measure | Proportion of Successful Replications | | |
|---|---|---|---|
| Significance Pattern, $S(\alpha = .05)$ | | .37 | |
| Difference Test, $DT(\alpha_R = .05)$ | | .71 | |
| Equivalence Threshold ($\delta_E$) | $\delta_E = .1$ | $\delta_E = .3$ | $\delta_E = .5$ |
| Equivalence Test, $ET(\delta_E, \alpha_R = .05)$ | 0 | .08 | .40 |
| Correspondence Test, $CT(\delta_E, \alpha_R = .05)$ | | | |
| *Equivalence* | 0 | .07 | .32 |
| *Difference* | .29 | .27 | .21 |
| *Trivial Difference* | 0 | .01 | .08 |
| *Indeterminacy* | .71 | .64 | .40 |

Note that proportions for the correspondence test ($CT$) may not sum to 1 due to rounding.

**Figure 1**

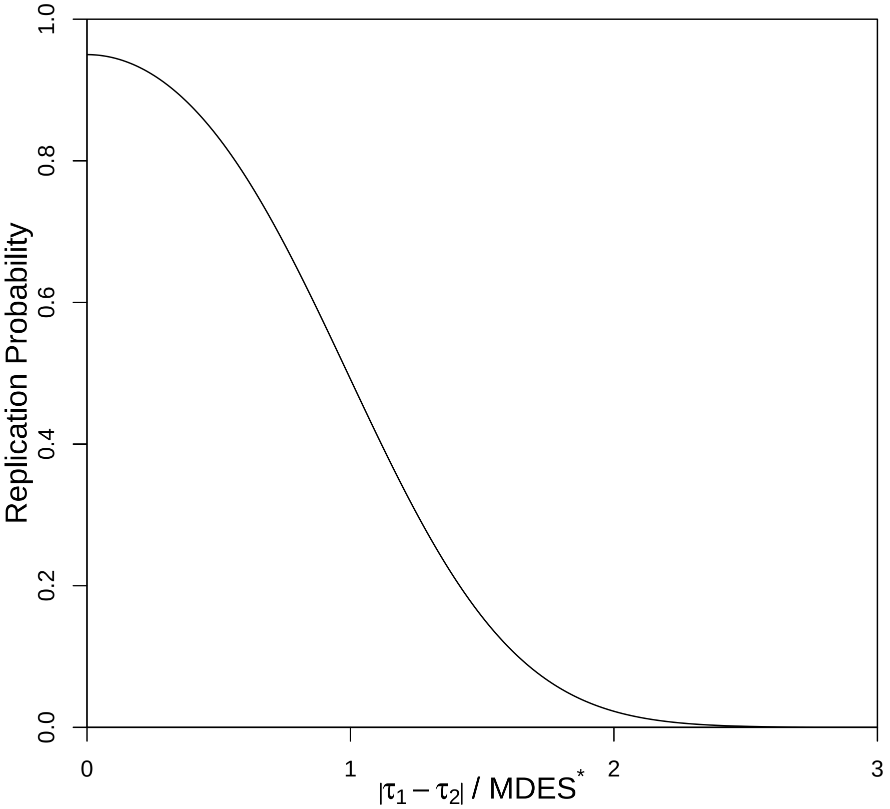
*Replication probability of the correspondence in significance pattern*



Replication probabilities are shown as function of the ratio of the true effect to the minimum detectable effect sizes, $\tau_k/MDES_k$, for studies $k = 1, 2$. The contour lines indicate specific replication probabilities.

**Figure 2**

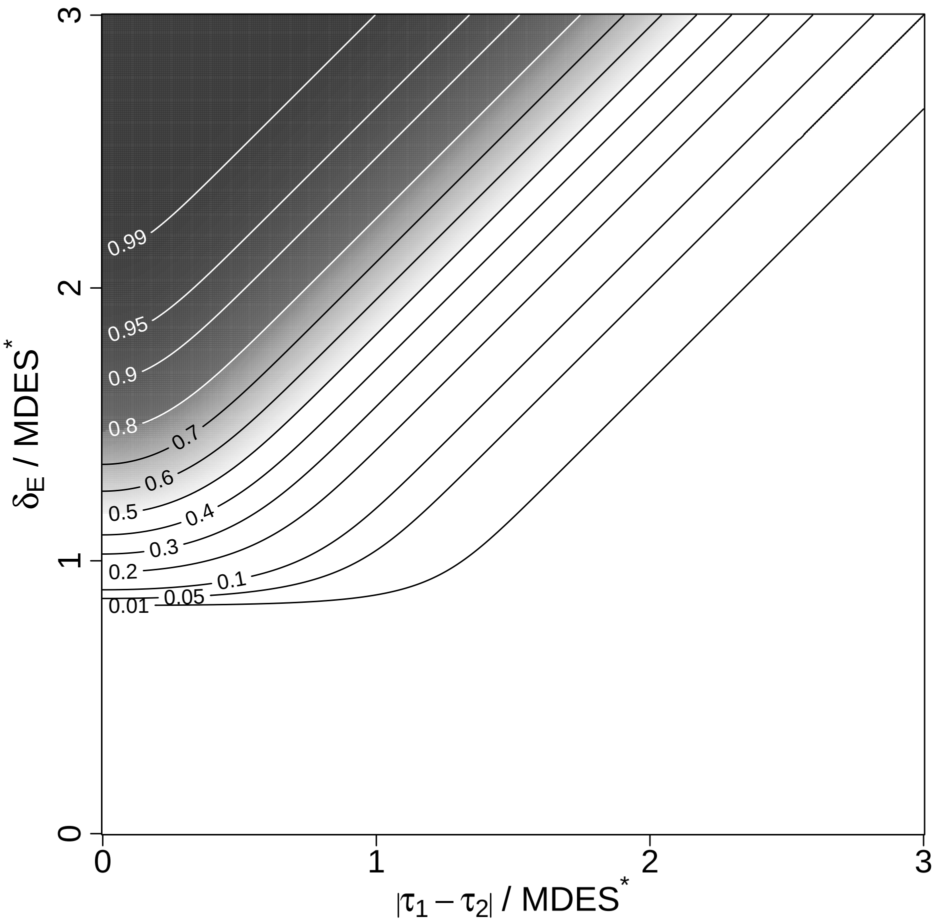
*Replication probability of the difference test*



Replication probabilities are depicted as function of the ratio of the absolute value of the difference in true effects to the average minimum detectable effect size, $(|\tau_1 - \tau_2|)/MDES^*$.

**Figure 3**

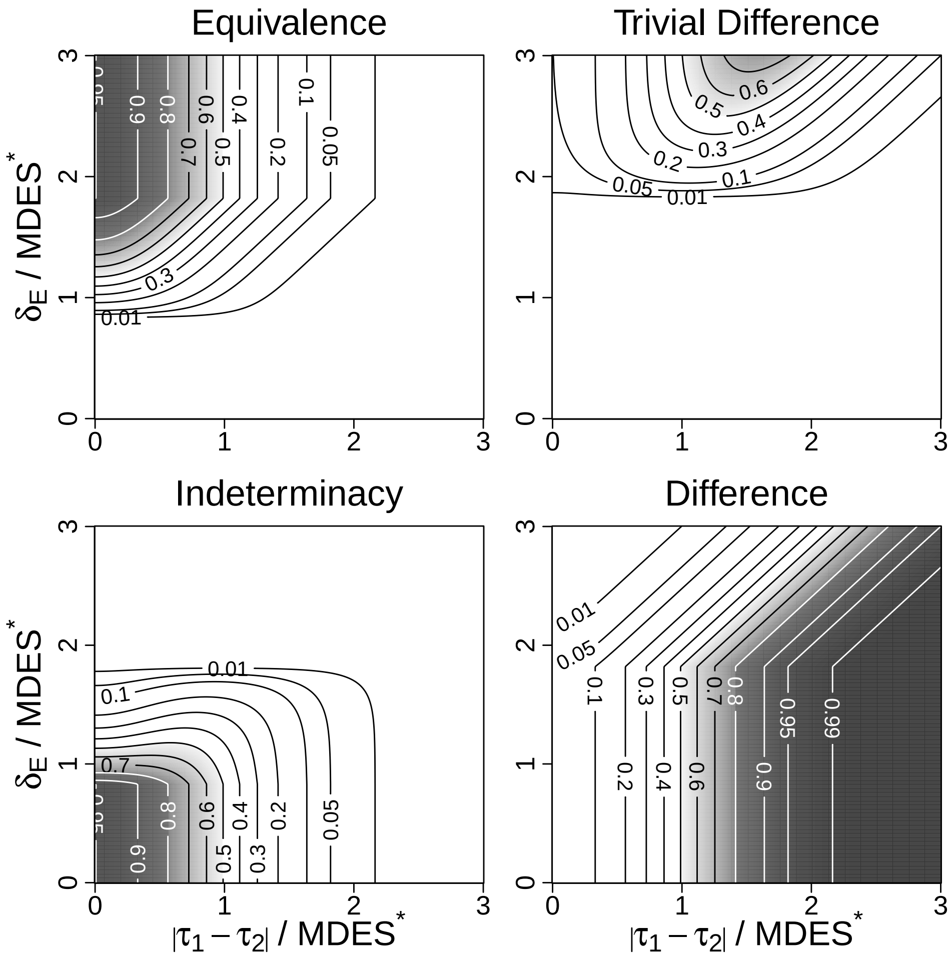
*Replication probability of the equivalence test*



Replication probabilities are depicted as function of the ratio of the absolute value of the difference in true effects to the average minimum detectable effect size, $|\tau_1 - \tau_2|/MDES^*$, and the ratio of the equivalence threshold to the average minimum detectable effect size, $\delta_E/MDES^*$. The contour lines show the replication probability attained for a given pair of ratios.

**Figure 4**

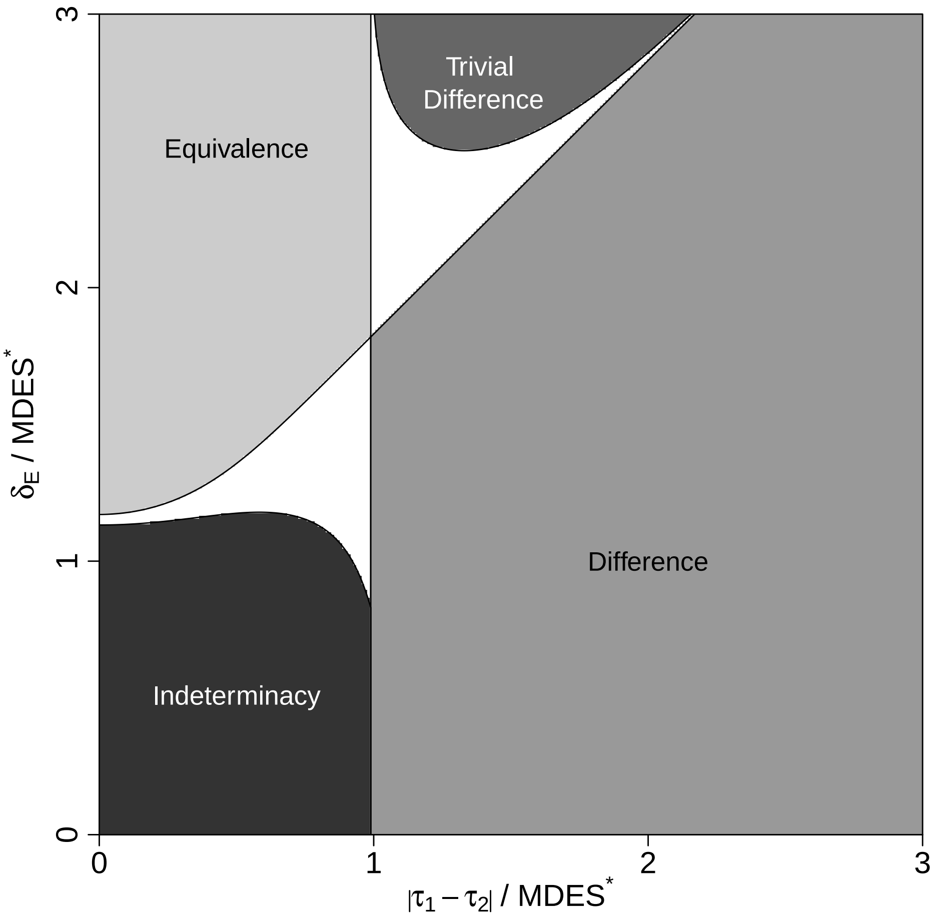
*Replication probability of the correspondence test*



Each panel shows the replication probabilities as function of the ratio of the absolute value of the difference in true effects to the average minimum detectable effect size, $|\tau_1 - \tau_2|/MDES^*$, and the ratio of the equivalence threshold to the average minimum detectable effect size, $\delta_E/MDES^*$. The contour lines show the replication probability attained for a given pair of ratios.

**Figure 5**

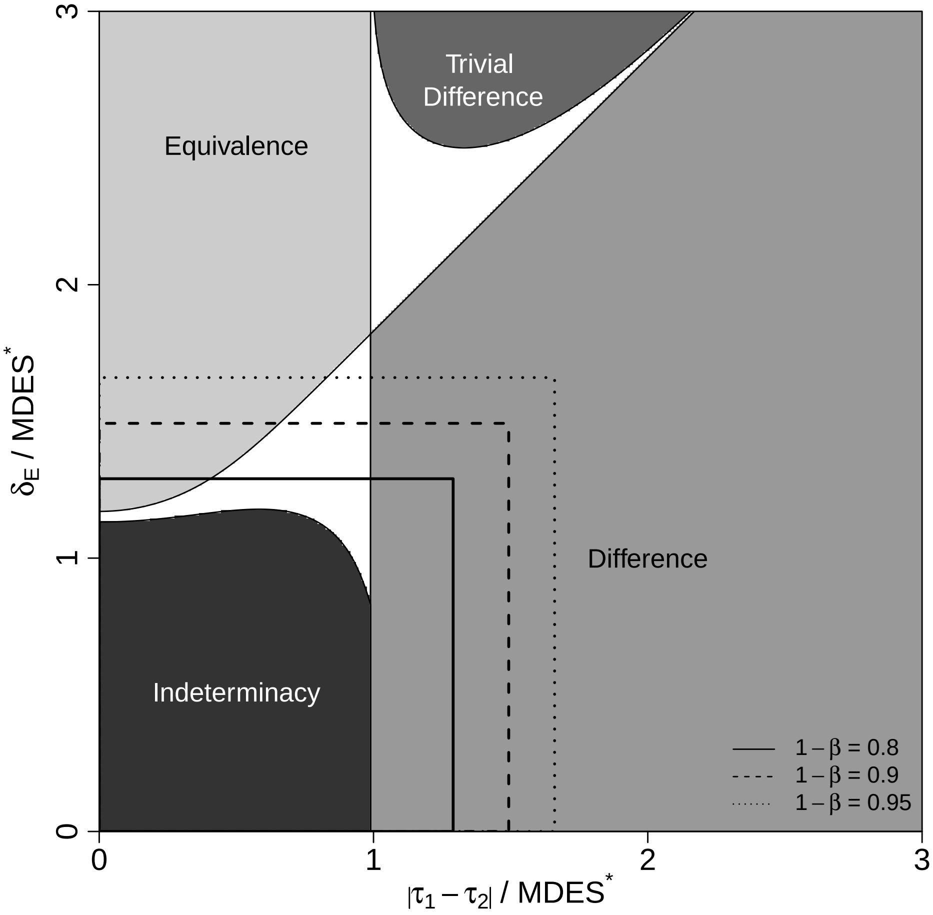
*Probability boundaries of the outcomes of the correspondence test*



The grey regions indicate ratio combinations for which the denoted outcome has a probability of at least .5. In the two white regions, the probability of any particular outcome is less than .5. These probabilities are based on the ratio of the absolute value of the difference in true effects to the average minimum detectable effect size, $|\tau_1 - \tau_2|/MDES^*$ and the ratio of the equivalence threshold to the average minimum detectable effect size, $\delta_E/MDES^*$.

**Figure 6**

*Probability boundaries of the outcomes of the correspondence test with area marking presumed current replication efforts*



See notes on as Figure 5. The areas demarcated by the solid black line, dashed line, and dotted line represent the presumed dominating outcomes when the individual minimum detectable effect sizes of replication studies are powered with .8, .9, and .95, respectively.

# Appendix A

## Correspondence in Magnitude and Sign

### Definition and Correspondence Probability

The simplest measure for comparing two effect estimates is to assess whether they both exceed or fall below a pre-specified magnitude threshold, $\delta_M$ (Wilde & Hollister, 2007). If the threshold is chosen to be zero, assessing correspondence of the two effect estimates is equivalent to comparing the estimates' sign. Importantly, correspondence in magnitude or sign does not account for any sampling uncertainty in effect estimates.

For a given threshold $\delta_M$, correspondence in magnitude, $M(\delta_M)$, can be formalized as an indicator function, 1[.], that returns 1 either if both effect estimates are greater than or equal to $\delta_M$ or both are less than $\delta_M$; otherwise, the indicator function returns 0 and indicates a lack of correspondence:

$$M(\delta_M) = 1[(\hat{\tau}_1 \geq \delta_M \;\&\; \hat{\tau}_2 \geq \delta_M) \vee (\hat{\tau}_1 < \delta_M \;\&\; \hat{\tau}_2 < \delta_M)], \tag{A1}$$

The replication probability, that is, the probability that the correspondence measure indicates replication success, $P(M(\delta_M) = 1)$, is the product of probabilities that each effect estimate equals or exceeds the magnitude threshold plus the product of probabilities that each will fall below the magnitude threshold. Assuming that the effect estimators' sampling distributions are asymptotically normal with expectations $\tau_1$ and $\tau_2$ and variances $\sigma^2_{\hat{\tau}_1}$ and $\sigma^2_{\hat{\tau}_2}$, the replication probability is given by

$$\begin{gathered} P(M(\delta_M) = 1|\tau_1, \tau_2, \sigma_{\hat{\tau}_1}, \sigma_{\hat{\tau}_2}, \delta_M) = \\ [1 - \Phi(\frac{\delta_M - \tau_1}{\sigma_{\hat{\tau}_1}})][1 - \Phi(\frac{\delta_M - \tau_2}{\sigma_{\hat{\tau}_2}})] + [\Phi(\frac{\delta_M - \tau_1}{\sigma_{\hat{\tau}_1}})][\Phi(\frac{\delta_M - \tau_2}{\sigma_{\hat{\tau}_2}})], \end{gathered} \tag{A2}$$

where $\Phi$ is standard normal distribution function.

A2 indicates that the replication probability depends on the threshold $\delta_M$, the unknown true effects $\tau_1$ and $\tau_2$, and the sampling variance of the estimators, $\sigma^2_{\hat{\tau}_1}$ and $\sigma^2_{\hat{\tau}_2}$. While the estimator variances can be directly inferred from the two studies, the magnitude of the two true effects remains unknown. Since the replication probability also depends on the threshold $\delta_M$, choosing the threshold in advance is crucial, and should be based on subject matter expertise and the minimum effect size of substantive interest. The default choice of $\delta_M = 0$ might not always be the best choice, particularly so when sampling uncertainty is not taken into account or when small effects have no policy relevance.

### Correspondence Probability Plot

Figure A1 shows the contour plot for two studies where the ratio $(\tau_k - \delta_M)/MDES_k$ may differ across the two studies. The $x$- and $y$-axes show the ratios for study 1 and 2, respectively. In interpreting the plot, we focus on the correspondence in sign where the magnitude threshold is set to zero, $\delta_M = 0$, such that the ratio directly reflects the size and sign of the unknown true effects in relation to the study-specific *MDES*s. The probability plot indicates that satisfactory replication probabilities of $P(M(\delta_M = 0) = 1) > .8$ are guaranteed only when both studies' ratio $\tau_k/MDES_k$ is greater than 0.43 or less than -0.43. If one study's ratio is larger, then the other study's ratio may be slightly smaller, but must always exceed about 0.3. This suggests that neither study needs to be sufficiently powered to actually detect the true effect (via NHST). This is so, because correspondence in sign does not rely on NHST and thus ignores sampling uncertainty. However, if the sign of the true effects differs, then the replication probability will always be below .5.

From Figure A1 it becomes clear that the values of the true effects greatly impact the replication probability. Even with the *MDES*s known, researchers still need to have reliable knowledge about the magnitude of the true effects to be able to estimate the replication probability or to sufficiently power the two studies to guarantee a replication probability greater than .8.

**Figure A1**

*Replication probability of correspondence in sign*

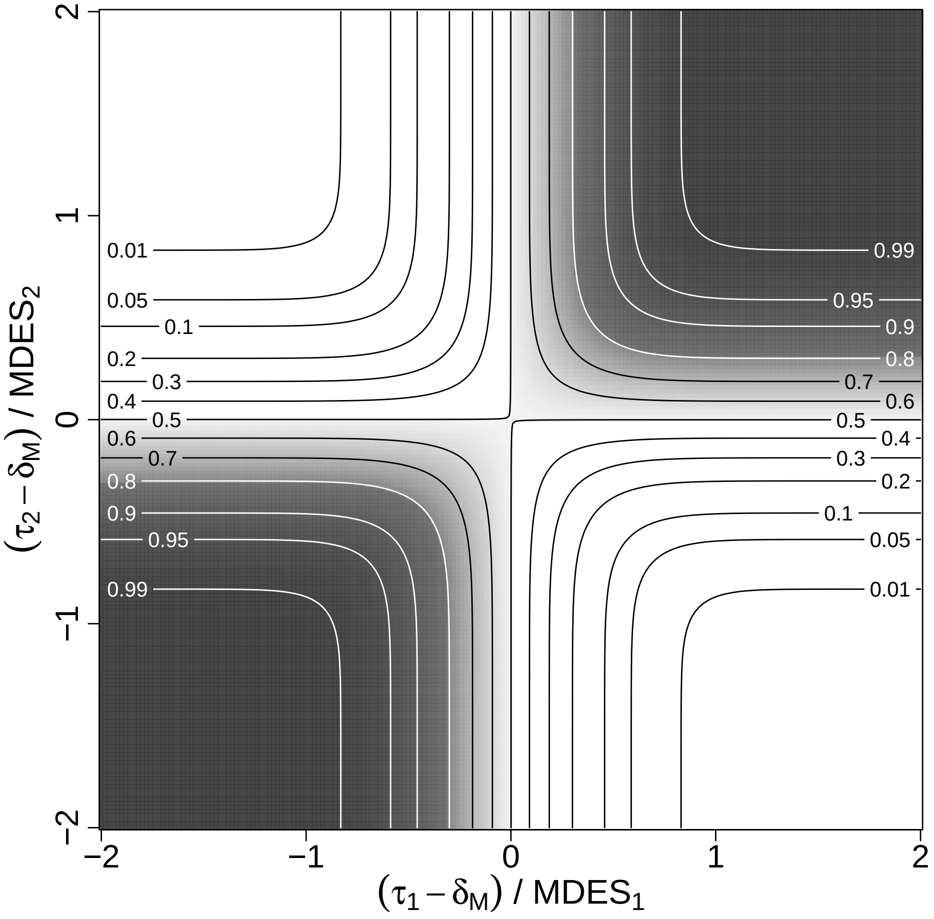


Figure A1 shows the replication probability for correspondence in sign dependent on the ratio of the true effect minus the magnitude threshold of both studies to the minimum detectable effect sizes, $(\tau_k - \delta_M)/MDES_k$, for studies $k = 1, 2$. The contour lines indicate specific replication probabilities.

## MDES Formula for Sample Size Calculations

Based on the formula for the replication probability in Equation (A2), the necessary *MDES* to obtain a fixed replication probability $p_R$ is given by

$$MDES = \frac{\delta_M - |\tau|}{\Phi^{-1}\left(\frac{1-\sqrt{1-2(1-p_R)}}{2}\right)} \times (z_{1-\alpha/2} + z_{1-\beta}), \tag{A3}$$

assuming $\delta_M < |\tau|$ to ensure the *MDES* is positive.

# Appendix B

## Proofs for replication probabilities

### Correspondence in magnitude and sign

The replication probability for correspondence in magnitude is the product of the probabilities that each effect estimate equals or exceeds the magnitude threshold plus the product of probabilities that each will fall below the magnitude threshold:

$$P(M(\delta_M) = 1|\tau_1, \tau_2, \delta_M) = P_1(\hat{\tau}_1 \geq \delta_M)P_2(\hat{\tau}_2 \geq \delta_M) + P_1(\hat{\tau}_1 < \delta_M)P_2(\hat{\tau}_2 < \delta_M), \qquad \text{(B1)}$$

where $P_1$ and $P_2$ are the randomization or sampling distributions of $\hat{\tau}_1$ and $\hat{\tau}_2$, respectively. While $P_1$ and $P_2$ also depend on the true effects and the magnitude threshold, we explicitly indicate this dependence only in the expression for the replication probability after the conditioning bar.

The probability of any given effect estimate $\hat{\tau}_k$ being larger than the magnitude threshold $\delta_M$, with $k = 1, 2$ and assuming $\hat{\tau}_k$ is distributed normally, is

$$\begin{aligned} &P(\hat{\tau}_k > \delta_M), \quad \hat{\tau}_k \sim N(\tau_k, \sigma^2_{\tau_k}) \\ &= P(\frac{\hat{\tau}_k - \tau_k}{\sigma_{\hat{\tau}_k}} > \frac{\delta_M - \tau_k}{\sigma_{\hat{\tau}_k}}) \\ &= 1 - \Phi(\frac{\delta_M - \tau_k}{\sigma_{\hat{\tau}_k}}), \end{aligned}$$

where $\Phi$ represents the standard normal probability density function.

To find the replication probability for two studies, this value is computed for both studies. The total replication probability is then the sum of the product of these two values and the probability that neither effects are larger than $\delta_M$. Thus, the overall replication probability is

$$P(M(\delta_M) = 1|\tau_1, \tau_2, \delta_M) = P(\hat{\tau}_1 > \delta_M)P(\hat{\tau}_2 > \delta_M) + (1 - P(\hat{\tau}_1 > \delta_M))(1 - P(\hat{\tau}_2 > \delta_M)).$$

### Correspondence in significance pattern

To determine the replication probability, we first compute for each study the probability that the effect estimate is positive and significant (denoted $ps_k^+$) and the probability that effect estimate is negative and significant (denoted $ps_k^-$). These two probabilities depend on the critical $z$-value for the selected $\alpha$, and the observed $z$-score:

$$\begin{aligned} ps_k^+ &= P_k(z_k \geq z^*_{1-\alpha/2}) \\ ps_k^- &= P_k(z_k \leq z^*_{\alpha/2}), \end{aligned} \tag{B2}$$

where, again, $P_k$ is the randomization or sampling distribution of the estimator $\hat{\tau}_k$. For $k = 1, 2$ studies both testing the null hypothesis $H_0 : \tau_k = \tau_0$ with study-specific effect estimators $\hat{\tau}_k$ with standard errors $\sigma_{\hat{\tau}_k}$, a critical $z$-value $z^*_{1-\alpha/2}$ for the two-tailed significance test, and conditional on the true value of the effect $\tau_k$, $ps_k^+$ is given by

$$\begin{aligned} ps_k^+ &= P(\frac{\hat{\tau}_k - \tau_0}{\sigma_{\hat{\tau}_k}} \geq z^*_{1-\alpha/2}) \\ &= P(\frac{\hat{\tau}_k - \tau_0}{\sigma_{\hat{\tau}_k}} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}} \geq z^*_{1-\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}) \\ &= P(\frac{\hat{\tau}_k - \tau_k}{\sigma_{\hat{\tau}_k}} \geq z^*_{1-\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}) \\ &= 1 - \Phi(z^*_{1-\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}). \end{aligned}$$

The probability that the observed effect is significant in the left tail, $ps_k^-$, is calculated in a similar way, with a critical value of $z^*_{\alpha/2}$:

$$\begin{aligned} ps_k^- &= P(\frac{\hat{\tau}_k - \tau_0}{\sigma_{\hat{\tau}_k}} \leq z^*_{\alpha/2}) \\ &= P(\frac{\hat{\tau}_k - \tau_0}{\sigma_{\hat{\tau}_k}} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}} \leq z^*_{\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}) \\ &= P(\frac{\hat{\tau}_k - \tau_k}{\sigma_{\hat{\tau}_k}} \leq z^*_{\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}) \\ &= \Phi(z^*_{\alpha/2} - \frac{\tau_k - \tau_0}{\sigma_{\hat{\tau}_k}}). \end{aligned}$$

The overall probability of showing correspondence in significance pattern for two studies is equal to the sum of the probability that both studies are significant in the right tail, the probability that both studies are significant in the left tail, and the probability that neither are significant. Assuming that the probability of obtaining a certain significance result in study 1 is independent of the probability of finding a certain significance result in study 2, the replication probability is:

$$\begin{aligned}P(S(\alpha)=1|\tau_1,\tau_2,\sigma_{\hat{\tau}_1},\sigma_{\hat{\tau}_2},\alpha) &= ps_1^+ps_2^+ + ps_1^-ps_2^- + (1-ps_1^+-ps_1^-)(1-ps_2^+-ps_2^-)\\ &= [1-\Phi(\zeta_1^+)][1-\Phi(\zeta_2^+)] + [\Phi(\zeta_1^-)][\Phi(\zeta_2^-)] + [\Phi(\zeta_1^+)-\Phi(\zeta_1^-)][\Phi(\zeta_2^+)-\Phi(\zeta_2^-)],\end{aligned}$$

where $\zeta_k^+ = z^*_{1-\alpha/2} - \frac{\tau_k-\tau_0}{\sigma_{\hat{\tau}_k}}$ and $\zeta_k^- = z^*_{\alpha/2} - \frac{\tau_k-\tau_0}{\sigma_{\hat{\tau}_k}}$

**Difference test**

The probability of finding replication due to a non-significant difference test is equivalent to the probability that the observed test statistic based on observed effects $\hat{\tau}_1$ and $\hat{\tau}_2$ and the standard error of their difference $\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}$ is less extreme than the critical z-values $z^*_{1-\alpha_R/2}$ for a given Type I error rate for the replication test $\alpha_R$:

$$\begin{aligned}&P(-z^*_{1-\alpha_R/2} < \frac{\hat{\tau}_1-\hat{\tau}_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} < z^*_{1-\alpha_R/2}|\tau_1,\ \tau_2)\\ &= P(-z^*_{1-\alpha_R/2} - \frac{\tau_1-\tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} < \frac{\hat{\tau}_1-\hat{\tau}_2-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} < z^*_{1-\alpha_R/2} - \frac{\tau_1-\tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}})\\ &= \Phi(z^*_{1-\alpha_R/2} - \frac{\tau_1-\tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}}) - \Phi(-z^*_{1-\alpha_R/2} - \frac{\tau_1-\tau_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}})\end{aligned}$$

**Equivalence test**

The probability of replication due to a significant equivalence test is the probability that the observed difference in effects $\hat{\tau}_1 - \hat{\tau}_2$ within the bounds set by the equivalence threshold $\pm\delta_E$, rescaled to be on the same scale as the test statistic:

$$\begin{aligned}&P(-\frac{\delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} + z^*_{1-\alpha_R} \le \frac{\hat{\tau}_1-\hat{\tau}_2}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} \le \frac{\delta_E}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} - z^*_{1-\alpha_R}|\tau_1,\ \tau_2)\\ &= P(-\frac{\delta_E-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} + z^*_{1-\alpha_R} \le \frac{\hat{\tau}_1-\hat{\tau}_2-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} \le \frac{\delta_E-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} - z^*_{1-\alpha_R}|\tau_1,\ \tau_2)\\ &= \Phi(\frac{\delta_E-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} - z^*_{1-\alpha}) - \Phi(\frac{-\delta_E-(\tau_1-\tau_2)}{\sqrt{\sigma^2_{\hat{\tau}_1}+\sigma^2_{\hat{\tau}_2}}} + z^*_{1-\alpha})\end{aligned}$$

Note that, depending on the choice of $\delta_E$, the replication probability here can be negative. $\delta_E$ must be sufficiently larger than the true difference in effects $\tau_1 - \tau_2$ to ensure that the first term in the final expression is larger than the second. In the applications here, if the replication probability is calculated to be negative for the equivalence test, it is set to 0.

**Correspondence Test**

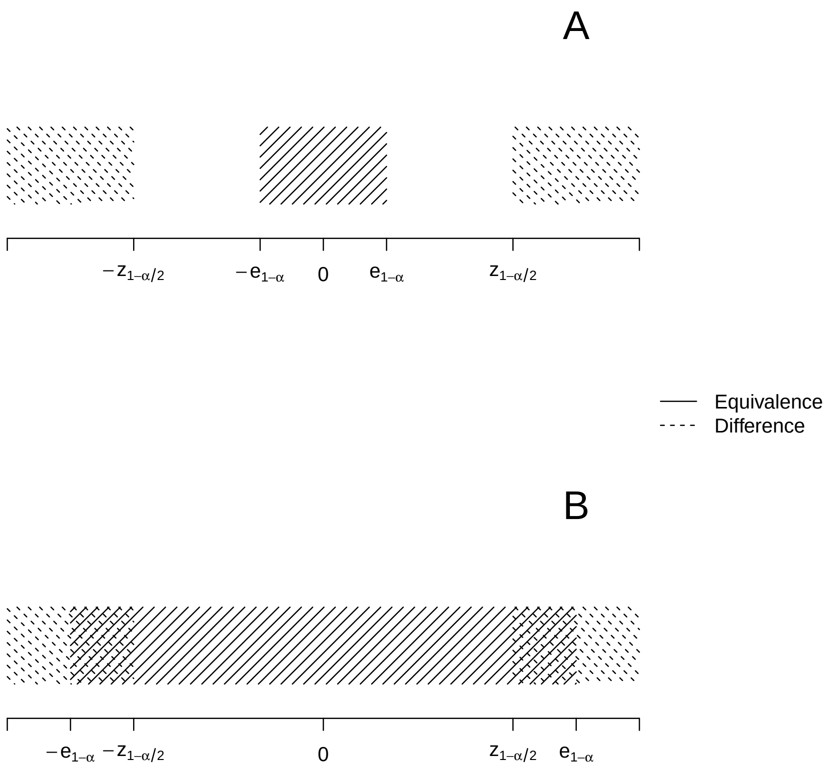
For the condition where $P(ET = 1) \leq P(DT = 1)$, the probabilities for the outcomes of *Equivalence* and *Difference* are the same as the outcome of the equivalence test and difference test. The probability of *Indeterminacy* is then equal to whatever is necessary so that the three probabilities sum to 1, i.e. $P(DT = 1) - P(ET = 1)$.

When $P(ET = 1) > P(DT = 1)$, the sum of these two probabilities exceeds one, indicating an overlapping region where both the equivalence test and difference test can be significant. Thus, the probability of *Equivalence* is equal to the probability of the difference test showing correspondence $P(DT = 1)$, as this is a necessary condition for the *Equivalence* result. The probability of *Difference* is simply $1 - P(ET = 1)$. Then, the probability of a *Trivial Difference* is the degree of overlap between $P(DT = 1)$ and $P(ET = 1)$, $P(ET = 1) - P(DT = 1)$.

Figure B1 visualizes the possible outcomes of the correspondence test. Panel A shows the three possibilities when $P(DT = 0) + P(ET = 1) \leq 1$. Here, when the difference between effects is close to 0, *Equivalence* is likely, while when the difference is far from 0, *Difference* is likely. The space between these outcomes leads to an outcome of *Indeterminacy*. Panel B of Figure B1 shows the possibilities when $P(DT = 0) + P(ET = 1) > 1$. Here, the regions of *Equivalence* and *Difference* overlap, allowing for the possibility of a *Trivial Difference* outcome.

**Figure B1**

*Visualization of possible results of the correspondence test*



Panel A visualizes the *Indeterminacy* situation. The *Equivalence* region is marked by the solid line, and bound by two critical thresholds $\pm e_{1-\alpha}$, that is, the one-sided $z$-values of the equivalence test. The *Difference* regions are marked by the dashed line, and begin at the critical values $\pm z_{1-\alpha/2}$. These regions do not overlap, and thus the empty white space between them is the region of *Indeterminacy.* Panel B shows the *Trivial Difference* situation, when the regions do overlap.

# Appendix C

## Proofs for Required Minimum Detectable Effect Sizes

### Correspondence in Sign or Magnitude

Here, we assume that the effect estimates for studies 1 and 2 will have the same standard error, $\sigma_{\hat{\tau}_1} = \sigma_{\hat{\tau}_2} = \sigma_{\hat{\tau}}$, and the same true effect $\tau$. Let $p = \Phi(\frac{\delta_M - \tau}{\sigma_{\hat{\tau}}})$ and $p_R$ be the replication probability:

$$\begin{aligned} p_R &= p^2 + (1-p)^2 \\ 0 &= 2p^2 - 2p - (p_R - 1) \\ p &= \frac{2 \pm \sqrt{4 - 4(2)(1-p_R)}}{4} \\ \Phi(\frac{\delta_M - \tau}{\sigma_{\hat{\tau}}}) &= \frac{1 \pm \sqrt{1 - 2(1-p_R)}}{2} \\ \sigma_{\hat{\tau}} &= \frac{\delta_M - \tau}{\Phi^{-1}\left(\frac{1 \pm \sqrt{1-2(1-p_R)}}{2}\right)} \end{aligned}$$

If the numerator is positive, then the addition part of the denominator will produce a valid (i.e., non-negative) standard error. If the numerator is negative, then the subtraction part will produce a valid standard error. To ensure the result is always positive, the absolute value of $\tau$ can be taken. Multiplying this result by $z_{1-\alpha_R/2} + z_{1-\beta_R}$ produces the required *MDES*:

$$MDES = (z_{1-\alpha/2} + z_{1-\beta}) \frac{\delta_M - |\tau|}{\Phi^{-1}\left(\frac{1 \pm \sqrt{1-2(1-p_R)}}{2}\right)}$$

### Correspondence in significance

For a set of one-sided tests with equal, positive effects $\tau_1 = \tau_2 = \tau$ and probabilities of being significant $ps_1^+ = ps_2^+ = ps_+ = \Phi(z_{1-\alpha/2} - \frac{\tau}{\sigma_\tau})$, the required *MDES* for a specified replication probability $p_R$ is

$$\begin{aligned}
p_R &= ps_+^2 + (1 - ps_+)^2 \\
0 &= 2ps_+^2 - 2ps_+ + 1 - p_R \\
ps_+ &= \frac{1 \pm \sqrt{(1 - 2(1 - p_R))}}{2} \\
\Phi(z_{1-\alpha/2} - \frac{\tau}{\sigma_{\hat{\tau}}}) &= \frac{1 \pm \sqrt{1 - 2(1 - p_R)}}{2} \\
\sigma_{\hat{\tau}} &= \frac{\tau}{z_{1-\alpha/2} - \Phi^{-1}(\frac{1 \pm \sqrt{1-2(1-p_R)}}{2})} \\
\mathit{MDES} &= (z_{1-\alpha/2} + z_{1-\beta})\frac{\tau}{z_{1-\alpha/2} - \Phi^{-1}(\frac{1 \pm \sqrt{1-2(1-p_R)}}{2})}.
\end{aligned}$$

The use of $\alpha/2$ in this formula will result in the required *MDES* to achieve correspondence in significance for a two-sided test with Type I error $\alpha$, rather than a one-sided test with Type I error rate $\alpha$.

**Equivalence Test**

For a given replication probability $p_R$, the necessary standard error for both studies (assuming equal standard errors, $\sigma^2_{\hat{\tau}_1} = \sigma^2_{\hat{\tau}_2} = \sigma^2$) and average $\mathit{MDES}^*$ (assuming standardized effects) is found by:

$$\begin{aligned}
p_R &= \Phi(\frac{\delta_E}{\sqrt{2\sigma^2}} - z_{1-\alpha_R}) - \Phi(\frac{-\delta_E}{\sqrt{2\sigma^2}} + z_{1-\alpha_R}) \\
p_R &= 2\left(\Phi(\frac{\delta_E}{\sqrt{2\sigma^2}} - z_{1-\alpha_R}) - .5\right) \\
\Phi^{-1}(\frac{p_R}{2} + .5) &= \frac{\delta_E}{\sqrt{2\sigma^2}} - z_{1-\alpha_R} \\
\sqrt{2\sigma^2} &= \frac{\delta_E}{\Phi^{-1}(\frac{p_R}{2} + .5) + z_{1-\alpha_R}} \\
\sigma &= \sqrt{\frac{1}{2}\left(\frac{\delta_E}{\Phi^{-1}(\frac{p_r}{2} + .5) + z_{1-\alpha_R}}\right)^2} \\
\mathit{MDES}^* &= (z_{1-\alpha_R/2} + z_{1-\beta_R})\sqrt{\frac{1}{2}\left(\frac{\delta_E}{\Phi^{-1}(\frac{p_R}{2} + .5) + z_{1-\alpha_R}}\right)^2}.
\end{aligned}$$

Translational Abstract.

Research reproducibility and effect replication has become a topic of major concern throughout the social and behavioral sciences. During the last decade, low replication rates of published research findings became a major issue, leading to the public proclamation of a "replication crisis". Given the importance of the issue, it is crucial to consider which methods researchers use to assess replication success because replication success or failure strongly depends on the chosen measure for assessing correspondence in effect estimates. This article discusses the statistical properties of selected correspondence measures for comparing the results of two independent replication studies. One of the measures is the novel correspondence test that combines the outcomes of a difference and significance test. To facilitate the computation of correspondence measures and the calculation of sample size requirement to achieve a predetermined probability for replication success R and Stata tools are provided.

R functions

Stata functions