Journal of Methods and Measurement in the Social Sciences,
Vol. 13, No.2, 41-69, 2022

Using Correspondence Tests to Assess Replicability of Open Science Collaboration Results: Inferences from a SMART Design-based, Meta-analytic Approach

William H. Yeaton
Florida State University

Gertrudes Velasquez
Florida State University

The logic of the SMART (Sequential Multiple Assignment Randomization Trial) design was applied to assess the replicability of original-replicate study pairs for Open Science Collaboration (OSC) intervention studies. Within SMART, we utilized both subtests of the correspondence test (CT) to assess study pair comparability. First, we implemented a CT difference test to determine if an effect size difference between the original study and its replicate pair was close to zero; second, we implemented a CT equivalence test to determine if the effect size difference of that study pair was within a designated threshold. In Stage 1 of SMART, each study pair was randomly assigned to one of two alphas (.01 and .05), thereby creating two, probabilistically similar subsets of study pairs. Within each alpha subset, successful difference tests (test of significance was not significantly different than zero) and unsuccessful difference tests were then determined. In Stage 2 of SMART, study pairs in each combination of alpha level and successful or unsuccessful difference tests were randomly assigned to one of two thresholds (±.25 *SD*, ±.50 *SD*). Equivalence tests were then conducted for all study pairs in each of these four subsets. Successful equivalence occurred when the distance between an original and its replicate pair was statistically significantly less than a given threshold. Thus, initial randomization followed by a second randomization was used to gauge comparability of each OSC original study and its replicate, for two alpha levels and two thresholds. In the first set of results, to mirror the common replicability assessment case in which only difference tests are conducted, 16 of 96 difference tests (16.7%) conducted in Stage 1 were successful. In the second set of results, for initially successful difference tests, two thresholds were used to determine the percent of study pairs that also passed the equivalence test. Depending on α and threshold, 8.0%-13.8% of studies successfully passed both difference and equivalence CT subtests. In the third set of results, using SMART, after randomization to two α-values and contingent on success or lack of success of a difference test, study pairs were randomized to two thresholds and a statistical test of equivalence conducted. Using meta-analysis methods within SMART-based subsets of study pairs, original-replicate average effect size differences were compared to differences in the second set of results. We found a similarly-sized 10.3% of study pairs passed both CT subtests (nine of 87 study pairs successfully passed the difference test at either alpha and successfully passed the equivalence test at either threshold). Reflecting the importance of incorporating both CT subtests, of 16 study pairs that initially passed the difference test, nearly half (43.7%) failed the equivalence test. Thus, for CT success, we found that α choice had little impact, while threshold choice was an important determinant. In all three sets of results, the percent of successful replications was substantially smaller than the 36% of OSC replicates that were statistically significant. To confirm this study's replicability, we found very similar patterns of CT success and lack of success for two, SMART-based tables, one for alpha = .01 and one for alpha = .05. The current research extends the utility of CT established by Steiner and Wong (2018) in which results were based on simulation data.

Keywords: Correspondence test, Open Science Collaboration, Replication, SMART

Despite the central role of replication to the foundation of scientific inquiry, favorable evidence for replicability has been mixed. Worse, within the last decade multiple sources have characterized replication as in "crisis" (e.g., Anderson and Maxwell, 2016; Gelman & Vazire, 2021; Williams, 2019). This characterization reflects the selection of a wide range of criteria to gauge replicability, and currently no clear consensus exists that identifies which measures best judge replication success. Current measures used to assess replicability include those in the exemplary Open Science Collaboration (OSC) whereby statistical tests were used to determine the percent of replicates studies that were significant in the original direction (Open Science Collaboration, 2015) and Bayesian methods (Camerer et al., 2018) were used to judge whether replicate-original study differences yielded evidence favoring an effect or favoring the null hypothesis of no effect.

Given the lack of a single, best measure of replicability, Anderson and Maxwell (2016) suggested that researchers report multiple methods. The authors described five tactics in which confidence intervals might be utilized to better judge the consistency or inconsistency of effect-size differences between original and replicate pairs. Fabrigar and Wegener (2016) preferred to focus upon meta-analysis procedures that combined effect sizes of original and replicate studies. Adding to the set of elements that influence replication success, Anderson and Maxwell (2017) noted that effect sizes in original studies were likely to be *over*estimates of population effects, given that publication bias existed in the original study but not in the replicate. As a partial remedy, Schäfer and Schwarz (2019) argued that preregistered, replication studies would be less likely to show effect size inflation and that sample sizes in replication studies should be increased to reflect statistical power comparable to that of originals (see also, Schauer and Hedges, 2020).

Schauer and Hedges (2021) examined multiple replicability measures to address comparability of original-replicate pairs. These authors introduced false success rate and false failure rates to assess replicability for multiple equivalence thresholds (.20, .50, and .80 *SD*s). Using OSC data, for both confidence interval overlap and prediction interval, they found high, false success and false failure rates. For correspondence in sign and statistical significance, only for power > .50 in the original study did higher power in the replicate reduce false failure rate.

As Schauer and Hedges (2021) had placed particular emphasis on the role of statistical power, they noted that small sample size was especially salient when a single, original-replication study pair was examined. Fortunately, this "n = 2" argument was avoided in the current study, given the large set of original study-replicate pairs (100) in the OSC database.

Open Science Collaboration

Perhaps the most extensive, direct replication project was performed by the Open Science Collaboration (2015) research team. A total of 100 experimental and correlational psychology study pairs were analyzed, and 270 authors contributed to the project. Replication materials and procedures were intended to duplicate the design, analysis, and conduct of each original study.

In OSC, three journals were selected to represent prominent research in cognitive and social-personality psychology. Given the lack of consensus of a best, single indicator of replicability, multiple measures of replication success were reported, including: statistical significance of the replicate (in the same direction); determination of whether the original effect size was within the 95% confidence interval of the replicate; the difference between average original and average replicate effect sizes; a meta-analytic estimate that combined effect sizes for each original-replicate pair; subjective assessment of replicability; and correlates of replicability (e.g., was the effect size magnitude of the original study predictive of replication success?).

Despite redundancy in methods and materials, "[a] large portion of replications produced weaker evidence for the original findings" (Open Science Collaboration 2015, p. 943). A scant 36% of replicates were statistically significant in the original direction, only 47% of the 95% confidence intervals of replicate studies contained the original's effect size, and the average magnitude of replication effect size was reduced to approximately half that of the original's effect size. These findings added fuel to the pessimism regarding replication success.

An effect size measure used for subtests of the correspondence test

The current paper used the effect size difference between an original OSC study and its replicate as the basis for assessing replicability. Two tests were conducted with this effect size difference; a difference test and an equivalence test. Steiner and Wong (2018) incorporated both tests to establish the correspondence test (CT), and both tests must be successfully passed to successfully achieve correspondence. More formally: 1) the difference between the effect size of the original and replicate study is not statistically significantly different from zero; and 2) that same effect size difference is statistically significantly smaller than some *a priori* constant (threshold) as evaluated in the equivalence test. Thus, CT seeks to establish that the effect size difference between an original-replicate pair is small (near zero) and not too large (less than some threshold). While an effect size difference of .25 *SD* may be statistically non-significantly different than zero

(perhaps due to small sample size), its apparent, large magnitude should not be too big (effect size difference must be less than the threshold).

Correspondence test

In the context of an historically static set of methods for conducting replications, building upon previous work by Tryon (2001), the current research extends the Steiner and Wong (2018) approach to a multi-study context as these authors had introduced CT in a within-study comparison (WSC) context to judge outcome similarity for a single RCT and a single quasi-experimental analogue rather than to address comparability of original-replicate results for a large set of studies. A primary strength of CT is that it utilizes two statistical subtests and that low (or high) statistical power is not a simultaneous advantage of both subtests. For CT, the difference test posits a nil difference (literally, a difference of zero) between the original study and the replicate, then judges whether the replicate was "not too far" from the original (success occurs when one "fails to reject" the nil). However, a critical problem with the difference test is that low statistical power may lead to no-difference findings. A conclusion that the difference between the original-replicate pair was not significantly greater than zero could be attributable to small sample size.

To counteract this weakness, an equivalence test is also conducted. This portion of the CT requires an *a priori* threshold. That threshold choice can focus on a theoretical dimension (e.g., Is the quasi-experimental bias too large?) or the choice can address a practical, policy-relevant difference in results (e.g., Are replication results "close enough" to originals? Can we accept replication findings and implement the interventions more generally?).

For an equivalence threshold of ±0.10 *SD*, the composite null hypothesis posits that the original-replicate effect-size difference is greater than or equal to 0.10 *SD*, and less than or equal to -0.10 *SD* (two, one-sided tests). The composite null hypothesis is rejected when the difference is statistically significant, within the ±0.10 *SD* range, thus concluding replication and original ESs are "similar enough." Given that the two subtests effectively counterbalance the impact of low (or high) power, CT's multiple subtest feature also acts as an antidote to the single replicate (n = 2) weakness noted by Schauer and Hedges (2021).

In summary, to assert correspondence, the difference test must be statistically non-significant (effect size difference is close to zero), and the equivalence test must be statistically significant (effect size difference is not too large). Inclusion of both difference and equivalence subtests in CT is intended to not allow a given study feature to simultaneously advantage both subtests. For equivalence, low power works against statistical

significance; for difference, low power works in support of comparability (as the aim is to fail to reject the nil hypothesis).

Four correspondence test outcome categories

The CT is successful when both difference and equivalence tests are successful. Of the four possible combinations of test results, two may lead to ambiguity. When the difference test is successful (non-significant difference perhaps due to small sample size) but the equivalence test is not successful (non-significant, perhaps due to a small threshold), the CT yields an "indeterminant" result. This outcome is particularly problematic, since indeterminacy can only be concluded if an equivalence test has been conducted; unfortunately, most researchers do not test equivalency. Without an assessment of equivalency, a successful difference test is deficient.

The second ambiguous case occurs when the difference test is unsuccessful and statistically significant (there is large sample size which is bad for a successful difference test) and the equivalence test is successful (there is a large threshold which is good for a significant equivalence test). Here, a "trivial difference" results; the difference between original and replicate may be small, but CT's conclusion is suspect due to large sample size and large threshold. Case four is unambiguous; CT fall into the "difference" category as both CT tests are unsuccessful. The difference test is significant (the original and replicate are too far apart), and the equivalence test is nonsignificant (original and replicate are not close enough to each other).

This study closely examines two of these four cases. First, it addresses factors that contribute to successful correspondence. Second, it reveals the extent of mis-inference stemming from successful difference tests followed by unsuccessful equivalence tests (indeterminacy). Technical details regarding the equivalence test are presented in a later section.

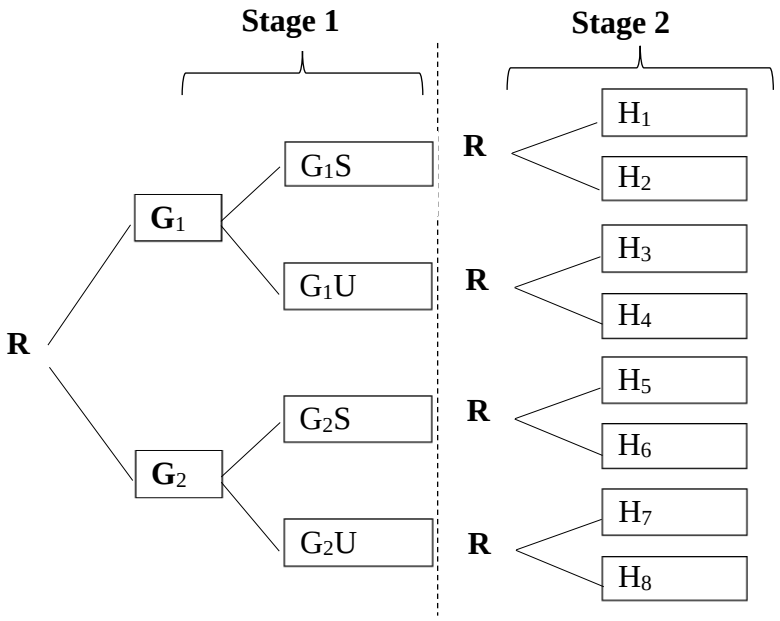
Overview of the SMART design

A SMART (Sequential Multiple Assignment Randomized Trial) design (Murphy, 2005) applies two randomizations, in sequence. In SMART, a second randomization follows the usual random allocation of an experiment. After initial randomization, effect size results for treatment and control group participants are compared to a pre-established standard. Scores above (or below) the standard are judged successful and those below (or above) are deemed unsuccessful. For all four subgroups, random allocation is applied, resulting in eight equivalent subgroups.

In the SMART/CORR application of the general SMART design used in this study, the correspondence test was implemented. In SMART/CORR,

individual "participants" are original-study replicate pairs, not persons, as reflected in the graphic, below. Two subgroups of OSC original-replicate study pairs, G_1 and G_2, result from random assignment (R) in Stage 1. Based on some quantitative cut score, successful (S) and unsuccessful (U) subgroups emerge for both G_1 and G_2 (yielding four subgroups: G_1S, G_1U; G_2S, G_2U). Study pairs in each of these four subgroups are then randomized (R) during Stage 2, resulting in eight subgroups, H_1–H_8, each yielding either a successful or an unsuccessful outcome. (Note: randomized, subset pairs such as H_1 and H_2 are probabilistically similar; subset pairs such as H_1 and H_3 may not be not similar.)

Thus, in the SMART/CORR, individual, original-replicate OSC study pairs serve as units of analysis rather than individual participants. This SMART/CORR adaptive strategy appears as:



The original-replicate pairs of the current study utilize OSC treatment studies which were randomly divided into two groups (G_1 and G_2) during Stage 1. Based on a prespecified alpha level (for the difference test), successful (S) and unsuccessful (U) subgroups emerge for both G_1 and G_2), yielding four subgroups. OSC study pairs in each of these subgroups were rerandomized (R) during Stage 2, resulting in eight subgroups, H_1–H_8, each yielding successful or unsuccessful effect size difference outcomes according to a prespecified threshold (for the equivalence test).

Most SMARTs determine initial treatment success during Stage 1 based on the size of the difference between treatment and control group. Second-stage interventions are then used to maintain or alter the course of initial, successful and unsuccessful treatment. However, in this study, the

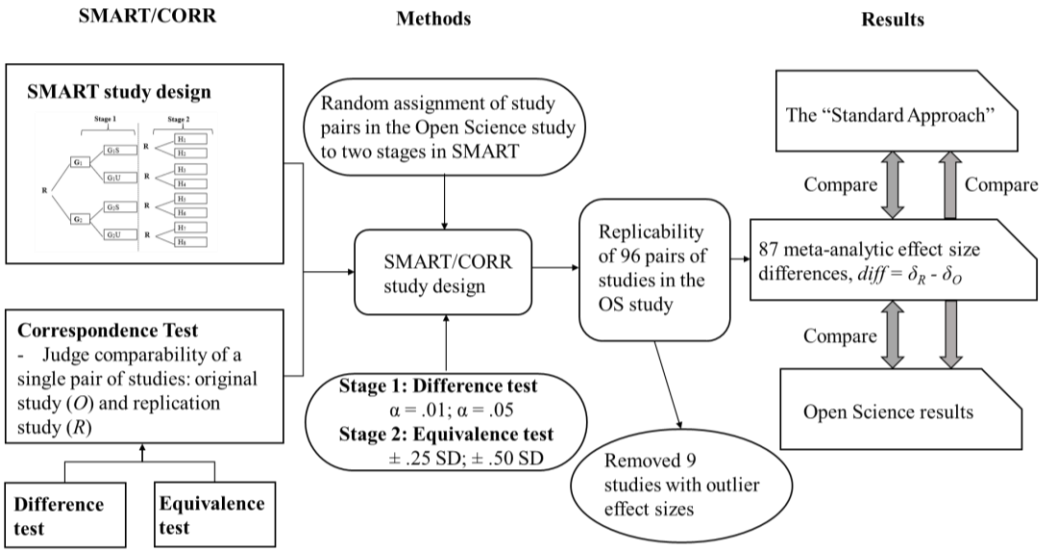
definition of success at each stage was determined by the results of a statistical test. We randomly assigned study pairs to two, different alpha levels in Stage 1 (for the difference test) and then established success for study pairs in each subgroup. Subsequently, we randomly assigned study pairs to two, different thresholds in Stage 2 (for the equivalence test) and then established success within each subgroup.

Thus, in SMART/CORR, one goal was to assess the impact of randomly assigned conditions rather than treatments where conditions (alpha levels and thresholds were assigned to study pairs, not to individuals. In Stage 1, each study pair was randomized to either a .05 or a .01 α-level to test differences for statistical significance. Study pairs were categorized as having successfully or unsuccessfully passed difference tests, and meta-analytic methods were used to calculate average effect size differences. We expected that smaller *p*-values would yield more successful difference tests (more statistically non-significant differences).

In Stage 2, equivalence thresholds of ±.25 *SD* or ±.50 *SD* were randomized to study pairs for successful or unsuccessful difference tests, for two, Stage 1, α-values. An equivalence test was conducted for each resulting subgroup, again using meta-analytic methods to assess average effects. It was expected that larger thresholds would yield more successful equivalence tests (more statistically significant differences within a given threshold). Successful CTs were based on coincident success of both difference and equivalence findings. Of particular interest was the combination of results when the difference test was successful and the equivalence test was unsuccessful. These results reflected indeterminacy and identified instances in which replication conclusions based only on difference tests may potentially be flawed. In Figure 1, an overview of SMART/CORR is provided, along with an outline of related methods and results.

In the far-right portion of Figure 1 labelled "Results," in the top tab, the "Standard Approach" yielded 87 results with a successful or unsuccessful correspondence test for each original-replicate OSC pair, using a given α-value for the difference test and a given threshold for the equivalence test. In the middle tab, SMART/CORR yielded measures of replicability success for multiple subsets of randomly assigned study pairs. SMART/CORR applied meta-analytic techniques to assess effect size in subsets and to gauge the impact of α-value and threshold on success of the correspondence test. In the bottom tab, OSC (2015) results were reported.

Figure 1. Overview of SMART/CORR study design.



Choice of α-values for difference test and thresholds for equivalence test

For SMART/CORR difference tests, two, common but arbitrary α-levels, .05 and .01, were chosen to test statistical significance. Each alpha value produced a 2 x 2 correspondence table in which individual cells represented each of the four CT results. Second, while *a priori* choice of thresholds for the equivalence test was arbitrary (Kruschke, 2018), the initially selected threshold values, ±0.10 *SD* and ±0.20 *SD*, require further explanation.

The value ±0.05 *SD* used by the What Works Clearinghouse for baseline equivalence in an RCT (Institute of Education Sciences, 2020) was first considered as a stringent threshold. However, this small value was judged to likely produce too few, successful CT tests. Instead, the ±0.10 *SD* threshold was tentatively chosen.

Kruschke (2018), in designating a range of equivalence thresholds considered "good enough" for practical, policy-related decisions (a region of practical equivalence: ROPE), characterized an effect size of 0.20 *SD* as small (see Cohen, 1988), then argued for an arbitrary threshold equal to half of this small effect. In a meta-analysis of within-study comparison (WSC) findings that compared results in RCT and RD designs addressing the same research question, Chaplin et al., (2018) typically used 0.10 *SD* as a gauge of practical, outcome consistency across designs. Steiner and Wong (2018) utilized *SD* values ranging from 0.10 to 0.60 *SD* in their simulation studies but judged *SD*s of 0.10 or lower to be "reasonable" to judge equivalence "because it minimizes the probability of an incorrect equivalence

conclusion." (p. 28) To assess false failure and false success replication rates in OSC data, Schauer and Hedges (2021) used substantially-sized thresholds: .20, .50., and .80 *SD*s. Clearly, threshold magnitude varies widely depending upon researcher preference and study-specific questions.

Given unclear guidance regarding threshold choice based on the existing literature, an empirically-based step was used to better guide this choice. Consistent with Bonett's (2020) recommendation, multiple thresholds were considered in the current study. This empirically-based step reduced the chances that only a very small percent of replicates would pass both the difference subtest and the equivalence subtest of a CT, with the ±0.20 *SD* lenient threshold.

Preliminary analyses were performed, and equivalence test results are shown for both successful and unsuccessful difference tests (Table 1). The category "conceptual" reflected reliance on literature-based, threshold recommendations (±.10 and ±.20 *SD*). As the number of successful difference study pairs (at $\alpha = .05$) with a successful equivalence test (at $\alpha = .05$) increased between ±.20 and ±.25 *SD*, the slightly larger ±.25 *SD* was used as a stringent threshold; twice that value, ±.50 *SD*, was chosen as a lenient threshold, along with a still larger ±.75 *SD* threshold. The term "empirical" designated the three largest thresholds used to determine counts of successful and unsuccessful difference and equivalence tests.

A theoretical framework for the application of design to a replication context

Recently, Steiner et al., (2019) presented a research design framework within which one could assess replication efforts. After defining replication as "a research design that tests whether two (or more) studies produce the same causal effect within the limits of sampling error" (p. 280), the authors provide five assumptions under which successful replication is likely to occur. These assumptions include: 1) close comparability of contrasting conditions and measures used in the original and replicate; 2) the same causal estimand was employed to calculate effects in both original and replicate; 3) correct identification of causal estimands (e.g., randomization correctly conducted, little to no differential attrition in both studies); 4) unbiased estimation of causal estimands; and 5) correct reporting of study outcomes.

This theoretical framework was applied to the current context in which methodological redundancy of the original-replicate OSC study pairs had been planned. In these direct replications, researchers went to great lengths to ensure that general methods, particular interventions, and specific measures used in original studies were closely followed in replicates. The analyses in this study yielded results redundant with or calculable from statistics reported in the publicly available, OSC archive.

Table 1
Cells display cumulative number and cumulative percent of successful equivalence tests, for conceptual- and empirical-based thresholds, based on successful (n = 16) and unsuccessful (n = 80) difference tests

Equivalence Test	Difference Test[a]: n=96	
Threshold (*SD*)	Successful: n=16 (%)	Unsuccessful: n=80 (%)
Conceptual		
±.10	0 (0.0)	0 (0.0)
±.20	1 (6.3)	3 (3.8)
Empirical[b]		
±.25	4 (25.0)	5 (6.3)
±.50	10 (62.5)	25 (31.3)
±.75	13 (81.3)	40 (50.0)

Note: Preliminary results were conducted to inform threshold choice. Equivalence tests stem from two, one-sided, 97.5% confidence intervals. *SD* = standard deviation.
[a] Conducted at $\alpha = .05$, across replicates.
[b] Since counts are cumulative, to ensure independence for Row 1 and Row 2, Fisher's Exact Test was conducted ($p = .393$) on empirical counts in the 2 x 2 matrix $\begin{bmatrix} 4 & 5 \\ 6 & 20 \end{bmatrix}$.

Unbiased estimates from well-done experiments and estimates from replicate experiments were calculated in the same way (assumption 4). However, there was a slight possibility that differences in original and replicate participants and in the precise manner that measures were implemented in original and replicate settings might have altered study outcomes (assumption 1). While the effect size causal estimand (an average treatment effect, when experiments were conducted) was the same in the OSC replicate pairs as in this study (assumption 2), a primary purpose was to determine if the CT yielded replicability findings consistent with those of OSC. Importantly, identification (assumption 3) was likely established, as selection bias was not in question; most originals and their replicate pairs were experiments. Reporting quality of the OSC was high; few instances of missing data were present in the OSC (assumption 5). Of the original studies in the OSC database, only three did not provide sufficient information to calculate effect sizes. In summary, when efforts were made to directly replicate original studies, it is reasonable to conclude that replication assumptions were closely adhered to.

In this application of SMART, random assignment created covariate equivalence by expectation in Stage 1. However, inference made by comparing successful or unsuccessful Stage 1 subgroups (e.g., G_1S and G_2S) must be made cautiously, as covariate values (e.g., follow-up length) may be confounded across studies.

In Stage 2, comparisons within each of the four randomized pairs of conditions were likely to produce unbiased estimates. For example, a comparison of H_1 and H_2 estimates will probabilistically be unbiased. However, comparisons of other group results (e.g., subsets H_1 versus H_5) may not be unbiased due to between-condition confounding (though both groups were successful in Stage 1, success reflected different α-level conditions).

Study Aims

This research served multiple purposes. First, empirical data based on CT were used to evaluate replicability of the large set of OSC studies. Our empirical results extended CT's range of applicability previously based on simulation findings reported by Steiner and Wong (2018).

The second purpose reflected the contingent nature of SMART by determining the rate of indeterminacy; first a successful difference test occurs, then an unsuccessful equivalence test follows. This case is particularly important as most researchers conduct only a difference test.

A third purpose of this study was enabled by the two-stage structure of SMART. The random allocation features of SMART yielded unconfounded contrasts between probabilistically similar subsets of study pairs for successful and unsuccessful difference tests with different α-values (Stage 1) and for successful and unsuccessful equivalence tests with different thresholds (Stage 2). Further details regarding the SMART/CORR application are provided below.

Linking study aims to study practices

The first aim was achieved by determining the probability of successful and unsuccessful difference and equivalence tests in OSC study pairs. These subtest results were subsequently used to determine the rate in which study pairs fell into four CT outcome categories.

The second aim focused upon the probability of unsuccessful equivalence tests. As noted, indeterminacy occurs when the difference test is successful and the equivalence test is unsuccessful. As this difference-test-only case is common, a gauge of the degree of typical mis-inference becomes possible by adding the equivalency test.

The third study aim was accomplished via the SMART/CORR design. How often is the difference test successful, for different alphas (Stage 1).

How often is the equivalence test successful, for different thresholds, contingent upon successful and unsuccessful difference study pairs (Stage 2).

Methods

Re-analysis of Open Science Collaboration data

Data for this research were obtained from the Open Science Framework (OSF) (https://osf.io/fgjvw/). In this SMART/CORR application, for the 100 pairs of original and replication studies used in the OSC database, 96 of these pairs provided effect size measures or relevant statistics to enable calculation of standardized mean differences for the *d* statistic. The statistical software *R* v. 4.0.1 (R Core Team, 2021) was utilized to randomly assign each OSC original study and its paired replicate to conditions (two α-values and two equivalence thresholds), to generate graphical displays, and to conduct statistical analyses. For each randomization, two sets of study pairs were created (without replacement) by applying the *set. seed* function along with the *sample* function in *R*.

In summary, in this SMART/CORR application of SMART, the utility of CT was extended to a multiple-study context rather than limited to a single original study and its replicate pair. To gauge successful replicability: 1) a CT was conducted for each OSC original study-replicate pair to assess the replicability of the entire dataset; 2) SMART/CORR was applied to assess the impact of α-value and threshold upon successful difference and equivalence tests for probabilistically similar subsets of study pairs.

Identifying and addressing outliers in OSC data

In SMART/CORR, to minimize the impact of outliers on cell means and differences between cell means from the set of 96 original-replicate OSC pairs, outliers were removed. Diagnostic methods described by Viechtbauer and Cheung (2010) were employed to detect outlier and influential cases in the meta-analytic, random-effects models conducted at Stages 1 and 2 of SMART/CORR. At each stage, influential studies were first identified using case deletion diagnostics based on Cook's distance and DFFITS statistics (regression-based measures for assessing potential outlier or influential cases adapted to the meta-analytic context). Studentized deleted residual statistics were then examined to identify outliers from among these influential studies. Effect sizes with residual statistics values greater than the absolute value of 1.96 in either Stage 1 or Stage 2 of the SMART/CORR were identified as outliers, resulting in eight studies which met this criterion. A ninth study was identified as an outlier as it had the same

studentized residual value (-1.94) in Stage 1 as a study with values of -1.94 and -2.01 in the two SMART/CORR stages.

In the primary analyses shown in Figure 3 and Table 2, these nine outliers were omitted from the set of 96 study pairs. A sensitivity analysis was conducted in which results for the 87 study pairs were compared with results for all 96 study pairs. Differences in tests of statistical significance and effect-size estimates for the resulting subsets of study pairs were noted.

Calculation of effect size difference and standard error for OSC study pairs

Here, computational details are provided for the specific procedures implemented to calculate effect size difference and standard error, for each OSC original study and its replicate.

Initially, the Fisher's *z* correlations of original and replication studies reported in the Open Science dataset were transformed to *r* correlations and then to *d* effect sizes. A *d* effect size estimate was calculated for each original and replication study in 96 pairs of studies. For each original and replication pair of studies, the difference between *d* effect size estimates is calculated by $diff = d_R - d_O$ where d_R and d_O are *d* effect size estimates in each pair. The variance of *diff* is given by $v_{diff} = \frac{n_R + n_O}{n_R n_O} + \frac{diff^2}{2(n_R + n_O)}$, where n_R and n_O are the sample sizes of the groups in the replication and original studies (Borenstein, 2009). The standard error of *diff* is the square root of v_{diff}, $se_{diff} = \sqrt{v_{diff}}$. The *d*-based, effect size results were independently verified by comparing them to those in the OSC dataset. The weighted average of *d*, based on 96 studies, was -0.46. The absolute value of this *d*, when transformed to an *r* effect size, yielded 0.224, which was close to the overall *r*-based difference of 0.206 found in OSC, based on 97 studies.

SMART/CORR application of effect size difference measures in Stage 1 and Stage 2

A two-stage process was implemented to randomly assign OSC, original-replicate pairs of intervention studies to cells in the SMART/CORR. In Stage 1, 96 study pairs were randomly assigned to the two α conditions (.01 and .05) for the difference-test component of the CT. The nil and alternate hypotheses of the difference test are given by H_O: $\delta_R - \delta_O = 0$ and H_A: $\delta_R - \delta_O \neq 0$, where δ_R and δ_O represent population *d* effect sizes for the replication and original studies, respectively. The *t*-test for the difference test is given by $t = \frac{diff}{se_{diff}}$. For each study pair, a difference test was conducted at the assigned α, .01 or .05. Failure to reject the nil hypothesis yielded a "successful" difference test.

In Stage 2, study pairs in the "successful" and "unsuccessful" cells of the two α conditions of the difference test were randomly assigned to stringent and lenient threshold conditions of the equivalence-test component of CT. The composite null hypothesis for equivalence states that the absolute difference between the replication and original effects is larger than or equal to the selected threshold h: H_O:$|\delta_R - \delta_O| \geq h$. This null hypothesis is also conceptualized as two one-side null hypotheses: H_{O1}: $\delta_R - \delta_O \leq -h$ and H_{O1}: $\delta_R - \delta_O \geq h$, where $-h$ and h are the lower and upper bounds of the selected threshold. One test rejects the null if the difference is larger than the lower bound of the threshold: H_{A1}: $\delta_R - \delta_O > -h$ and the other rejects the null if the difference is smaller than the upper bound of the threshold: H_{A2}: $\delta_R - \delta_O < h$. The t-test statistics for the two one-sided tests are given by $t_L = \frac{diff - (-h)}{se_{diff}}$ and $t_U = \frac{diff - h}{se_{diff}}$. For study pairs, the equivalence test was conducted at α = .05. Rejections of both null hypotheses establish original-replicate equivalence.

For different α-values and thresholds, successful and unsuccessful outcomes for difference (Stage 1) and equivalence (Stage 2) subtests of CT are shown within relevant cells of SMART/CORR. For samples of study pairs within each cell, meta-analytic methods were used to calculate standardized, mean effect size (d) differences and 95% confidence intervals.

Results

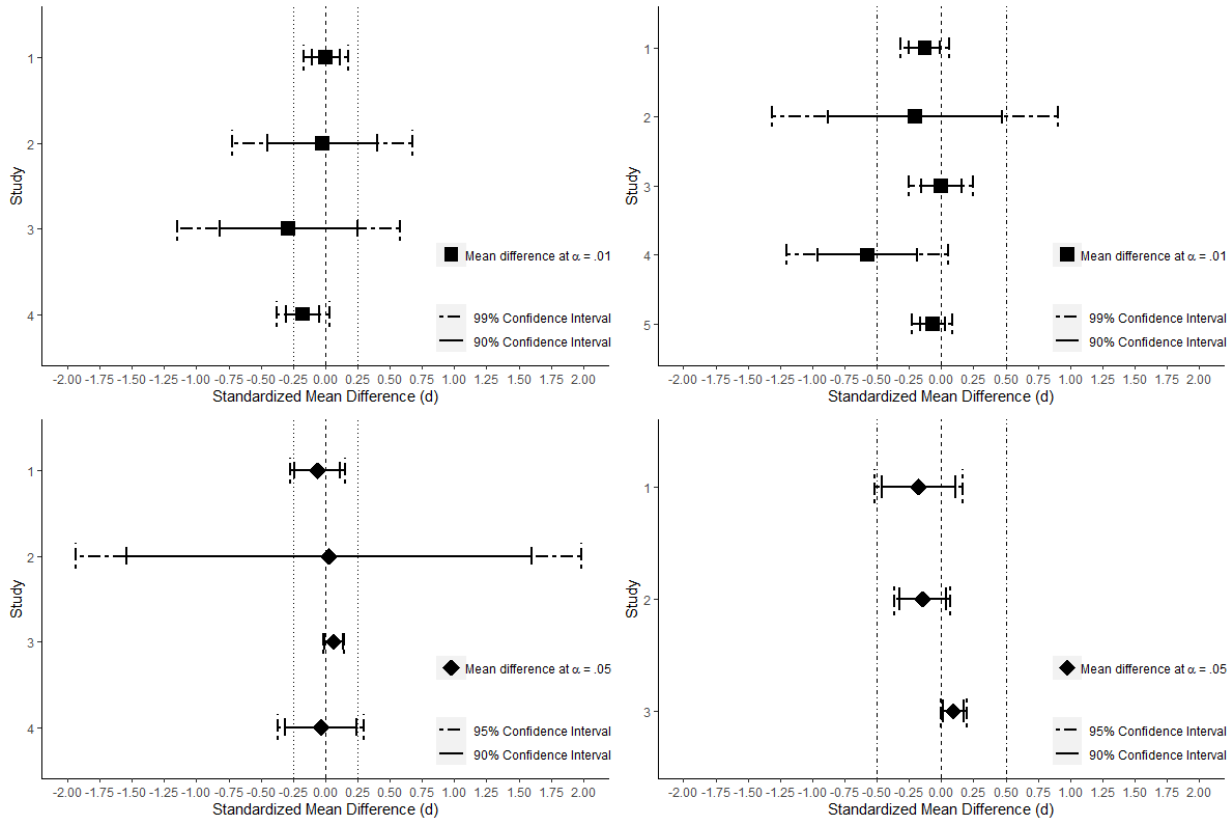
Preliminary analyses: using successful difference tests to explore threshold choice

For preliminary analyses (n = 96), results for 16 study pairs (16.7%) that had successfully passed the difference test at α = .05 are displayed in Figure 2. The primary purpose of the initial analyses was to provide an empirical basis for subsequent choice of thresholds as these analyses established the number of successful equivalence tests for different combinations of alphas and thresholds. This preliminary step aimed to avoid choice of thresholds that led to few, successful equivalence tests and, therefore, few successful CTs.

The four graphics in Figure 2 portray effect size difference for two, randomly assigned α-values and two, randomly assigned thresholds. In the rows, difference tests were conducted at either α = .01 (boxes, plots in row one) or at α = .05 (diamonds, plots in row two). In the columns, equivalence tests were conducted at either ±0.25 *SD* in column one or ±0.50 SD in column two; vertical lines represent these thresholds. The dashed, horizontal lines represent 99% (row one) and 95% confidence intervals (row 2) for nil-hypotheses that test difference. The solid, horizontal lines (row

one and row two) represent 90% confidence intervals for two, one-sided tests for equivalence, at α = .05.

Figure 2. Initial examination of empirical mean differences (replication ES – original ES) and confidence intervals yielded 16 study pairs passing the difference test (α = .05). Equivalence tests were then conducted on these successful difference test study pairs.



Each α-value was randomly assigned to 48 study pairs. Successful difference test study pairs for both α = .05 and α = .01 were identified, followed by random allocation to equivalence tests at ±0.25 *SD* and ±0.50 *SD* thresholds. (Naturally, preliminary analyses that included study pairs for difference tests successful at α = .05 would also be successful at α = .01.)

The non-statistically-significant effect size difference between these two, alpha-level subsets of study pairs (results not shown) reflected successful random allocation. In all but three cases, differences were negative, reflecting smaller effects in replicates than in originals.

Across the four quadrants of Figure 2, nine of the 16 study pairs (56.3%) that had produced a successful, difference test were also successfully equivalent. As expected, the largest number of study pairs (n = 5, upper right quadrant) fell within the CT cell that combined the more lenient

threshold (±0.50 *SD*) and the smaller alpha (.01). Nine study pairs fell within the .01 alpha (row 1) while seven fell within the .05 alpha (row 2). Eight study pairs fell in both column 1 (threshold ±.25) and column 2 (threshold ±.50).

As summarized in Table 1, for 96 study pairs, it was possible to determine how many of the 16 successful difference test study pairs would then successfully pass the equivalence test for two conceptual and three empirical thresholds. Starting at ±0.20 *SD*, larger thresholds substantially increased the number study pairs that successfully passed the equivalence test.

For conceptual thresholds, using two, one-sided tests with α = .025, one study fell within the ±0.20 *SD* threshold, while none fell within ±0.10 *SD*. A slightly larger, empirical threshold (±0.25 *SD*) yielded additional, successful equivalence tests (n = 4; 25%). A total of 10 of 16 study pair differences (62.5%) fell within the ±0.50 *SD* threshold, and 13 of 16 (81.3%) fell inside the ±0.75 *SD* threshold. Thus, in these 16 successful difference study pairs, for thresholds ranging from ±.10 to ±.75, the percent of study pairs that successfully passed the equivalence test ranged from 0.0% to 81.3%. Fisher's Exact Test found no association between counts of successful and unsuccessful outcomes for equivalence and difference tests (p = .393).

For unsuccessful difference tests, the number and percent of successful equivalence tests was determined for these same thresholds: ±.10 (0/80 = 0%); ± .20 (3/80 = 3.8%); ±0.25 *SD* (5/80 = 6.3%); ±0.50 *SD* (25/80 = 31.3%); and ±0.75 *SD* (40/80 = 50.0%). As threshold increased, the percent of unsuccessful equivalence tests also increased. For each of three empirically-based thresholds, percent of successful equivalence tests was substantially larger for replicates exhibiting a successful difference test vs. those with an unsuccessful difference test.

Results for "standard approach," using 87 correspondence tests, without SMART

Without SMART randomization, the "standard approach" results (not shown in tables or figures) reflected the usual way in which a CT would be conducted. That is, the first study pair was classified as successfully or unsuccessfully passing the CT, and this process was repeated for each of the remaining 86 study pairs. Results were based on an aggregate of these 87 study pairs. By way of validation relative to the 97 OSC study pairs, when the r-based difference between the sets of original (-0.403) and replicate studies (-0.197) was converted to a d (-0.421), that d result was comparable to the overall difference, d = -0.450, based on 87 study pairs.

A count of successful CT tests was made for each alpha-threshold combination. For ±0.25 *SD*, for an α = .01 difference test, seven study pairs

passed the CT; for that same strict threshold, seven study pairs also passed the CT, for α = .05. For a lenient ±0.50 *SD* threshold, counts were 12 and 11, for α = .01 and .05, respectively. For each of the four combinations, the average effect size difference between original and replicate studies was not statistically significantly different than zero. In the case least likely to yield replication success (α = .05, threshold = ±0.25 *SD*), seven study pairs (8.0%) passed CT; in the context most favorable to replication success (α = .01, threshold = ±0.50 *SD*), 12 study pairs (13.8%) passed CT.

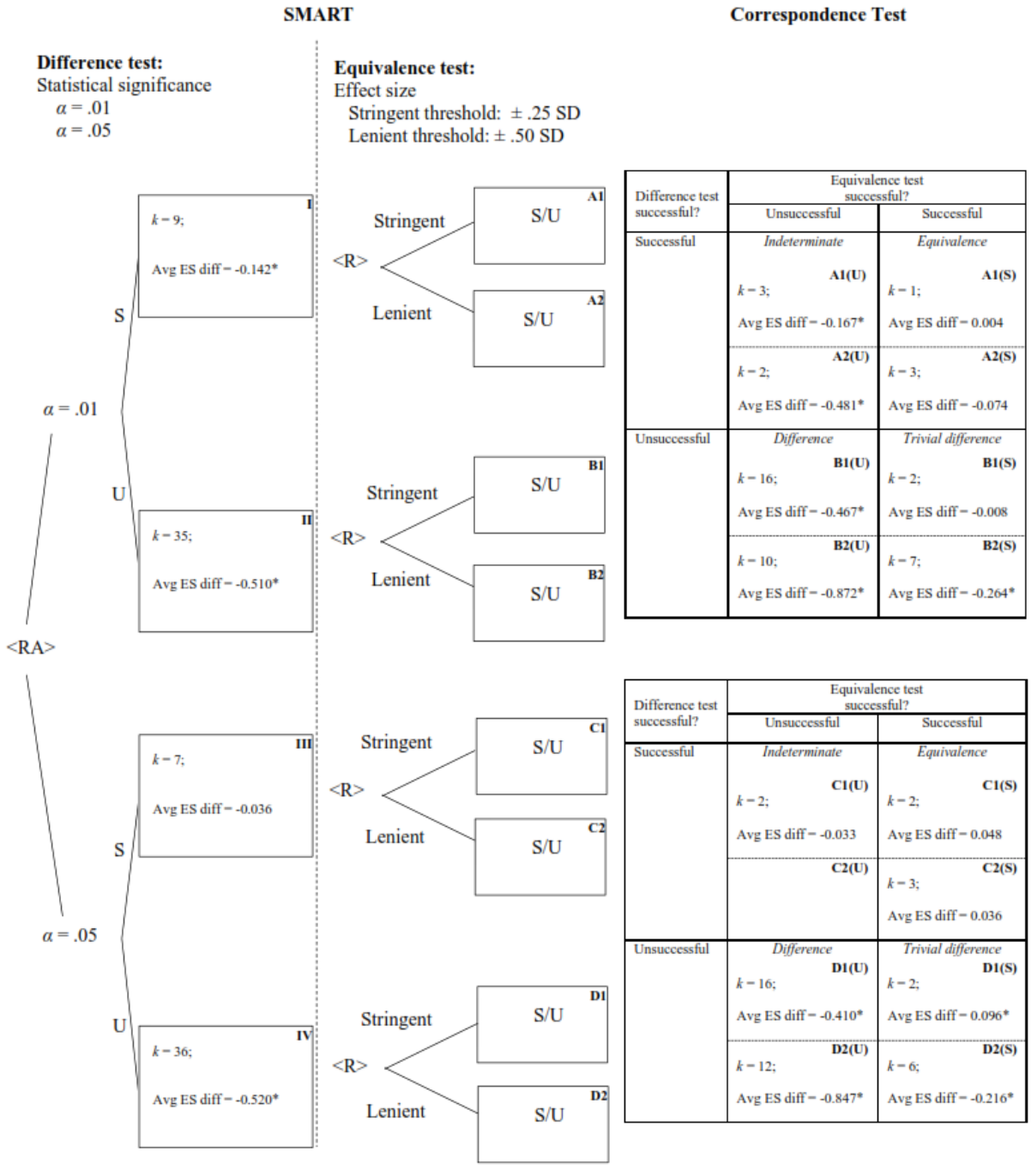
Results for subsets of study pairs in CT cells, using SMART/CORR

The SMART/CORR design flow of 87 original-replicate study pairs is displayed in Figure 3. This flow followed the structure of SMART/CORR in which random allocation occurred twice, once per stage. Two CT tables were created, one using α = .01 and an analogous table using α = .05. Cellular results within each table were based on successful and unsuccessful tests of equivalence for two thresholds, depending upon successful and unsuccessful difference tests. SMART/CORR's contingent structure allowed an assessment of the probability of the critical case of indeterminacy (a successful difference test followed by an unsuccessful equivalence test).

In Figure 3, we used the following notation: Avg ES diff was the average effect size difference between each pair of replication and original study (subtraction in this order), CI was confidence interval, <R> was randomization, S was successful replication, U was unsuccessful replication, cells I– IV were for the difference test in Stage 1, and cells A1 – D2 were for the equivalence test in Stage 2. An * reflected statistical significance, p =.05 or p = .01, for difference and p = .05 for equivalence. For the equivalence test, the null hypothesis was that the absolute effect size difference between replication and original study was greater than the threshold. For the difference test, the nil hypothesis was that the effect size difference between the replication and original study was zero.

To clarify placement of study pairs into the cells of one correspondence table in Figure 3 using SMART/CORR, we describe the allocation sequence that produced entries falling into = .01, nine (I) successfully passed the difference test; for 35, the difference test was unsuccessful. Four (A1) of the nine successful difference test study pairs were randomly assigned to a stringent threshold, whereby one study pair (A1(S)) successfully passed the equivalence test. Five (A2) the top portion of the uppermost table. In Stage 1, 44 study pairs were randomly assigned to .01 alpha, while 43 study pairs were randomly allocated to .05 alpha. Of the 44 study pairs with alpha study pairs were randomly assigned to a lenient threshold, whereby three (A2(S)) successfully passed the equivalence test.

Figure 3. Application of SMART using Steiner and Wong's (2018) correspondence test.



Thus, for both correspondence tables, it was also possible to examine the effect size for each of the four CT categories. For subsets of successful and unsuccessful difference and equivalence test results, meta-analytic techniques were used to compute average original-replicate effect size differences, and inferential statistics were applied.

Within each cell of both correspondence tables, two results were reported: 1) number of original-replicate study pairs falling in a particular cell; and 2) average effect size difference between original and replicate pair ("Avg ES diff"). To illustrate study flow within a table, original-replicate study pairs that fell into the right column (A1(S) and A2(S)) of the upper, right quadrant of the upper correspondence table had a successful α = .01 difference test, used both stringent and lenient thresholds, A1 and A2, and had a successful equivalence subtest.

In Figure 3, 16 study pairs (16/87 = 18.4%) successfully passed the difference test, with similar counts from the two tables: nine from the α = .01 table and seven from the α = .05 table. 26 study pairs (26/87 = 29.9%) successfully passed the equivalence test, with an equal number of study pairs (n = 13) from the α = .01 table and the α = .05 table. For these 26 equivalent study pairs, the stringent test was passed by seven (7/87 = 8.0%), while the lenient test was passed by 19 (19/87 = 21.3%). Of 61 study pairs that unsuccessfully passed the equivalence test, 37 tests were at the stringent level and 24 at the lenient level. Nine study pairs (9/87 = 10.3%) successfully passed both correspondence subtests.

Within Stage 1, for subsets of study pairs with difference tests of α = .01 and α = .05, a similar number of successful (k = 9, k = 7) and unsuccessful (k = 35, k = 36) study pairs were noted. Within Stage 2, for the nine study pairs judged to show correspondence, four (k = 1 + k = 3) were from the upper correspondence table (alpha = .01) and five (k = 2 + k = 3) were from the lower table (alpha = .05). Of the nine successful, difference study pairs, for alpha = .01 (upper table), five were "indeterminant." Of seven successful, difference study pairs, for alpha = .05 (lower table), two were "indeterminant." Together, seven of 16 (43.4%) successful difference study pairs were indeterminant.

In summary, with a focus on the equivalence test and its impact, 29.9% (26/87) of OSC study pairs were equivalent, while 10.3% (9/87) of the OSC study pairs successfully passed the CT. Seven of the 16 (43.4%) successful, difference test study pairs were indeterminant.

Finally, more than half of replicates, 54 (62.1%), were classified as "Difference," with similar counts in the upper (k = 26) and lower (k = 28) tables. Seventeen study pairs (19.5%) fell in the "Trivial difference" category, nine from the upper table and eight from the lower table.

Effect size difference for subsets with successful, unsuccessful CTs, in SMART/CORR

Using Stage 1 results, a random effects model was fitted to address within-study and between-studies error variances. Type I error rates were controlled via the Knapp and Hartung (2003) variance estimator in conjunction with the restricted maximum likelihood (REML) estimation method. The null hypothesis for homogeneous effect sizes was rejected, with $Q(83) = 32625.98$, $p < .001$, prompting planned, pairwise comparisons. Approximately 99.7% of total variation in observed difference effects between replication and original studies ($I^2 = 99.7\%$) can be considered as true variance rather than chance.

In Figure 3, using average effect size in cells of the CT, it was possible to assess the impact of alpha and threshold in greater detail. For study pairs with unsuccessful difference tests in Stage 1, both alpha = .01 and .05 yielded substantially large, average effect size results (-.510 and -.520, respectively) that were statistically significantly different than zero (using a 95% confidence interval). For study pairs with successful difference tests, average effects were closer to zero. For alpha = .05, average effect size (-.036) was not significantly different than zero; for alpha = .01, average effect size (-.142) was significantly different than zero.

Table 2 reflects cell-based results shown in Figure 3 such that associations between alpha and threshold could be examined with respect to average effect size and percent of successful CTs. The comparison between Stage 1 study pairs with successful difference tests (I vs. III), for α-values of .05 or .01, led to a small, non-statistically significant effect size difference of 0.106 *SD* (see "Does alpha matter?"). This small effect size difference was consistent with the small probability of difference in successful difference tests; nine of 44 study pairs (20.5%) for .01 alpha, and seven of 43 (16.3%) for .05 alpha, a nonsignificant difference of 4.2%.

A random effects model was fitted in Stage 2 using the same model fit procedures as in Stage 1. Effect sizes were heterogeneous across all study pairs, $Q(79) = 21511.8$, $p < .001$, prompting planned, pairwise comparisons.

The role of threshold was further examined in Stage 2 by determining both average effect size difference and percent of study pairs that had successfully passed the equivalence test, for a series of four stringent and lenient thresholds (see "Does threshold matter"). With alpha = .01, for A1 vs. A2 and B1 vs. B2, and with alpha = .05, for C1 vs. C2 and D1 vs. D2, differences between lenient and stringent thresholds, for successful equivalent original study and replicate pairs, were large (35.0%, 30.1%, 50.0%, and 22.2%, respectively), generally consistent in their magnitude, and in each case favored the lenient threshold.

Table 2
Differences in average effect size difference, statistical significance, 95% confidence intervals, and proportion successful for SMART/CORR

Comparisons	Differences in Average Effect Size Difference	95% CI	Proportion Successful (%)
Stage 1 difference tests			
Does alpha matter, if difference is S?			
I vs III (S vs S)	0.11	(-0.36, 0.57)	9/44 (20.5) vs. 7/43 (16.3)
Stage 2 equivalence tests			
Does threshold matter, if equivalence is S, U?			
A1 (Stringent) vs A2 (Lenient): difference = S	-0.06	(-0.76, 0.63)	1/4 (25) vs. 3/5 (60)
B1 (Stringent) vs B2 (Lenient): difference = U	-0.2	(-0.51, 0.12)	2/18 (11.1) vs. 7/17 (41.2) *
C1 (Stringent) vs C2 (Lenient): difference = S	-0.06	(-0.84, 0.71)	2/4 (50) vs. 3/3 (100)
D1 (Stringent) vs D2 (Lenient): difference = U	-0.05	(-0.36, 0.26)	2/18 (11.1) vs. 6/18 (33.3)
Aggregate of Stage 2 results			7/44 (15.9) vs. 19/43 (44.2) *
Replication of two correspondence tables			
Equivalence [A1(S) + A2(S)]/44 vs [C1(S) + C2(S)]/43	-0.13	(-0.88, 0.62)	4/44 (9.1) vs. 5/43 (11.6)
Trivial Difference [B1(S) + B2(S)]/44 vs [D1(S) + D2(S)]/43	0.02	(-0.30, .34)	9/44(20.5) vs. 8/43(18.6)
Difference [B1(U) + B2(U)]/44 vs [D1(U) +D2(U)]/43	0.1	(-0.65, 0.86)	26/44(59.1) vs. 28/43(65.1)
Indeterminate [A1(U) + A2(U)]/44 vs [C1(U) + C2(U)]/43	-0.06	(-0.36, 0.25)	5/44(11.4) vs. 2/43(4.7)

Note: ES = effect size; CI = confidence interval; S = successful; U = unsuccessful; *$p < .05$, **$p < .01$; I-IV, A1-D4, from Figure 3. Successful = fulfills requirements of difference or equivalence test. Negative effect size differences reflect larger effect sizes in the original set of study pairs. For example, from Figure 3, I vs II was calculated as: -0.510 – (-0.142) = -0.368.

Thus, in SMART/CORR as in the standard approach, choice of threshold was important for successful CTs. However, in only one instance (B1 stringent vs. B2 lenient, for $\alpha = .01$) was there a statistically significant difference between the percent of successful lenient and stringent equivalence study pairs, likely due to generally small sample sizes (9, 35, 7, and 36) within contrasts. However, when the four stringent versus lenient results were aggregated, the substantial 15.9% vs. 44.2% difference was statistically significant. With respect to average ES, no statistically significant difference occurred in any of the four stringent vs. lenient contrasts.

Agreement between correspondence tables

With both alpha levels in Table 2, there was close effect size agreement between analogous equivalence, indeterminate, difference, and trivial difference cells in the two correspondence tables (see "Replication of two correspondence tables"). No cell in one correspondence table differed from its analogue cell in the second table by more than 0.13 *SD*, and all effect size differences were statistically non-significant. Differences in "proportion successful" for each of the four analogous cells were also small and not statistically significant.

In addition, nested, log-linear models were used to examine the pattern of counts across analogous cells of the correspondence tables in SMART/CORR. Patterns of counts for the difference and equivalence tests were conditionally independent across the two tables ($\chi^2(2) = .258$, $p = .879$). Therefore, cell patterns in the two correspondence tables were successfully replicated using the SMART/CORR design.

A1 vs. B1, C1 vs. D1, A2 vs. B2, and C2 vs. D2 comparisons were not addressed. These contrasts could potentially be assessed using the regression discontinuity design. However, as sample sizes were insufficiently large, RD analyses were not conducted.

Discussion

Compared to the 36% of original-replicate study differences reported in OSC to be statistically significant ($\alpha < .05$ and in the original study's direction), 16.7% (16/96) of study pairs in this study passed the difference test for $\alpha < .05$. Both standard (without randomization to alpha and threshold) and SMART/CORR approaches that used the CT yielded conclusions that were less positive regarding successful replication. For the standard approach in which a simple compilation of successful difference and equivalence tests results was tabulated for 87 OSC study pairs, it was found that 8%-13.8% of study pairs successfully passed CT, depending on

choice of α and threshold. Using a SMART/CORR design, a similarly-sized 10.3% (9/87) of study pairs were successful using either the .01 or .05 alpha or either the ±.25 *SD* or ±.50 *SD* threshold. Thus, both standard and SMART/CORR approaches produced comparable results that were substantially different than OSC results.

From preliminary analyses based on 96 OSC study pairs and from SMART/CORR design analyses based on 87 OSC study pairs (with ES outliers removed), separate difference and equivalence test results were reported. Sixteen study pairs (16.6% of 97, and 18.4% of 87, respectively) passed the difference test. In contrast, the equivalence test was successful in 29.9% (26/87) of study pairs. Nine of 16 (56.3%) original-replicate pairs that passed the difference test passed both CT subtests.

The contingent nature of SMART/CORR (a difference test followed by an equivalence test), enabled determination of the percent of indeterminant cases (a successful difference test followed by an unsuccessful equivalence test). Nearly 50 percent (7/16 = 43.7%) of successful difference tests were indeterminant, calling into question conclusions in which only a difference test had been conducted to assess replicability.

Using a design-based SMART/CORR approach to randomly allocate original replicate study pairs from OSC, it was possible to gauge the independent impact of both alpha and threshold on successful CTs. Generally, alpha had little impact on both the chance of a successful difference test in Stage 1 of the SMART/CORR or the subsequent chance of a successful equivalence test in Stage 2. In contrast, a larger, lenient threshold was consistently linked to a higher percentage of successful equivalence tests and, thus, successful CTs.

SMART/CORR (see Figure 3) yielded two prototypic correspondence tables, one for α = .05 and one for α = .01, by randomly assigning thresholds to after successful or unsuccessful difference tests. These tables yielded cell results with comparable average effect size differences and similar percentages of study pairs successfully passing CT for equivalence, trivial difference, difference, and indeterminate study pairs. A total of 54 of 87 (62.1%) study pairs in these two correspondence tables fell in the difference category. Result similarity across tables (a replication within a study of replication) represents a strong argument for SMART/CORR's veridicality.

Consistency of current results with other re-analyses of Open Science Collaboration results

Etz and Vandekerckhove (2016) reanalyzed OSC results using Bayesian methods. Unlike the current approach which examined effect size differences in original-replicate pairs, the authors examined sets of original and replicate studies while also considering several models of publication bias within original studies. As the authors note, replication studies do not

face the pressure of publication bias, therefore it was expected that replication studies would produce generally smaller effect sizes than originals. The resulting evidence was characterized as "often weak," defined as having a Bayes factor less than 10, and it was noted that most (64%) of the replication and original studies did not provide strong evidence for either the null (zero effect size) or the alternative (non-zero effect size). Furthermore, "In only eight cases (11%) did both the original study and replication study strongly support the alternative hypothesis" (p. 8/12). Very comparable, low levels of replicability success were found in this SMART/CORR research.

Comparing SMART/CORR results with and without outliers

In Stage 1, when nine outlier study pairs were removed from analyses, the Figure 3 comparisons between I vs. II (not shown in Table 2), with an average effect size difference that equals -.368, and between III and IV (also not shown), with average effect size difference that equals -0.485, were both statistically significant. When 9 outliers were included, these effect size differences changed little in size (became -0.369 and -0.502, respectively), though both of the latter results were non-significant. With and or without outliers, the proportion of study pairs for which the difference test was successful was statistically significantly lower for replicates in both I vs. II and III vs. IV comparisons. For the remaining Stage 1 and Stage 2 comparisons, in only one instance was the pattern of statistical significance different for the set of study pairs with and without outliers. In the reduced set of 87 study pairs, the percent successful in the B1 vs. B2 comparison was now statistically significant.

Strengths and limitations

In their within-study comparisons approach, Steiner and Wong (2018) compared RCTs and quasi-experiments and acknowledged the possibility of differing degrees of bias for studies with different designs. In contrast, in this replication study nearly all OSC studies were experiments; thus, the extent of bias was likely low and similar, thereby not a source of confounding in original-replicate pairs.

In the standard method of judging replicability, CT success can be gauged using different combinations of alpha and threshold. However, this comparison style of results for different alphas and thresholds is not based on separate, independent studies. For example, the same, successful difference study pairs at alpha = .05 will be successful at .01, and the same, successful equivalence study pairs using a strict threshold will be successful using a lenient threshold. In contrast, SMART/CORR-based results utilized separate, randomly assigned, independent subsets of study pairs.

Consequently, conclusions regarding the impact of alpha and threshold are not confounded by differences between subsets of study pairs.

The SMART/CORR design's contingent, sequential structure (randomization of both successful and unsuccessful difference study pairs after an initial randomization) provides a natural fit to address important replicability questions. For example: what percent of study pairs have successful equivalence tests after a successful difference test (what percent of studies pass CT?); what percent of study pairs are indeterminant (have successful difference tests followed by unsuccessful equivalence tests)?

The current SMART/CORR approach is not without limitations, however. The choice of two α-values and two thresholds restricts the generality of these findings. While traditional wisdom dictates small thresholds of ±0.10 *SD*s or less (e.g., Chaplin et al., 2018), use of small, conceptually-based thresholds (e.g., ±.10 *SD*) would have placed this study's findings yet further from the evidence for successful replicability shown in the OSC. Evidence from empirically-based thresholds strongly suggested that, even when original study methods were closely adhered to, as in OSC, larger thresholds such as ±.25 *SD* were needed to realistically reflect comparability between original and replicate studies. This limitation should be qualified with the knowledge that CT is more demanding than existing comparability measures, as two filters are applied, and each must be satisfied to yield favorable replication.

In contrast to utilizing many alphas and thresholds, replication conclusions based on a single alpha and threshold would also be subject to criticism. Imagine a study whose original-replicate effect size difference was ±.35 *SD*. For equivalence, if the assigned threshold was ±.50 *SD*, the null hypothesis would be rejected (recall, rejection indicates success). However, this study would not have rejected the null at the ±.25 *SD* threshold, thus leading to an "unsuccessful" equivalence test (a CT undercount). For difference, if the assigned α was .01 and the study produced a $p = .03$, the nil hypothesis would not be rejected (a difference test success) but would be rejected if the study had been assigned $\alpha = .05$ (an unsuccessful difference test), leading to a difference test success (CT overcount).

Unlike Steiner and Wong's simulations, statistical power was not systematically controlled in the current case, though features of power such as sample size were likely equivalent in subsets of study pairs due to the random allocation aspect of SMART/CORR. Thus, cells created by the SMART/CORR design likely had similar though small sample sizes, which would increase the indeterminacy rate. Despite differences between the current study and Steiner and Wong, threshold choice was consistently important in establishing a successful CT.

Even with a large, initial dataset of 87 study pairs, SMART/CORR-based tests of statistical significance often utilized small samples of study pairs, leading to reduced power. Successful difference tests in Stage 1 produced

subsets of seven and nine study pairs. When followed by a second randomization, the subset sample size was essentially halved. Thus, Stage 2 comparisons between different thresholds were sometimes based on very few study pairs (e.g., C1 vs. C2: four vs. three study pairs). Increasing the number of replication studies in studies such as OSC or increasing the proportion of successful replications would mitigate this limitation.

Finally, it is important to note the fundamental role of chance within the SMART/CORR approach. The current study was based on a single series of randomizations. A second series of randomizations would lead to different subsets of study pairs with counts of successful CTs different than counts found in this study. However, this limitation is tempered by the similarity of the two correspondence tables in yielding similar rates for each of the four CT categories.

Conclusions

Previously, CT had been shown to be a viable method to assess replication success within a single study and its replicate. Now, it is reasonable to conclude that CT can be extended to include relatively large samples of replication studies. In addition, the current research corroborates replicability cautions noted by Steiner and Wong aimed to protect against indeterminacy by implementing the CT. In a SMART/CORR, the generality of results in unconfounded subsets of replication studies can be probed under ideal circumstances, in contrast to cases in which effects of studies from different publications or from artificial-simulation results were compared. Fortunately, both Steiner and Wong's simulation-based approach and the current study's approach utilizing actual study results lend credence to the important role of CT in defending against incorrect replicability conclusions.

In future studies, the SMART/CORR method would allow high-quality testing of multiple study conditions other than α-values and equivalence thresholds. For example, one could randomize studies to different statistical analyses or sample weighting strategies (though this approach requires access to original data) to evaluate their impact. While SMART designs have typically implemented simple random sampling, stratified random allocation might be applied by partitioning original studies into strata of different sample sizes.

The study results reported here do not alleviate growing concerns regarding the replicability crisis. They do affirm that, even under the best of circumstances, when methods in replication studies have been carefully crafted to match those of original studies, researchers cannot anticipate close effect size comparability. Instead, one can argue that replicability expectations should be reexamined. In both standard and SMART-based assessments, using CT to establish comparability, relatively large thresholds

were necessary to yield equivalent original-replicate pairs. Under such circumstances, one is led to reconsider the magnitude of thresholds used to create reasonable demonstrations of equivalence. At the least, researchers are encouraged to implement sensitivity tests that include multiple thresholds.

The SMART/CORR approach is not without precedent. Such analogs enable the further application of design thinking to the study of replicability. In two-stage randomization (e.g., Shadish, et. al., 2011), a WSC method, participants were first randomized to a particular design (an RCT and a quasi-experiment), and half are either randomized again or assigned to treatment by some other mechanism (e.g., self-selection or cut score). The closeness of estimates in two-stage randomization is quite analogous to the closeness of estimates within a SMART/CORR design. However, the extension of appropriate meta-analytic methods to original-replicate pairs in a SMART context as a gauge of replicability is novel. Conceptual connections of this sort are precisely the kind that sparks the addition of tools to our current kit of replication methods.

Author notes: Correspondence should be addressed to William H. Yeaton, Department of Educational Psychology and Learning Systems, College of Education, Florida State University, Stone Building, 1114 West Call Street, Tallahassee, FL 32306. Email: bill.yeaton@yahoo.com
Thanks to Betsy Becker, Warren Tryon, Bernd Weiss, and members of the Synthesis Research Group (SynRG) at FSU for helpful comments on earlier drafts of this paper.

References

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

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*, 305–324. DOI: 10.1080/00273171.2017.1289361.

Bonett, D. G. (2020). Design and analysis of replication studies. *Organizational Research Methods, 24*, 513–529. DOI: 10.1177/1094428120911088.

Borenstein, M. (2009). Effect sizes for continuous data. In J. C. Valentine, L. V. Hedges, & H. M. Cooper (Eds), *The Handbook of Research Synthesis and Meta-Analysis*. UPCC Book Collections on Project MUSE. New York: Russell Sage Foundation, pp. 221–235.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho. T.-H., Huber, J. Johannesson, M., Kirchler, M., Nave, G. Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer. L., Imai, T., Isaksson, S., . . . Wu, H. (2018). Evaluating the replicability of social science

experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour, 2*, 637–644. DOI: 10.1038/s41562-018-0399-z.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & Morris, L. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management, 37*, 403–429. DOI: 10.1002/pam.22051.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE*, D. Marinazzo (Ed.) *11*, e0149794. DOI: 10.1371/journal.pone.0149794.

Fabrigar, L. R., & Wegener D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66*, 68–80. DOI: 10.1016/j.jesp.2015.07.009.

Gelman, A., & Vazire, S. (2021). Why did it take so many decades for the behavioral sciences to develop a sense of the crisis around methodology and replication? *Journal of Methods and Measurement in the Social Sciences*, *12*, 27-31.

Institute of Education Sciences. (2020). What works clearinghouse standards handbook, version 4.1. Available at: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693–2710. DOI: 10.1002/sim.1482.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*, 270–280. DOI: 10.1177/2515245918771304.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies: Development of adaptive treatment strategies. *Statistics in Medicine*, *24,* 1455–1481. DOI: 10.1002/sim.2022.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, 1–8. DOI: 10.1126/science.aac4716.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology,* 10.813. DOI: 10.3389/fpsyg.2019.00813.

Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin, 146*, 701–719. DOI: 10.1037/bul0000232.

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

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological Methods, 16,* 179–191. DOI: 10.1037/a0023345.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. *Evaluation Review, 42*, 214–247. DOI: 10.1177/0193841X18773807.

Steiner, P. M., Wong, V. C., & Anglin K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie, 227*, 280–292. DOI: 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*, 371–386. DOI: 10.1037/1082-989X.6.4.371.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112–125. DOI: 10.1002/jrsm.11.

Williams, C. R. (2019). How redefining statistical significance can worsen the replication crisis. *Economics Letters, 181*, 65–69. DOI: 10.1016/j.econlet.2019.05.007.