Estimating the effect of publication bias in Behavioural research

Introduction

This paper uses the results of over 300 replication studies conducted as a part of eight large scale replication projects (henceforth ‘replication projects’) to estimate the effect of publication bias in the Behavioural Sciences research literature. Although the presence or absence of effects may be an interesting question in of itself, an understanding of the magnitude of effect sizes is essential to understanding a relationship or system. As such the discovery and precise estimation of associations and effects is essential to developing a coherent and reliable scientific literature is. A major effort among psychological researchers and methodologists in recent decades in behavioural research has been the movement away from focusing entirely on binary outcome statistical significance testing (e.g., Cohen, 1990; Cohen, 1994; Meehl, 1967, 1978).

Under conditions where results are selectively reported based on characteristics related to the size of the effect (e.g., statistical significance), the literature no longer provides an unbiased estimate of the true outcome effect (Hedges, 1992). There is good reason to think that reporting and publication biases lead to exaggeration of effect sizes in the behavioural sciences literature results (Lane & Dunlap, 1978; Mahoney, 1977; Murphy & Aguinis, 2017; Simmons, Nelson, & Simonsohn, 2011). The current paper examines a newly available resource, large scale replication studies which have systematically replicated bodies of research, in order to estimate the degree to which effects reported in the psychological literature are inflated.

All of these projects were primarily conducted in order to assess the degree to which their particular area of research contains results which are irreproducible, or to estimate variability in effects among subpopulations. All used non-random samples of the literature, and all show that the reproducibility of results is below what would be expected given that all experiments were being analysed and published without regard to the statistical significance of results. See Table 1 for a list of the included replication projects, along with their target populations, and the percentages of replication attempts with a statistically significant results in the same direction as the replicated result. This new body of literature makes it possible to assess the effect of publication bias on the size of the reported effects in addition to estimating the proportion of various bodies of research which successfully replicate.

Table 1.

Included large scale replication projects, along with the number of articles replicated, the number included in this analyses of each type, and how sample sizes were determined.

Reference	Number of included replication studies	Percent of replications statistically significant in the same direction as the original result
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi:10.1126/science.aac4716	100	36%
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating Variation in Replicability. *Social Psychology*, 45(3), 142-152. doi:10.1027/1864-9335/a000178	13	85%
Klein et al. (2018)	28	54%
Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82. doi:10.1016/j.jesp.2015.10.012	10	30%
Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644. doi:10.1038/s41562-018-0399-z	21	62%
Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433. DOI: 10.1126/science.aaf0918	18	61%
Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Colombo, M. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 1-36. doi: 10.1007/s13164-018-0407-2.	40	70%
Soto et al (2018)[a]	121	86%

Note: [a] Soto et al's (2018)'s replication rate and was recalculated on the "study" (i.e., using the number of replicated effects not the number of trait-outcome associations as is reported in the paper) using results disattenuated using the Spearman-Brown prediction formula and Spearman disattenuation formula (Lord & Novick, 1968) to account for less reliable shorter form measures used in the replication studies.

Publication and reporting bias's effect on reported effect sizes

Publication bias is the process by which studies that find results which support their hypotheses, usually by showing statistically significant results, are more likely to be published than those that do not (Lane & Dunlap, 1978; Mahoney, 1977). This is the traditional "file draw effect" (Rosenthal, 1979), the idea that non-significant results get placed in the file draw as opposed to being reported. If studies are more likely to be published given that they show statistically significant results, effect sizes in the literature will be, on average, exaggerated, and the number of false positives (i.e., true null effects showing statistically significant results) increased (Lane & Dunlap, 1978). This occurs because the smaller the sample size included in research, all else being equal, the larger the observed effect has to be to reach statistical significance (Oakes, 1986). When an effect under study is truly null, or practically indistinguishable from the null, and null effects are rarely published, this can create the appearance of true non-zero effect in the literature based on false positive results alone (Oakes, 1986). The degree of effect size exaggeration depends, primarily, on the true statistical power of studies (i.e., the effect size and sample size included in studies given the experimental design and analysis strategy) and the proportion of true nulls being investigated (Oakes, 1986). If anything, publication bias towards statistically significant results appears to be particularly acute in behavioural research, where the number of papers reporting at least one statistically significant result (estimates range from 75% to over 90%; Fanelli, 2010; Fanelli, 2012; Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016). Taking a recent estimate of the average power of psychology to detect reasonable estimates of the average effect sizes seen in psychology (44% to detect a cohen's d of .5; Szucs & Ioannidis, 2017), the proportion of significant reported results would mean that 1.7 to 2.05 studies would have to be conducted per published paper to account for the proportion of studies that report significant findings under favourable assumptions[1].

Selective reporting among measures and Questionable Research Practices (QRPs) can also lead to the same outcome, the appearance of support for a particular theory or the presence of a particular effect, when particular outcome measures are reported, emphasised, or not reported because of the results of statistical analyses. QRPs like p-hacking and Hypothesising After the Results are Known (HARKing) on the basis of the some outcome measure such as statistical significance or achieving large effect sizes (Kerr, 1998) can also lead to effect sizes being exaggerated and increased proportions of false positives in the scientific literature (Bakker, van Dijk, & Wicherts, 2012; Murphy & Aguinis, 2017; Simmons et al., 2011). Recent surveys in the behavioural research literature also

[1] $\frac{.75}{.44}$ to $\frac{.9}{.44}$. These estimates make the simplifying assumption that all studies performed examine true non-zero effects of a 'medium' effect size of .5 Cohen's *d*, and that all papers report a single. More realistic estimates of the proportion of null-experiments would lead to greater numbers of experiments having to be performed.

suggest that that questionable research practices activities like HARKing and p-hacking are common across fields of psychological research (Fiedler & Schwarz, 2015; John, Loewenstein, & Prelec, 2012). All of these activities lead to increased false positive errors, and equivalently exaggerated effect sizes as represented in the scientific literature. The current paper provides an estimate of the cumulative effect of publication bias and QRPs on published effect sizes, an essential piece of information in accurately reading and interpreting the scientific literature.

Previous efforts to estimate publication bias in the behavioural sciences literature

Previous efforts to assess this question have shown using simulation studies that under reasonable assumptions type one error rates could be as high as 40% and effect size exaggeration as high as d = .33 when questionable research practices are in place even, and even without QRPs as high as d = .16, in typical experiments in psychology (Bakker et al., 2012). Stanley, Carter, and Doucouliagos (2018) used WAAP-WLS, ELS and PET-PEESE estimators in 200 meta-analyses published in Psychological Bulletin and suggest that differences in effect sizes between replication and original studies can largely be explained by heterogeneity not selective reporting in either direction. They found 8 to 15% residual effect size bias depending on the meta-analytic bias reduction method used. However, two of three of these estimation methods are known to be downwardly biased (leading to underestimates of the amount effect size inflation (Stanley & Doucouliagos, 2015; Stanley, Doucouliagos, & Ioannidis, 2017)), and this literature could reasonably be expected to be less biased than others in that they only sampled studies for which enough papers had been published to perform a meta-analysis, and to be covered in the pages of Psychological Bulletin.

In order to estimate the effect of publication bias on effect sizes, the current study presents an exploratory analysis of this large set of data using three main approaches. The first approach is purely descriptive, reporting the raw change in effect sizes. The second approach is to use a multilevel or hierarchical meta-analytic approach to provide an estimate of the expected effect size change between original replication studies. The third approach is to use a Bayesian Mixture Model to quantify the degree of effect size change from original to replication effect sizes. As the replication studies inevitably include a large number of effects which are likely to be true null effects (or effects which are so close to true null effects to be practically dismissible), this study also presents a series of analyses designed to estimate the degree to which true effect sizes in the literature are decreased after excluding those effects which are likely to be true nulls.

Methods

Data extraction

All eight published or in press large scale replication projects performed within in the behavioral science research literature were included in the current research (see table 1 for a list of the included studies and their sample size determination methods). The original source of each replicated effect, reported test statistics, effect sizes, sample sizes, standard errors and p-values were extracted for each original and replication study. Several of the large scale replication projects did not present the original test statistics and p values (e.g., Many labs 1 and 3). In these cases, these values were manually extracted from the original articles. When sample sizes for original studies were not available they were manually extracted from original articles where possible. When the original and replication effect sizes were not reported as Fisher Z transformed correlation coefficients, effect sizes were converted from test statistics or effect sizes for analysis if the original results or those reported in the replication project were reported in correlation coefficients or Cohen's d, but not otherwise. This means that the current analysis follows the assumptions for conversion from a particular statistical output to correlation or Cohen's d following the replication projects, and that some analyses have been left out of the current analysis (e.g., Chi Square tests from (Open Science Collaboration, 2015)). In cases where sample sizes were not reported per group equal sample sizes among groups were assumed in these conversions. See Table one for the number of valid studies extracted from each project. All results are transformed from Fisher Z transformed correlation coefficients (which are used in all analyses) to correlation coefficients unless it is otherwise stated. This was done in part following (Open Science Collaboration, 2015), and in part to present results in a common and intuitively understandable format which should be familiar to most psychologists and behavioral researchers.

Extraction details

Three effects which did not report that their findings were indicative of a non-zero effect were excluded from (Open Science Collaboration, 2015). In some cases in the Nature Science reproducibility projects (Camerer et al., 2018) multiple replication studies were performed for a single effect. In each of these cases we performed a fixed effects meta-analysis using the metafor package (Viechtbauer, 2010) to estimate a meta-analytic effect size estimate. The effect size, standard errors and sample sizes used in the current study reflect this pooled estimate. This method leads to one study more “replicating” according to the ‘statistical significance in the same direction of the original study’ than was originally reported in (Camerer et al., 2018), where they using the largest performed study instead of a pooled estimate.

In [LOOPR study CITATION], some measures used shorter form version of the original questionnaire, all results presented have been disattenuated using the Spearman-Brown prediction formula and Spearman disattenuation formula to estimate the trait-outcome associations that would be expected if our outcome measure had used the same number of items as the original study (Lord & Novick, 1968). Following the other large scale replication studies, the signs of negative original correlations were set to positive (and the sign of the replication sample were switched too). The experimental philosophy reproducibility project included two original studies which were non-significant (and which were not claimed to provide evidence for the effects under test), these were removed from analysis. Four studies from Many labs 2 [CITATION] were removed because effect sizes could not be simply derived (the original and replication studies examined a difference in effect sizes seen in different conditions, and the effects were not directly tested against each other), and two additional were excluded because their effect sizes were only available in Cohen’s q.

Analysis

All analysis was performed in R version 3.5 (R Development Core Team, 2018). Mean raw differences along with Wald-type 95% confidence intervals around the mean difference, median effect size differences, and raw proportion decreases in effect sizes (i.e., $(originalES_i - replicaitonES_i)\ /\ originalES_i$) were calculated on the Fisher-Z transformed effect sizes. The reported Wald-type confidence intervals do not account for non-independence between effects taken from the same paper, or between studies from the same replication projects. In order to account for this non-independence, a multilevel-meta-analysis framework was used, see below for more details. Confidence intervals around binomial proportions are 95% Wilson Score intervals. All difference scores (i.e., proportion changes and mean differences) were calculated using Fisher Z transformed effect sizes. Any studies with missing data (e.g., missing effect sizes or sample sizes for the initial or replication studies) were excluded, and sample sizes are reported alongside each analysis in tables. All analyses were exploratory, and multiple models which were developed are not presented here. See https://github.com/fsingletonthorn/effectSizeAdjustment for a git repository with a record of all interim models and for all model code and data, and see https://osf.io/daj8b for the preregistration of this project.

Multilevel meta-analysis

In order to obtain a reasonable estimate of the change in effect size between original and replication studies, a multilevel random effects meta-analysis was performed on the difference in Fisher Z transformed correlations between original and replication studies. This treats each pair of effects, the original and replicated sample size, as one “study” in a meta-analytic framework. Standard errors for each effect were estimated as $se = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}$ with N_1 being the sample size in the original study and N_2 being the sample size in the replication study. Empirical Bayes estimates and 95% credible intervals for random effects were calculated following (Morris, 1983; Robinson, 1991). In order to account for non-independence between effects, this meta-analysis included random effects for each effect and for the replication project each replication attempt was performed as a part of. Meta-analyses were performed using the Metafor package (Viechtbauer, 2010).

Leave one out cross validation

To assess whether the main results of this study are sensitive to the inclusion of each of the replication projects and individual replicated effects within each replication project, the models were rerun using leave one out cross validation, excluding both the individual replication attempts and the replication projects one at a time. When leaving out individual studies the range of point estimates (i.e., the difference between the smallest and largest estimate of the difference between original and replication studies) for each of the LOO cross validation models did not exceed more than a Fisher z sore of 0.02. When excluding one replication project at a time, model estimate ranges did not exceed 0.05. See supplementary material [LOO] for a table of the proportion of model estimate p values below .05, and estimate quintiles for each model from the leave on out cross validation on the study and project levels. None of these changes would lead to substantially different conclusions being drawn from the model output.

Accounting for null effects

An important question in assessing the degree to which effects are attenuated in this literature is how much this effect is driven by the presence of a subset of null effects (or effects so small as to be effectively null). The average attenuation could be extremely high, and yet this effect be almost entirely driven by the presence of effectively-null effects. This aspect becomes especially important as the sampling of the literature is non-random, meaning it is plausible that some effects were chosen for replication to a greater or lesser extent as it was expected that they may not replicate, inflating or deflating the amount of effect size change seen in the entire sample. In order to account for this issue, three main approaches were taken.

Firstly, original studies were simply excluded using various exclusion criteria, and raw effect size differences calculated and multilevel meta-analysis models re-estimated. The exclusion criteria used are detailed below. Because all of these methods function by removing small effects, no significance testing was performed on the difference between the model estimates estimated decreases after accounting for small or near-null effects. It is certain that, at a population level, all of these actions would lower the size of the observed effect size decrease. The second method of estimating the effect size difference while accounting for the presence of null or effectively null effects was to include the p value of the replication studies as a moderator in the meta-analysis of the effect size differences detailed above. This means that the model estimate, the meta-analytic mean, is the predicted mean effect size decrease assuming a replication p value of 0. The third method of estimating the effect size difference in non-null effects was the a Bayesian mixture model adapted from (Camerer et al., 2018), more detail is provided below.

Exclusion rules

Multiple exclusion rules were used; excluding studies in which the replication study was not significant, removing statistically equivalent studies found using equivalence testing, and using a cut score from approximate Bayes Factors estimated from the reported correlation coefficient effect sizes.

Statistical significance of the replication study

The first method used to exclude likely effectively null effects is to only look at effects that reached statistical significance in the replication study and which had replication effects in the same direction as the original effect. This has the issue of meaning that effects for which the replication studies which were under-powered to detect the true effect size are likely to be excluded. Especially as in some of the replication projects the sample size in the second study was chosen using a power analysis of the observed effect in the original study (Open Science Collaboration, 2015), this method is likely to underestimate the amount of effect size exaggeration due to the exclusion of under-powered replications. In these cases, original studies which found large effects lead to follow up studies which have low power to detect the true, non-zero, but smaller effect size.

Equivalence tests

In order to avoid excluding under-powered studies erroneously, we also excluded studies based on whether the results of the replication study were statistically equivalent to the null (Lakens, 2017; Lakens, Scheel, & Isager, 2018), or significant in the opposite direction. As a requirement for equivalence testing is that a minimum effect size of interest is selected, we follow one suggestion in (Lakens et al., 2018) and use the lowest effect size that would be statistically significant to the original study as the smallest effect of interest (assuming an alpha of .05). Equivalence tests were performed used the Fisher Z transformed effect sizes, and approximated the standard errors of each study as $\sqrt{\frac{1}{n-3}}$, except for studies from (Camerer et al., 2018) which had more than a single replication attempts, where standard errors are those derived from the meta-analyses that produced the effect size estimate. Equivalence tests were performed using z tests, i.e., assuming a normal sampling distribution. Ideally, a full reanalysis would be performed for each original study using the original statistical test and full access to the original and replication data. However, it was not feasible to extract and reperform full analyses for the over 600 total original and replication studies. As a method of testing how closely this method of approximating standard errors matches the original replication projects results, significance tests for the replication and original studies were performed using this approximation. The results matched the significance or non-significance as reported in the replication projects in every case.

In interpreting the results of this analysis it is important to note that the minimum detectable effect was occasionally quite high as original sample sizes were often very small (mean = 0.17, SD = 0.12, 0th, 25th, 50th, 75th and 100th quintiles = [0, 0.1, 0.15, 0.23, 0.74]). This means that original studies were sometimes under-powered to detect even large effects using the current analysis (i.e., converting effects to correlation coefficients and estimating standard errors in this way). This means that this method may exclude studies which have effects the original authors may have considered important, but which they would have been unlikely to detect using this simplified analytic approach.

Approximate Bayes factors

Three different types of Bayes factors were developed for each study using default priors following (Wagenmakers, Verhagen, & Ly, 2016). Bayes Factors express the relative evidence for the null hypothesis compared to an alternative model, or equivalently the degree to which a Bayesian observer should update their prior beliefs in response to the receipt of new data in favor of one model or another. If a Bayes factor is greater than one the data is more likely under the alternative hypothesis than under the null hypothesis, and the opposite is true when a Bayes factor is below one. Conventional labels have been proposed, suggesting that Bayes factors between 1 and 3 provide little to no evidence (or 'anecdotal' evidence) and Bayes factors from 3-10 provide "substantial" evidence (Jeffreys, 1961; Wagenmakers et al., 2016).

Two of the developed Bayes Factors ignore the original study and express the relative evidence for and against the point null entirely based on results of the replication study, using a one (BF_{0+}) and and two tailed (BF_{01}) default alternative hypothesis (for details see (Wagenmakers et al., 2016)). Replication Bayes Factors (BF_{rep1}) were also developed, in which the prior for the replication correlation coefficient is the posterior based on the original research and a flat prior, for details see (Wagenmakers et al., 2016) and (Verhagen & Wagenmakers, 2014). This paper follows the typical notation where the order of the subscripts indicate whether a Bayes Factor represent evidence for the null (BF_{0+}, BF_{01}, BF_{0rep}) or for the alternative hypothesis (BF_{+0}, BF_{10}, BF_{rep0}).

The Bayes factors presented here were developed using the effect sizes as reported in correlation coefficients, regardless of the original study's experimental design. Importantly, these Bayes factors differ from those that would normally be developed using the closest Bayesian equivalents to each original replicated study's analysis, and should be viewed only as a coarse estimate of the degree of evidence provided for and against the null model. See table [bayesFactors] in supplementary materials [Bayes] for a table showing the differences between the values returned by this method compared to those reported in the Bayesian supplement to which were more appropriately calculated (Camerer et al., 2018), which demonstrates that the difference can be considerable, especially when the original analysis was idiosyncratic. Normally, One of the benefits of Bayes Factors is their continuous and interpretable scale, however in this case these approximate Bayes Factors are used as a heuristic to discard the studies which appear to likely be true (or effectively) null effects. Two different cut scores were used for each type of Bayes factor, discarding studies when Bayes factors suggested that the null model is either more than three times more likely than the alternative model (i.e., when there is more than 'anecdotal' evidence that the null is true), or when the alternative model is not at least three times more likely than the null model. A full Bayesian treatment of this issue is presented in the Bayesian Mixture Model below.

Simulations to assess exclusion criteria

All methods of excluding studies function by removing studies which have small effect sizes in the replication, so it was a forgone conclusion that the apparent amount of effect size reduction seen will go down as compared to the model which includes all effects. Because of the exploratory nature of the methods used to attempt to remove studies from this literature, a series of simulation studies were performed to assess how accurately the exclusion methods classify studies, and how accurately the raw estimates of effect size attenuation are under reasonable assumptions. Simulations took the original effect sizes, estimated a 'true' effect size from a normal distribution with a mean of the original effect a standard deviation equal to the standard error of the original study, and reduced this effect by an attenuation factor of 0 - 1 in steps of 0.1, and set a random proportion of ‘true’ effect sizes to 0 (again a proportion from 0 to 1 in steps of 0.1). Simulations were performed at least 10000 times for each analysis.

Accuracy (i.e., the proportion of studies which were accurately excluded as true negative or null effects, or for equivalence testing the proportion of studies which were at or below the minimum effect size of interest and which were statistically equivalent to the null) was assessed under this data generation process in 11958 simulations, showing that accuracy of these methods across all scenarios ranged from 0.75 to 0.84, with SDs of 0.07 to 0.19. See supplementary materials [simulations] table [SM accuracy] for full details on the performed simulations, including a table of the outcomes of these simulations, and heat maps of the mean error over these values. Note that these values are only valid under the simulated specific data generation process, where there is a consistent factor decrease in true effect size, and where the studies which are null are random and independent of the original effect and sample sizes. See supplementary materials [simulation] for a full description of the simulations, heat maps of the mean absolute error at each benchmark and full simulation output tables. The code used in these simulations is available from [OSFOSF.io].

Bayesian mixture model

The final approach to estimating the amount of effect size attenuation expected given that the effect under question is non-zero was the Bayesian mixture model presented in (Camerer et al., 2018). This model assumes that the each observed replication effect size comes from one of two components, either from the null hypothesis or from the alternative hypothesis. If the replication effect size is drawn from the null hypothesis, it is assumed to have come from a normal distribution with a mean of the true effect size (a value sampled from a distribution with a mean of 0 and a modeled standard deviation) and a standard deviation equal to the standard error of the replication study (estimated here as $\sqrt{\frac{1}{n-3}}$, n being the replication sample size). If the replication effect size comes from the alternative hypothesis, it is assumed to have been drawn from a normal distribution with a standard deviation equal to the standard error of the replication study, and a mean equal to the true effect size. In this case, the true effect size is sampled from a normal distribution with a mean equal to the original study's estimated true effect size attenuated by an "attenuation factor", equal to some value between zero and one and assumed to be equal across all studies. There are two main parameters of interest in this model; the "attenuation factor" (called a deflation factor in (Camerer et al., 2018)), the degree to which effect sizes are attenuated between original and replication study, and the overall rate at which studies are assigned to have come from the null hypothesis (the "assignment rate").

This model was estimated using four Markov chains from each of which 100,000 draws were taken (excluding a 11,000 draw burn-in period). Trace and density plots for the discussed parameters were assessed and the model appeared to have successfully converged. This analysis was performed in JAGS version 4.3.0 (Depaoli, Clifton, & Cobb, 2016) using the rjags interface (version 4.8.0; (Plummer, Stukalov, & Denwood, 2018)). See supplementary materials ["mixture model"] for model syntax and further analysis details.

Results

Raw decreases in effect sizes

Looking at the 314 replications for which both original and replication effect sizes were available, the effect size seen in the replication study was lower than that seen in the original study in 227 articles, 72% of the included articles. The average effect size for original studies was 0.38, and the mean effect size for replication studies was 0.27. There was an average decrease of r = -0.13 (Wald-type 95% CI [-0.16, -0.1], the slight discrepancy between values occurs because Fisher Z scores have been back transformed to correlation coefficients for presentation here). Notably, this represents an average decrease in effect sizes from the original to the replication study of -29.44%. See Table 2 for a comprehensive list of descriptives on the effect size differences seen, Figure 1 for a scatterplot of the replication effect sizes plotted against the original studies' and Figure 2 for a raincloud plot of the Fisher Z score change in effect sizes by replication project.

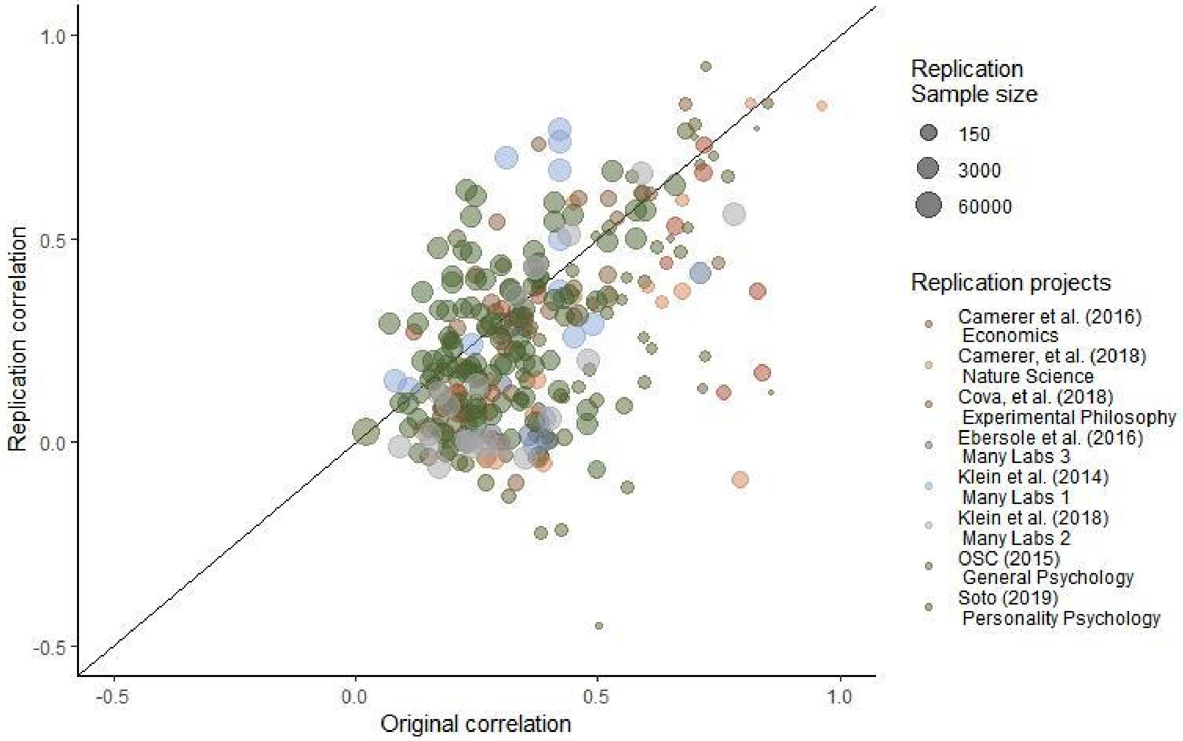


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of participants in the replication study, and the color of the points shows which replication project each effect size pair was from.

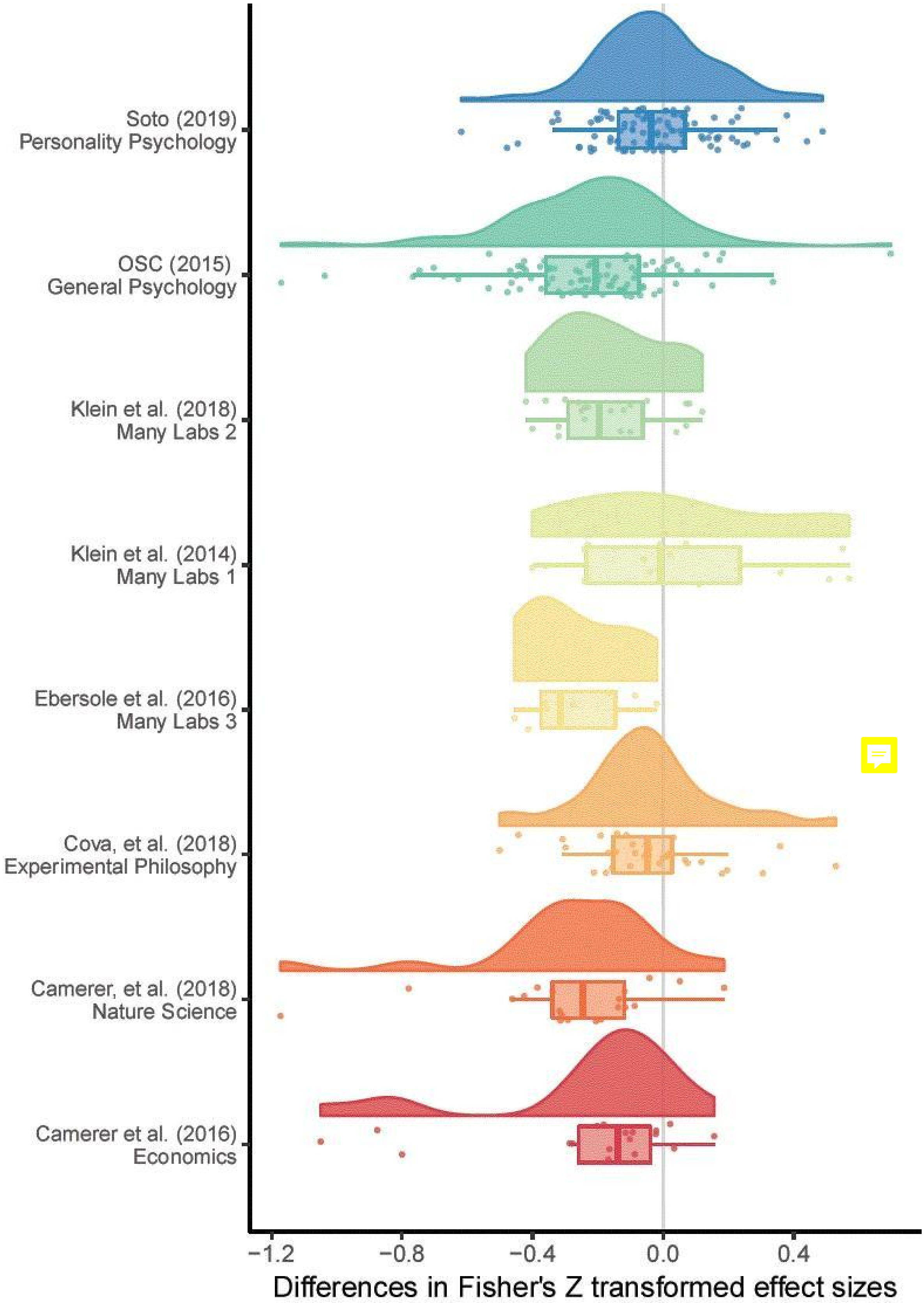


Figure 2. A raincloud plot of the change in effect sizes (here Fisher Z scores) from the original to the replication study by the replication project that each replication study was performed as a part of.

Excluding null results

Looking at the 201 replications in which the replication study was statistically significant 64% of all studies, the average effect for original studies was 0.4, and the mean effect size for replication studies was 0.38. There was an average decrease of r = -0.02, Wald-type 95% CI [-0.05, 0], an average decrease of 3.49%.

Excluding studies which were not statistically significant is likely to lead to an underestimate of the degree of effect size attenuation, as this exclusion rule will lead to the exclusion of under-powered replication studies as well as studies which are likely to be true null effects. In order to avoid this issue, equivalence tests were performed, meaning that the studies which are not-statistically equivalent to the null are included (using a bound of equivalence equal to the minimum detectable effect in the original study). This method is an attempt to not exclude the non-diagnostic replication studies, studies which are not statistically significant but which do not suggest that that the result is statistically equivalent to the null. Using these method 237 replications were not statistically equivalent to the null, 77.7% of studies for which equivalence tests could be performed. The average effect size in the original non-equivalent studies was 0.41, compared to a mean effect size for replication studies of r = 0.35. This is a mean decrease of r = -0.07, Wald-type 95% CI [-0.1, -0.04], an average decrease of -6.65%.

The results of the various Bayes Factors analyses generally support the results of the analysis removing statistically equivalent studies. Using this method, 177 to 232 replications were included, 58.22 to 76.32% of studies for which Bayes Factors tests could be estimated. See table 2 for full output.

Table 2. Differences between original and replication studies. All calculations were performed on Fisher's Z transformed correlations and back-transformed into correlation coefficients for interpretability.

	n included	n criteria calculable for	Mean original ES	Median original ES	Mean replication ES	Median replicaition ES	Mean ES difference	95% CI LB Mean ES Change	95% CI UB Mean ES Change	Median ES difference	SD difference	Mean proportion change	Median proportion change
Overall	314	314	0.38	0.33	0.27	0.20	-0.13	-0.16	-0.10	-0.11	0.25	-0.29	-0.35
StatisticalSignificance	201	314	0.40	0.35	0.38	0.32	-0.02	-0.05	0.00	-0.04	0.20	0.03	-0.07
Nonequivalence	237	305	0.41	0.35	0.35	0.30	-0.07	-0.10	-0.04	-0.06	0.24	-0.07	-0.16
BF0RepBelow3	220	303	0.40	0.34	0.37	0.32	-0.04	-0.06	-0.01	-0.05	0.20	-0.01	-0.13
BFRep0Above3	186	303	0.41	0.35	0.40	0.36	-0.01	-0.04	0.02	-0.01	0.20	0.09	-0.05
BF01Below3	221	304	0.42	0.36	0.36	0.32	-0.06	-0.10	-0.03	-0.05	0.25	-0.04	-0.13
BF10Above3	177	304	0.41	0.35	0.40	0.35	-0.01	-0.05	0.02	-0.01	0.20	0.08	-0.04
BF0PBelow3	232	304	0.42	0.35	0.36	0.31	-0.07	-0.10	-0.03	-0.05	0.24	-0.04	-0.14
BFP0Above3	186	304	0.41	0.35	0.40	0.35	-0.01	-0.04	0.02	-0.01	0.21	0.08	-0.05

Multilevel model

The model including all data estimates a r = -0.14 (95% CI [-0.2, -0.07]) decrease in effect sizes from the original to replication studies. This is represents a change equivalent to -34.34% (95% CI [-50.79%, -17.88%]) of the mean effect size in the original studies (r = 0.36).

Looking at the amount of variance explained, there was more variance attributable to the article (i.e., the original article) than too the project ($\sigma^2_{article}$ = 0.02 compared to $\sigma^2_{project}$ = 0.01), representing an intraclass correlation (ICC) of 0.22, meaning there is only a low correlation between the amounts of effect size change seen within projects. QE tests of heterogeneity suggest that there is a large amount of unexplained heterogeneity, QE(304) = 3527.86, p < .001, unsurprisingly given the heterogeneous sample included in the current sample.

Table [nice mod sum]. Model output from a multilevel random effects meta-analysis of the difference between original and replication effect sizes, with random effects for the project (i.e., which large scale replication project the replication was a part of) and the original (i.e., replicated) article or effect.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.14	-0.21	-0.07	0.03	< .001	
					Project variance = 0.007, n = 8
					Article variance = 0.025, n = 228
					QE(304) = 3527.86, p < .001

Table [BLUP]. Empirical Bayes estimates and 95% credible intervals for random effects, which are equivalent to 95% confidence intervals assuming that the studies are a random sample from a population with normally distributed average effect size difference.

	Estimate	Standard Error	95% PI lower bound	95% PI upper bound
Camerer et al. (2016) Economics	-0.06	0.05	-0.15	0.04
Camerer, et al. (2018) Nature Science	-0.07	0.05	-0.17	0.03
Cova, et al. (2018) Experimental Philosophy	0.08	0.04	-0.01	0.17
Ebersole et al. (2016) Many Labs 3	-0.06	0.06	-0.17	0.05
Klein et al. (2014) Many Labs 1	0.11	0.05	0.01	0.21

Klein et al. (2018) Many Labs 2	-0.03	0.05	-0.12	0.06
OSC (2015) General Psychology	-0.05	0.04	-0.12	0.03
Soto (2019) Personality Psychology	0.07	0.04	-0.01	0.15

The model was re-estimated using each the subsets of studies, excluding studies based on the exclusion criteria detailed above. See table [all model output] for the model estimates from each model. The estimates of the proportion of variance attributable to the article or replication project level did not change considerably in any of these models. There is a notable reduction in the estimated effect sizes under these different selection criteria, with estimates of the amount of effect size decrease from r = -0.04 to -0.09, representing 8.95% to 22.31% of the average effect in the original studies.

It is important to emphasize the degree of uncertainty in these results. For example, taking the highest effect size decrease using any of the exclusion criteria, estimating the decrease using only the results of the 228 experiments which did not provide have a BF_{0P} of greater than 3 (i.e., which did not have at more than "anecdotal" evidence for the null hypothesis compared to the one sided alternative hypothesis), showed that an estimated decrease of summarize r = -0.09, 95% CI [-0.16, -0.02]. Looking at the smallest estimated effect size difference under any exclusion criteria on the other hand, running the multilevel meta-analysis on just the results of the 186 experiments which had a BF_{rep0} of three or more (i.e., which had a replication Bayes factor which showed more than 'anecdotal' evidence for the alternative hypothesis), showed an estimated decrease of -0.04, 95% CI [-0.09, 0.02]. This is equivalent to a decrease of -8.95% of the average original effect size, with a 95% confidence interval that extends from a noticeable decrease to a moderate increase; 95% CI [-23.47, 5.57]. See supplementary materials [all exclusion Cartier output] for full model output and scatter plots of the data-set using each exclusion rule.

Table [all model output]

The number of studies included in each model, and the estimated correlation coefficient decrease from each model. Models were estimated using Fisher Z transformed correlation coefficients and back transformed for interpretability. Percentage attenuation gives the percentage attenuation for effect size differences as a percentage for the the mean original effect size (r = 0.36).

Inclusion rule	Model N	Model Estimate	95% CI lb	95% CI ub	Estimated % attenuation	LB % attenuation	UB % attenuation
All studies	305	-0.14	-0.20	-0.07	-34.3	-51	-17.88
Statistically significant	195	-0.05	-0.12	0.01	-13.2	-29	2.67
Non-equivalent	235	-0.08	-0.15	-0.01	-20.5	-38	-2.58
BF0P < 3	228	-0.09	-0.16	-0.02	-22.3	-40	-4.50
BFP0 > 3	182	-0.05	-0.11	0.02	-12.0	-28	3.79
BF01 < 3	217	-0.09	-0.16	-0.01	-21.5	-39	-3.67
BF10 > 3	173	-0.05	-0.11	0.02	-12.0	-28	4.33
BF0Rep < 3	220	-0.06	-0.12	0.00	-14.5	-29	0.01
BFRep0 > 3	186	-0.04	-0.09	0.02	-8.9	-23	5.57

Including replication p values as a moderator

Given that the above methods all rely on throwing away data, a preferable method of modeling these results may be including the p value of the replication study as a moderator. A total of 305 studies provided enough data to be included in this model (i.e., studies for which replication p values, as well as sample sizes and effect sizes for replication and original studies, were extracted). This analysis leads to similar conclusions to the models with exclusions, with an estimated effect size decease of r = -0.08, 95% CI [-0.14, -0.02]. This represents a decrease of -20.17% (95% CI [-34.28%, -6.06%]) of the average original effect size. The projects and article level differences are functionally identical to the model that does not include replication p values as a moderator.

Table [moderators]. Multilevel meta-regression results including the p value of the replication study as a moderator.

	Estimate	95% CI LB	95% CI UB	SE	p	Random effects
Estimate	-0.082	-0.14	-0.024	0.029	0.005	
p value	-0.327	-0.40	-0.256	0.036	< .001	
						Project variance = 0.005, n = 8
						Article variance = 0.021, n = 228

QE(304) = 3143.78, p < .001

Mixture model results

A total of 305 effects were included in the mixture model (i.e., all studies for which sample sizes and effect sizes for replication and original studies were extracted). The overall posterior assignment rate (i.e., the proportion of studies which are estimated to be from the non-null alternative hypothesis) is 88%, with a 95% highest probability density interval of [79%, 97%]. The overall attenuation factor is 19.3% with a 95% highest probability density interval of [11%, 28%]. Figure [mixture model], shows the original effect sizes plotted against replication effect sizes weighted by sample size, along with the posterior assignment rate. The coloring in this plot indicates the proportion of times each effect was assigned to the alternative hypothesis. As was seen and pointed out in the first use of this model in (Camerer et al., 2018), values close to the diagonal are reliably assigned to the alternative hypothesis whereas effects far below the diagonal are more reliably assigned to the null hypothesis, although the overall posterior assignment rate might be overly optimistic. In part, this optimism may be due to the fact that this model allows for true effect sizes to be estimated as being extremely low or near zero due to sampling variability alone and still assigned to the alternative hypothesis not the null, with 29% of the estimated true replication effect sizes being smaller than an r of .1.

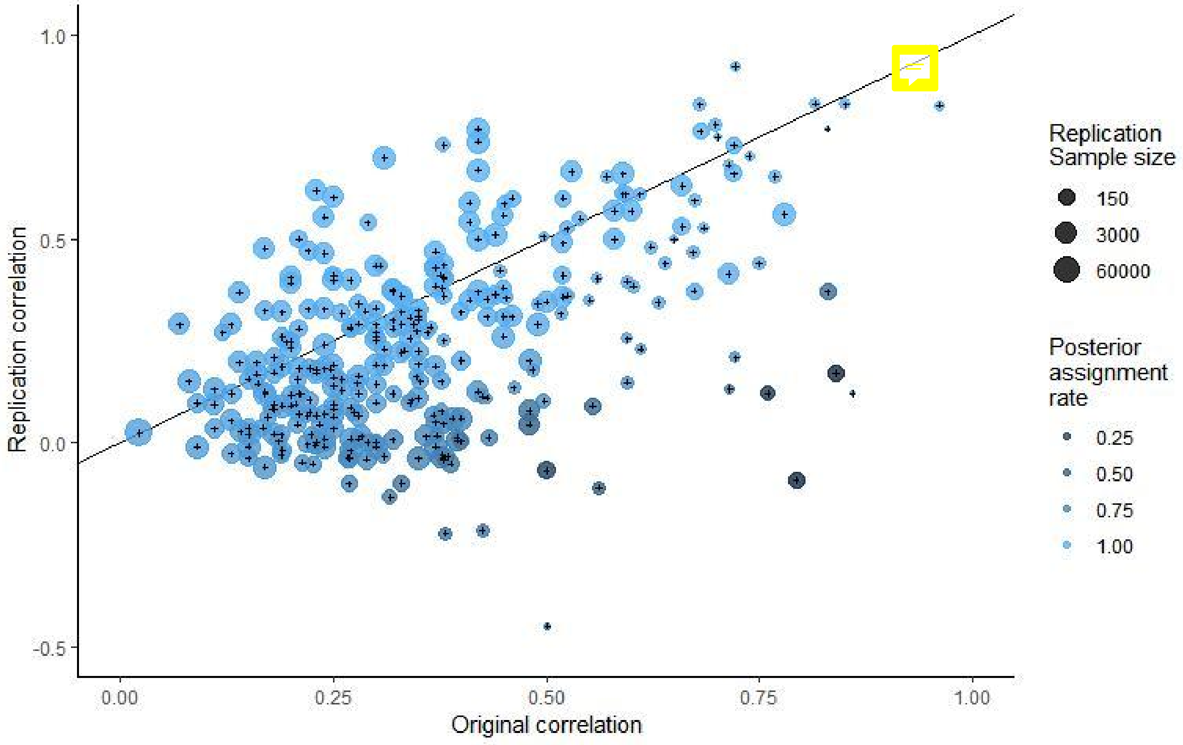


Figure [mixture model].

A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes, colored by the posterior assignment rate, the proportion of times each study was assigned to the alternative hypothesis. Points which fall on the the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of participants in the replication study.

Discussion

Overall, there was a substantial average decrease in effects sizes between the original and replication study. Taking the raw estimate, there was a moderate decrease of r = -0.13, approximately equal to a Cohen's d difference of -0.26. On average, replication sample sizes were -13% smaller in replication studies than they were in original studies. The results of the multilevel meta-analysis results agree with this estimate, showing an estimated mean decrease of r = -0.14, (95% CI [-0.2, -0.07]), equivalent to a -0.28 point Cohen's d decrease (95% CI [-0.41, -0.14]), or an estimated decrease of -34.34 (95% CI [-50.79%, -17.88%]) of the mean effect size in the original studies (a Fisher Z equal to r = 0.38). All of these estimates are higher than that suggested by (Stanley et al., 2018) of 8 - 15% reporting bias.

Arguably of more interest to researchers examining and planning research is the question of the degree of effect size attenuation expected under the assumption that the effect size is non-zero. All of the methods used here largely agree, although the degree of precision differs. The results of the two methods used to formally model this provide similar but more precise estimates, both estimating a decrease of approximately 20%, with a 95% highest probability density interval for the mixture model of a 95% highest probability density interval of [-28%, -11%] and a 95% confidence interval for the multilevel model including replication p Values as a moderator of [-34%, -6%]) of the average original effect size.

Although excluding data is a less than ideal way of investigating this issue, the results raw results and multilevel models performed excluding data also support these results. The raw mean decreases seen across the exclusion methods range from an decrease of r = -0.07 to -0.01 from the original to replication effect size. The multilevel models estimated excluding data lead to similar conclusions, although they highlight the degree of uncertainty in this result. Although all models estimated excluding data estimates show a lower effect size decrease when attempting to exclude null (or effectively null) results, the confidence intervals for all results extend from a decrease of -44.39% of the average original correlation coefficient, to an increase of 6.22%.

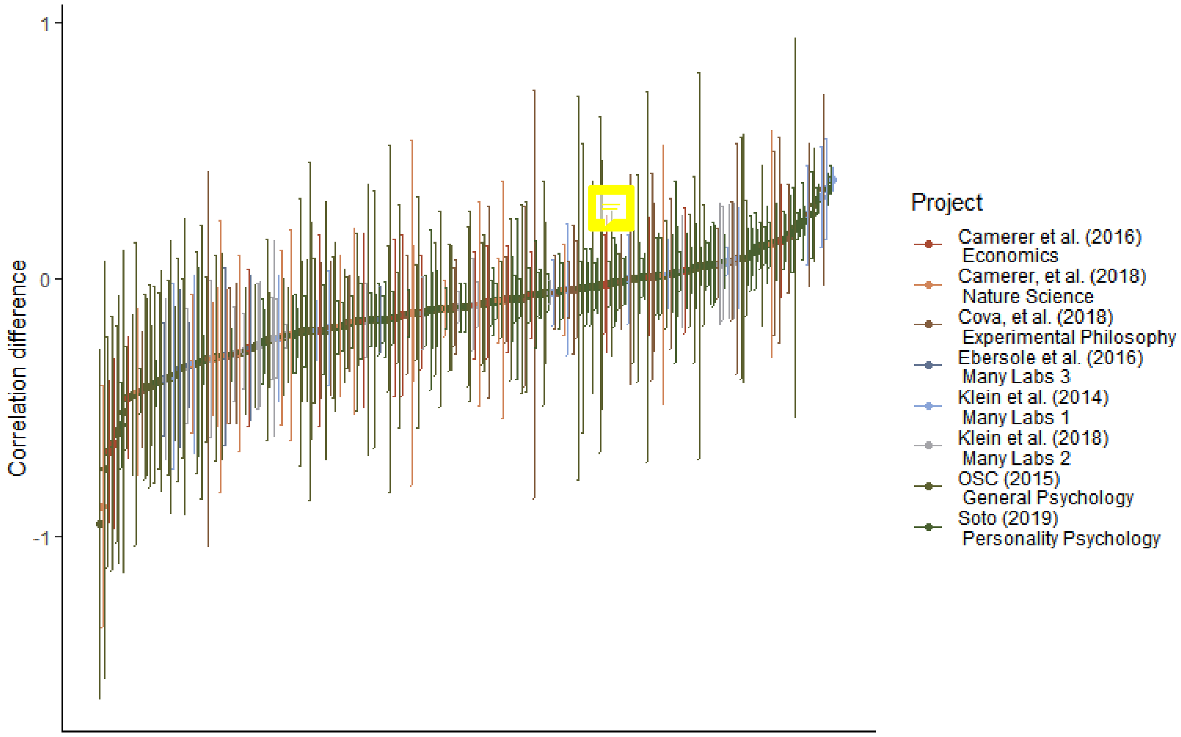


Figure x. A caterpillar plot of the effect size difference between original and replication study effect sizes ordered by magnitude, error bars are 95% confidence intervals around effect size differences.

Limitations and future directions

None of the projects included in this analysis were true random selections from the literature, and it is possible that the pattern in the selected sample may be different that that which would be seen in the literature overall. Additionally, although a number of different approaches were used to estimate the amount of effect size, all of the methods that were used to exclude replication studies from the multilevel modes which were bound to decrease the amount of effect size decrease that is seen. However, this preliminary analysis does provide suggestive evidence that the degree of effect size attenuation that is seen may be partially attributed to the presence of effectively-null results, and that the overall decrease in effect sizes in non-null studies is still considerable. The current study also does not attempt to distinguish between effect size heterogeneity (i.e., effect sizes that are different under different experimental conditions) and effect size attenuation. However, it seems reasonable to except in that effect size heterogeneity should lead to symmetrical effect size differences, and in so far as it might be expected to be negatively biased in replication studies, this could reasonably be termed effect size attenuation for the purposes of researchers hoping to replicate or plan future similar studies of the same same type of effects.

Furthermore, although the two alternative methods of estimating the amount of effect size reduction seen in non-null models (i.e., the model including replication p values as a moderator and the Bayesian mixture model) provide convergent evidence supporting the estimated effect size decrease assuming a non-null effect, both have issues. The model including p values as a moderator is best considered a heuristic guide, and is not straightforwardly interpretable (i.e., the model provides an estimate of the expected effect size difference given a replication p value of 0, an impossibility, and really acts as a convenience method for adjusting for the probability of the data being observed under the null model in the replication study). The Bayesian mixture model used here has two major issues. Firstly, it assumes independence between effect and that there is a uniform attenuation factor across all areas of psychological research. Secondly, the modeled true effect size can be negligible or even negative and the replication effect size still assumed to be sampled from the alternative distribution not the null. This later issue occurs because the model estimates the relationship between original and replication effect sizes using a two step errors-in-variables approach following (Matzke et al., 2017). Likely in large part due to this second issue, the model estimates the number of false positives at an extremely low rate (11.55% in contrast to other estimates of the number of false positive in this sample, which suggest a false positive rate of approximately 25%). For descriptive purposes, the current model does have a major benefit in that it estimates a single overall attenuation rate for non-null studies, the main goal of the current article. However, an important task in developing a more nuanced understanding of the data-generation process that leads to this data-set would be building a model that allows for the attenuation rate to change across replication studies, and possibly to allow it to include more components allowing for studies with negligible or negative but non-null effects in addition to the true alternative and null components modeled in the current studies.

Conclusion

These results highlight some major issues in the psychological research literature. Looking at the raw average proportion decrease, -29%, or at the results of the multilevel model, an r =-0.14 (95% CI [-0.2, -0.07]) average decrease differences which would make a considerable difference in most research scenarios. Researchers reading the literature should be aware of this large discrepancy, and plan their future experiments accordingly. Researchers who wish to ensure that they do not perform experiments that are unlikely to detect real effects should be aware that their experiments are likely to be under-powered if they plan their sample sizes using the effect size reported in a previous experiment. A researcher basing their intuitive estimates (or formal models) of the effect sizes that should be expected based on primary research findings are likely to be extremely optimistic. As a heuristic for researchers planning sample sizes, researchers could follow the advice given in (Camerer et al., 2018) and plan their experiments assuming that the original effect size is 50% of its reported value, a value matched by the more extreme 95% confidence interval of the estimated amount of effect size decrease using the multilevel meta-analytic framework. Alternatively, it may be preferable to rely on estimates of the smallest effect size of interest or use flexible analysis strategies that do not rely on precise a priori specification of the sample size to be collected (Albers & Lakens, 2018; Lakens, 2014; Schönbrodt & Wagenmakers, 2017) when practical constraints mean this approach is feasible.

Finally, this project emphasizes the importance of efforts to reduce publication and reporting biases. In many cases pre-registration with a sufficiently detailed analytic strategy may help to combat reporting bias by making it easier for researchers to avoid engaging in questionable research practices (Wicherts et al., 2016). In exploratory data analysis where an analytic strategy ends up being data-informed (e.g., like the current study), it is important to actively acknowledge this fact and to take other measures to reassure readers that the effects are not a product of a particular analytic strategy or due to other questionable research practices. One possible approach to this issue is to consider and report alternative analytic strategies without regard to whether their outcomes support a given theoretical position or empirical finding. Projects like registered reports, in which papers are reviewed before data-collection based on the design and analysis strategy as opposed to the results also show promise in developing a body of literature which is less effected by reporting and publication bias (Nosek & Lakens, 2014).

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187-195. doi:https://doi.org/10.1016/j.jesp.2017.09.004

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637-644. doi:10.1038/s41562-018-0399-z

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312. doi:10.1037/0003-066X.45.12.1304

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

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics, 41*(6), 628-649. doi:10.3102/1076998616664876

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE, 5*(4), e10068. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. doi:10.1007/s11192-011-0494-7

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. *Social Psychological and Personality Science, 7*(1), 45-52. doi:10.1177/1948550615612150

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: what is going on? *PeerJ, 4*, e1935. doi:10.7717/peerj.1935

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science, 7*(2), 246-255.

Jeffreys, H. (1961). Theory of probability, (3rd ed.). Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217. doi:10.1207/s15327957pspr0203_4

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

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances In Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31*(2), 107-112. doi:10.1111/j.2044-8317.1978.tb00578.x

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175. doi:10.1007/BF01173636

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology, 3*(1). doi:http://doi.org/10.1525/collabra.78

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), 103-115. doi:10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834. doi:10.1037/0022-006X.46.4.806

Miller, J. J. (1978). The Inverse of the Freeman – Tukey Double Arcsine Transformation. *The American Statistician, 32*(4), 138-138. doi:10.1080/00031305.1978.10479283

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association, 78*(381), 47-55. doi:10.1080/01621459.1983.10477920

Murphy, K. R., & Aguinis, H. (2017). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*. doi:10.1007/s10869-017-9524-7

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology, 45*(3), 137-141. doi:10.1027/1864-9335/a000192

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York, NY: Wiley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Retrieved from http://science.sciencemag.org/content/349/6251/aac4716.abstract

Peterson, R. (2018). bestNormalize. R package version 1.3.0.

Plummer, M., Stukalov, A., & Denwood, M. (2018). rjags: Bayesian Graphical Models using MCMC. R package version 4.8.0.

R Development Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science, 6*(1), 15-32. Retrieved from http://www.jstor.org.ezp.lib.unimelb.edu.au/stable/2245695

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1230-y

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin*.

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine, 34*(13), 2116-2127. doi:10.1002/sim.6481

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine, 36*(10), 1580-1598. doi:10.1002/sim.7228

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*(3), e2000797. doi:10.1371/journal.pbio.2000797

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology, 143*, 1457-1475. doi:10.1037/a0036731

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal Of Statistical Software, 36*(3), 48. doi:10.18637/jss.v036.i03

Wagenmakers, E. J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behav Res Methods, 48*(2), 413-426. doi:10.3758/s13428-015-0593-0

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832

Supplementary material

Approximate Bayes factors comparison

A comparison of the results of the Bayes Factors as estimated here and as reported in (Camerer et al., 2018) shows that they agree with each other in approximate magnitude and direction for the most part, although there are some notable discrepancies. See Table SM1 for the Bayes factors reported in (Camerer et al., 2018) and those reported in the current paper. The only large discrepancy included is seen in Balafoutas and Sutter (2012) in which the Bayes Factor reported in (Camerer et al., 2018) was based on a hypothesis test of ordered binomial probabilities, making it difficult to appropriately convert into a correlation coefficient, and likely accounting for the large difference.

Table SM1. One-sided and (BF_{plus1}) and replication (BF_{rep1}) Bayes Factors for as reported in (Camerer et al., 2018) and as estimated in the current paper, along with the reported correlation coefficients and sample sizes from the original and replication studies.

Article	Original_r	Original_N	Replication_r	Replication_n	Camerer_et_al._BFP0	Camerer_et_al._BFRep0	BFrep0	BF0plus	BF01
Ackerman et al. (2010), Science	0.27	54	0.09	858	5.4e-01	3.1e-01	2.6e+00	2.1e+00	1.0e+00
Aviezer et al. (2012), Science	0.96	15	0.83	14	4.5e+02	5.7e+01	2.3e+02	2.7e+02	1.4e+02
Balafoutas and Sutter (2012), Science	0.28	72	0.15	243	4.2e+00	4.3e+00	4.1e+00	2.1e+00	1.1e+00
Derex et al. (2013), Nature	0.52	51	0.36	65	3.1e+03	3.7e+03	3.3e+01	2.2e+01	1.1e+01
Duncan et al. (2012), Science	0.67	15	0.37	128	2.7e+03	2.5e+03	2.0e+03	2.3e+03	1.2e+03
Gervais and Norenzayan (2012), Science	0.29	57	-0.04	755	6.0e-02	3.0e-02	3.0e-02	2.0e-02	9.0e-02

Gneezy et al. (2014), Science	0.22	178	0.18	407	2.3e+02	4.9e+02	4.7e+02	1.1e+02	5.7e+01
Hauser et al. (2014), Nature	0.82	40	0.83	22	2.9e+03	1.0e+04	1.0e+05	2.5e+04	1.2e+04
Janssen et al. (2010), Science	0.63	63	0.34	42	5.9e+00	0.0e+00	1.9e+00	4.2e+00	2.1e+00
Karpicke and Blunt (2011), Science	0.60	40	0.38	49	1.5e+01	1.2e+01	1.4e+01	1.3e+01	6.5e+00
Kidd and Castano (2013), Science	0.27	86	-0.04	999	5.0e-02	1.0e-02	1.0e-02	2.0e-02	8.0e-02
Kovacs et al. (2010), Science	0.45	24	0.59	95	5.6e+07	1.3e+08	8.7e+07	5.4e+07	2.7e+07
Lee and Schwarz (2010), Science	0.39	40	-0.05	409	8.0e-02	1.0e-02	2.0e-02	3.0e-02	1.1e-01
Morewedge et al. (2010), Science	0.45	32	0.35	89	8.7e+01	1.6e+02	1.6e+02	8.1e+01	4.0e+01

Nishi et al. (2015), Nature	0.20	200	0.12	480	7.0e+00	7.8e+00	8.4e+00	2.9e+00	1.4e+00
Pyc and Rawson (2010), Science	0.38	36	0.15	438	6.8e+00	4.0e+00	1.7e+01	1.6e+01	8.0e+00
Ramirez and Beilock (2011), Science	0.79	20	-0.09	105	1.4e-01	0.0e+00	0.0e+00	7.0e-02	1.9e-01
Rand et al. (2012), Nature	0.14	343	0.03	3150	1.4e-01	1.0e-01	1.3e-01	1.3e-01	7.0e-02
Shah et al. (2012), Science	0.27	56	-0.04	897	7.0e-02	4.0e-02	4.0e-02	2.0e-02	8.0e-02
Sparrow et al. (2011), Science	0.37	69	0.07	338	1.5e-01	3.0e-02	6.0e-02	2.6e-01	1.5e-01
Wilson et al. (2014), Science,	0.67	30	0.59	39	6.0e+02	1.9e+03	1.9e+03	8.3e+02	4.2e+02

Plots and multilevel model output of the relationship between original and replication correlation coefficients using varied exclusion criteria

The following output shows scatter plots and model output for all of the multilevel meta-analyses performed using the varied exclusion criteria explained in the main text.

Table SM 2. Multilevel meta-analysis model estimates and random effects for all data.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.14	-0.21	-0.07	0.03	< .001	
					Project variance = 0.007, n = 8
					Article variance = 0.025, n = 228
					QE(304) = 3527.86, p < .001

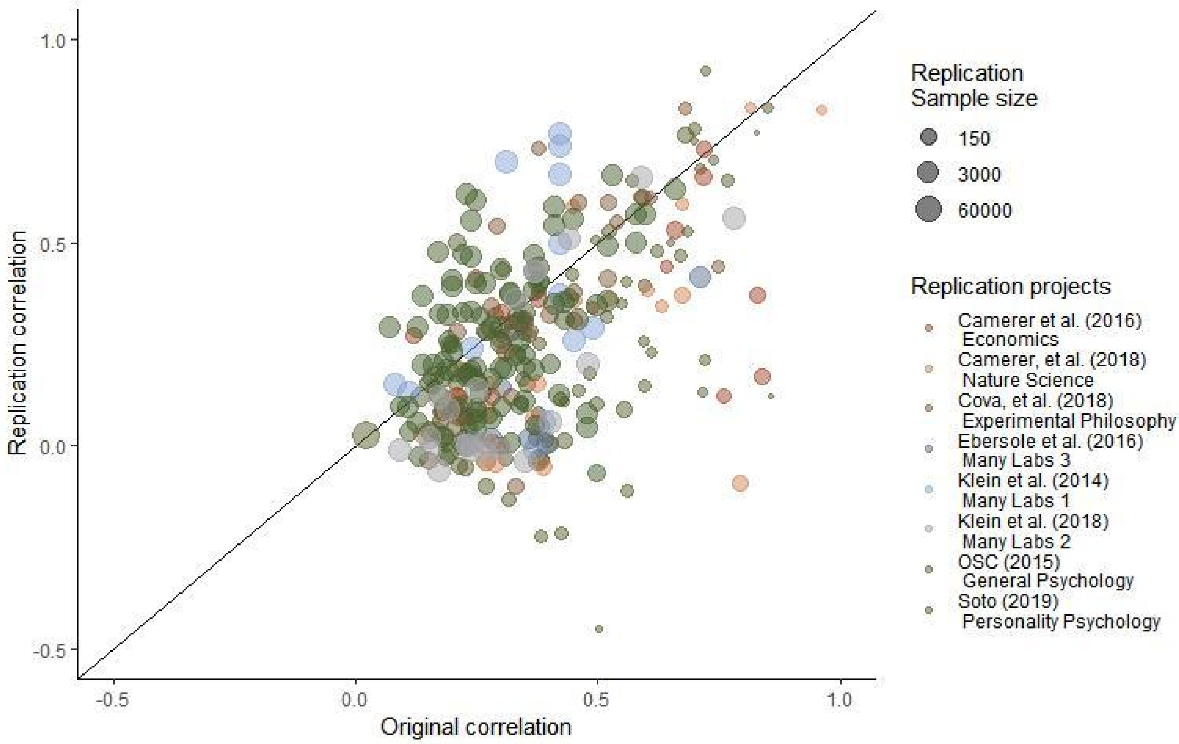


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

Table SM3. Multilevel meta-analysis model estimates and random effects including only statistically significant replications.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.05	-0.12	0.01	0.03	0.1	
					Project variance = 0.006, n = 8
					Article variance = 0.019, n = 129
					QE(194) = 2706.56, p < .001

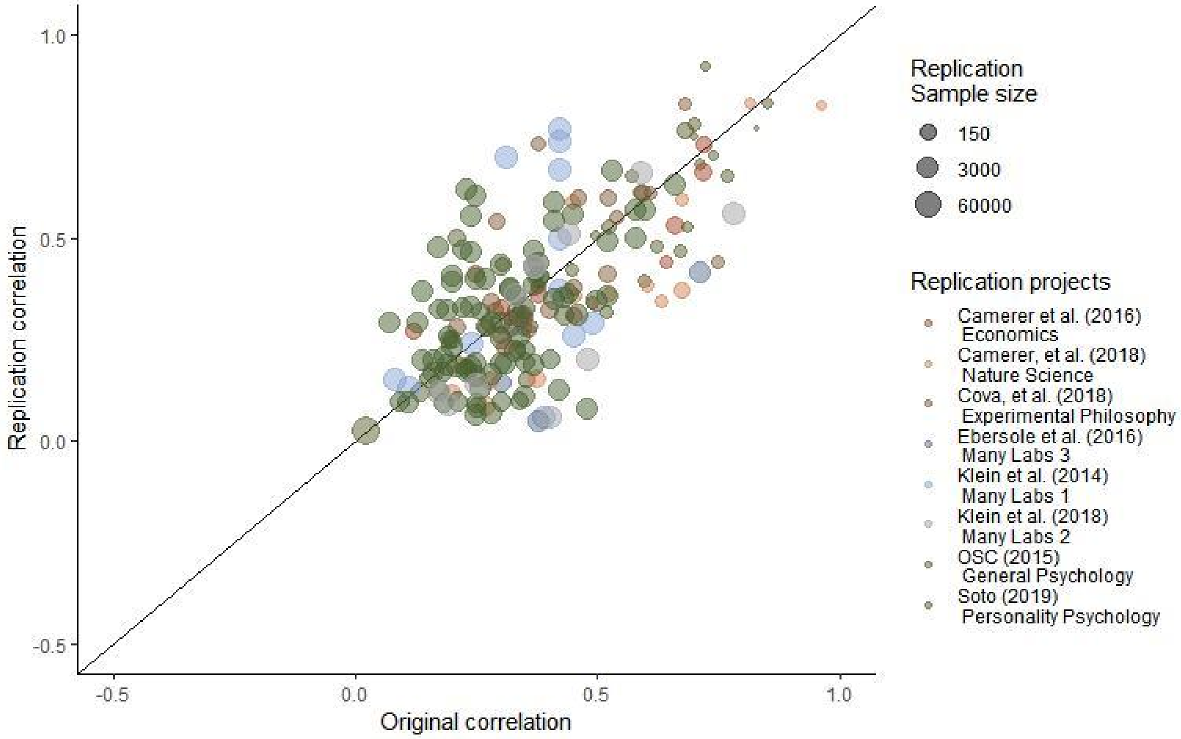


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

Table SM4. Multilevel meta-analysis model estimates and random effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original study.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.08	-0.16	-0.01	0.04	0.02	
					Project variance = 0.008, n = 8
					Article variance = 0.025, n = 167
					QE(234) = 3023.83, p < .001

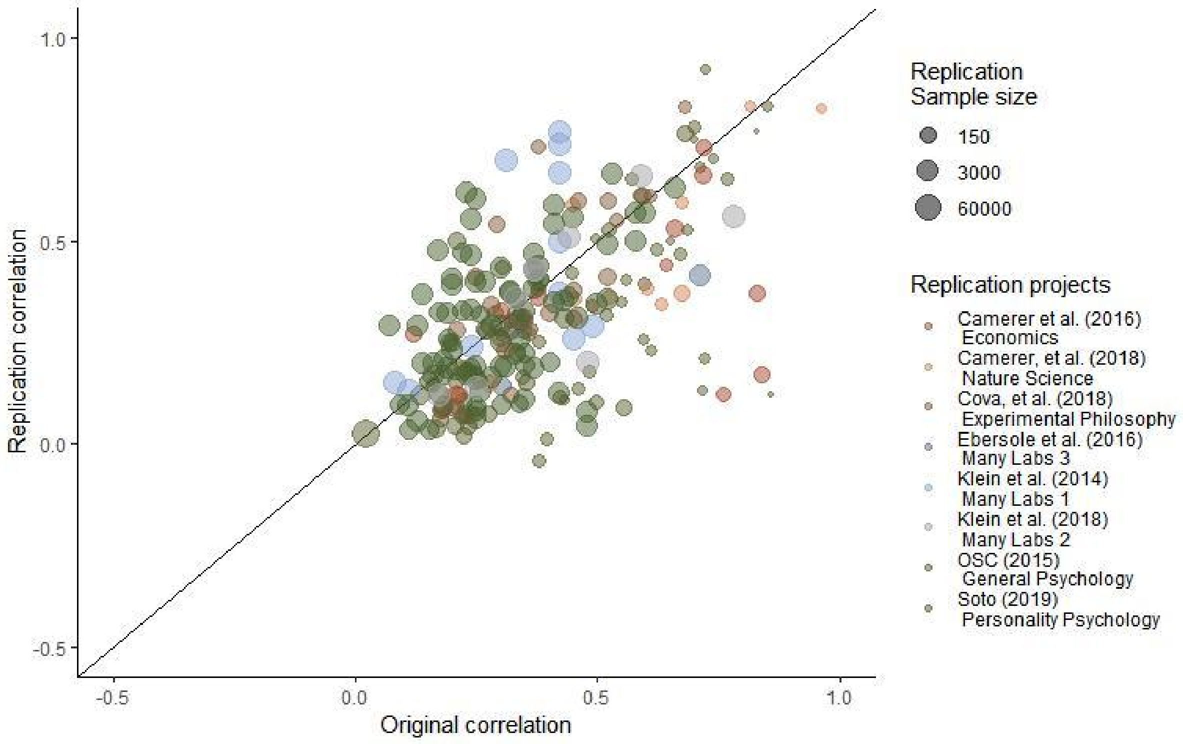


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original study.

Table SM5. Multilevel meta-analysis model estimates and random effects for studies with BF_{01} < 3.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.09	-0.16	-0.01	0.04	0.02	
					Project variance = 0.008, n = 8
					Article variance = 0.026, n = 151
					QE(216) = 2867.77, p < .001

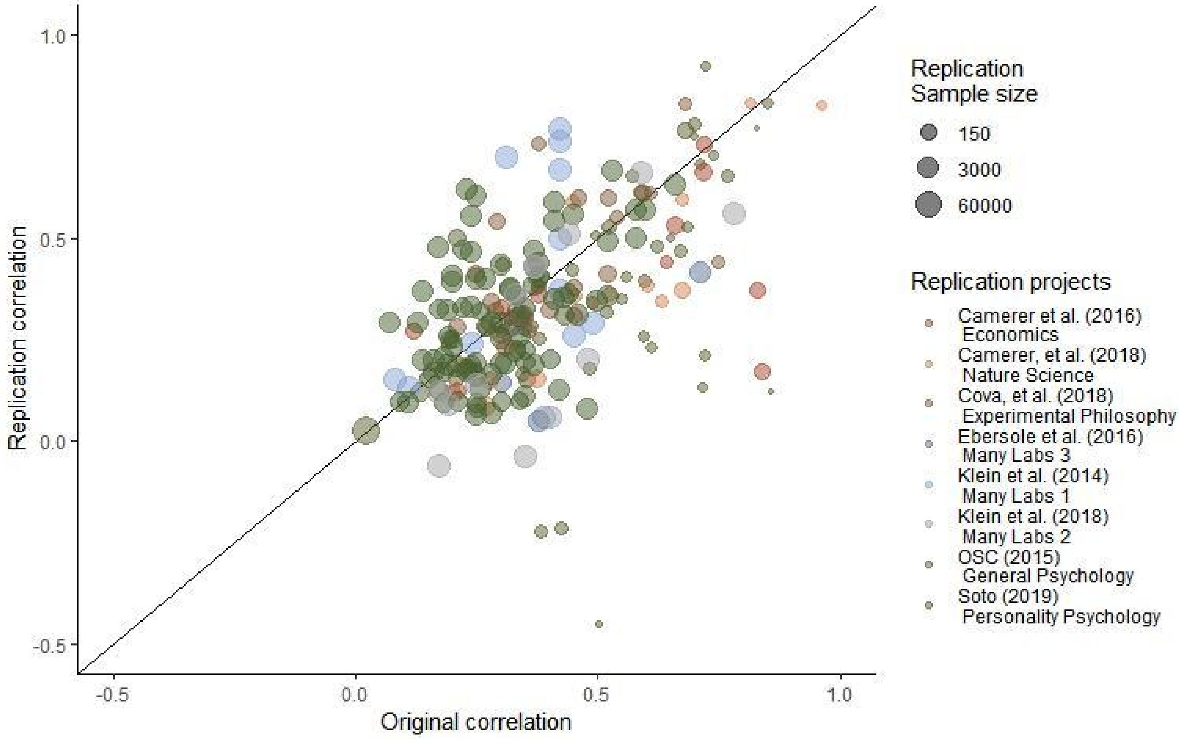


Figure SM4. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with BF_{01} < 3.

Table SM6. Multilevel meta-analysis model estimates and random effects for studies with BF_{10} > 3.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.05	-0.11	0.02	0.03	0.15	
					Project variance = 0.006, n = 8
					Article variance = 0.021, n = 115
					QE(172) = 2516.9, p < .001

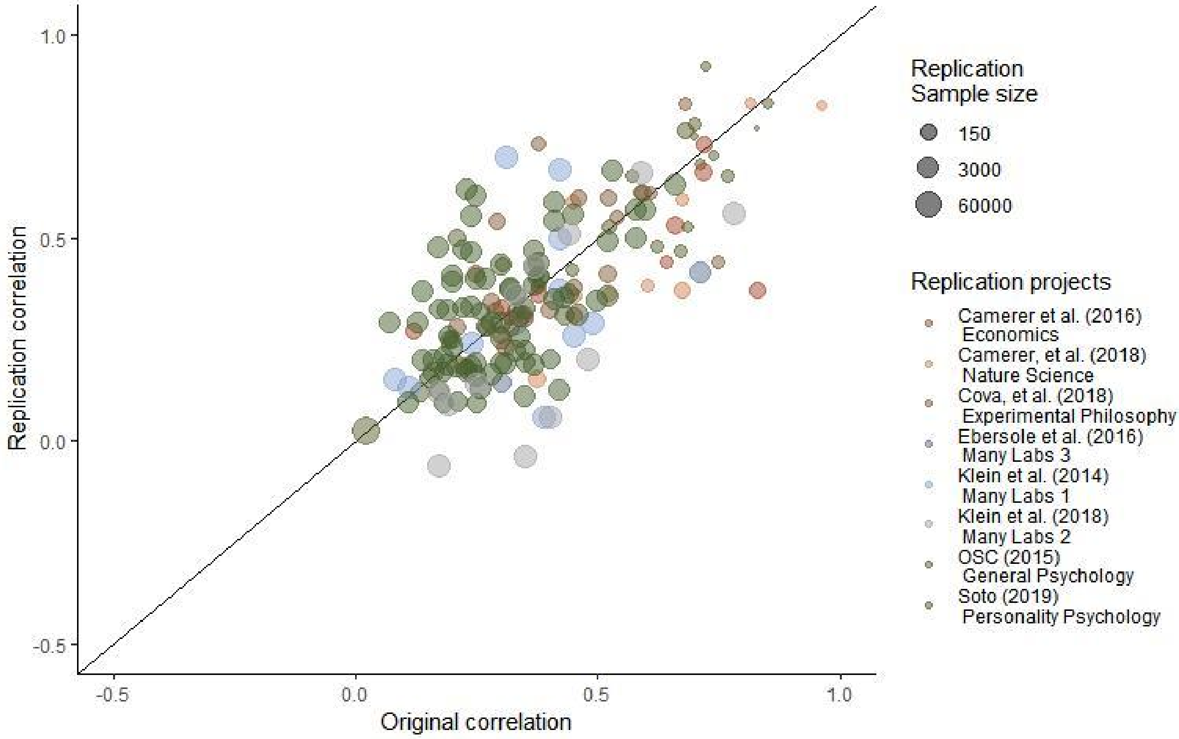


Figure SM5. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with BF_{10} > 3.

Table SM7. Multilevel meta-analysis model estimates and random effects for studies with BF_{0+} < 3.

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.09	-0.16	-0.02	0.04	0.01	
					Project variance = 0.008, n = 8
					Article variance = 0.025, n = 161
					QE(227) = 2885.86, p < .001

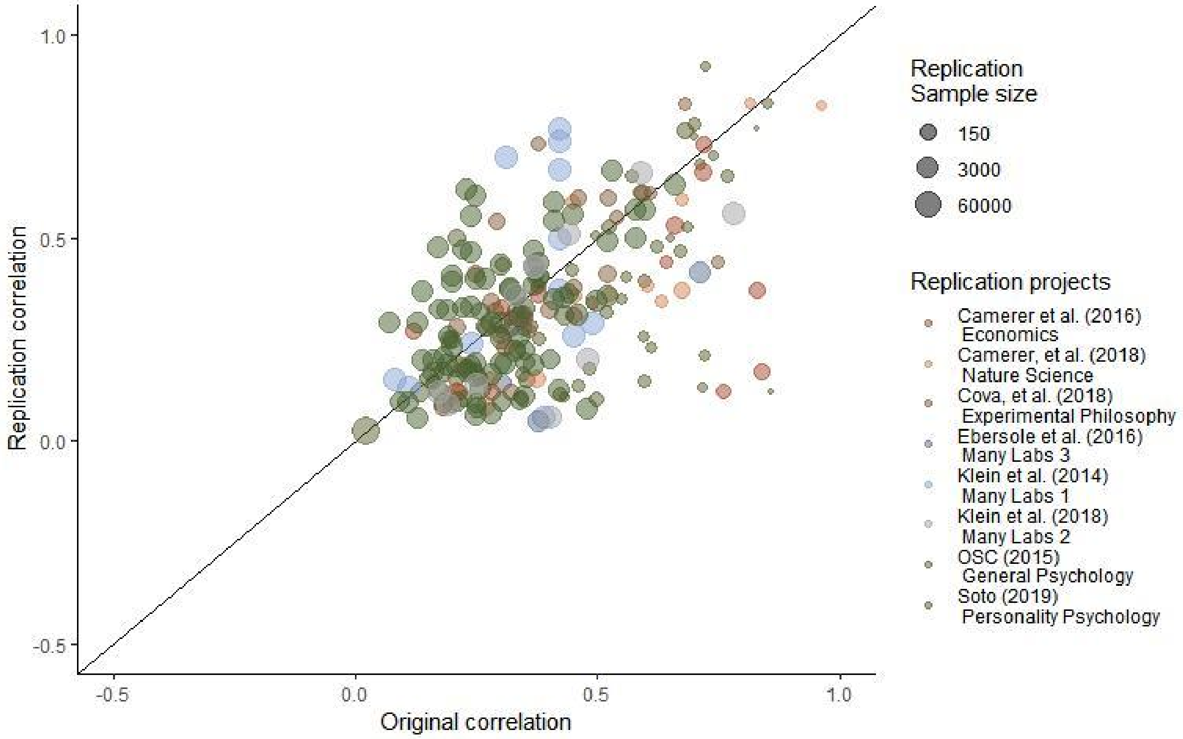


Figure SM6. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with BF_{0+} < 3.

Table SM8. Multilevel meta-analysis model estimates and random effects for studies with $BF_{+0} > 3$

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.05	-0.11	0.02	0.03	0.14	
					Project variance = 0.005, n = 8
					Article variance = 0.024, n = 120
					QE(181) = 2675.47, p < .001

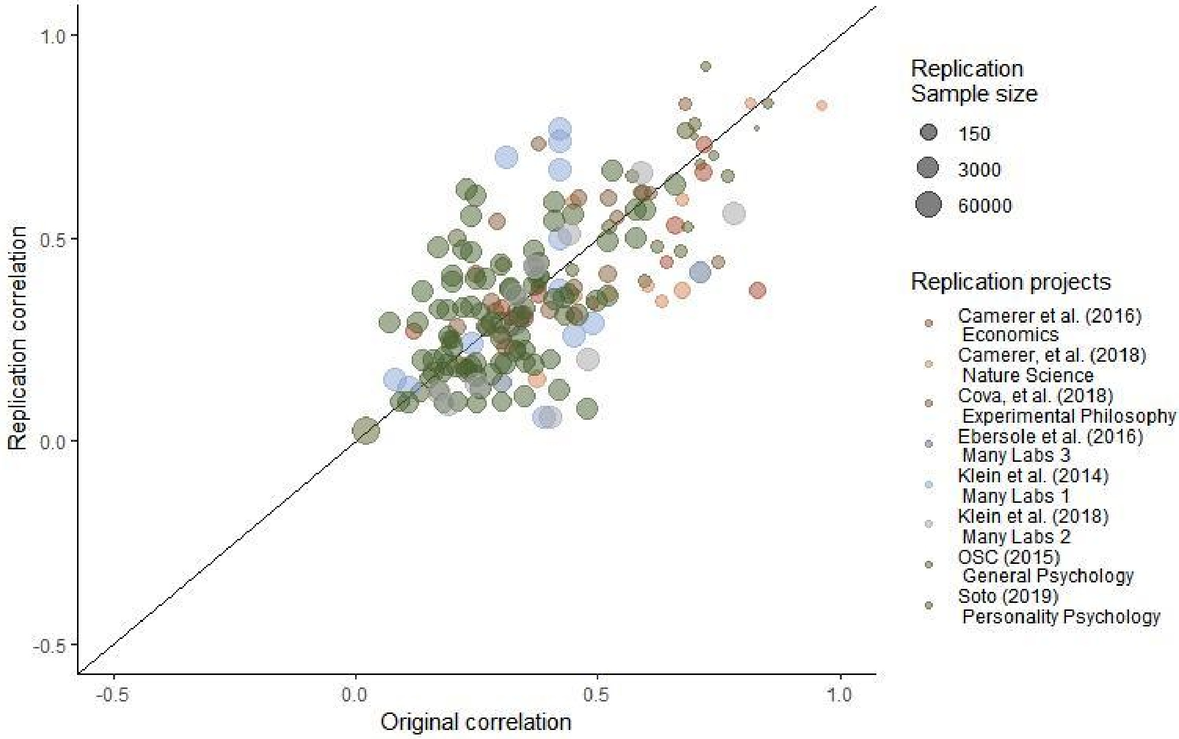


Figure SM7. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with $BF_{+0} > 3$

Table SM9. Multilevel meta-analysis model estimates and random effects for studies with $BF_{rep0} > 3$

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.04	-0.09	0.02	0.03	0.23	
					Project variance = 0.004, n = 8
					Article variance = 0.017, n = 126
					QE(185) = 2457.39, p < .001

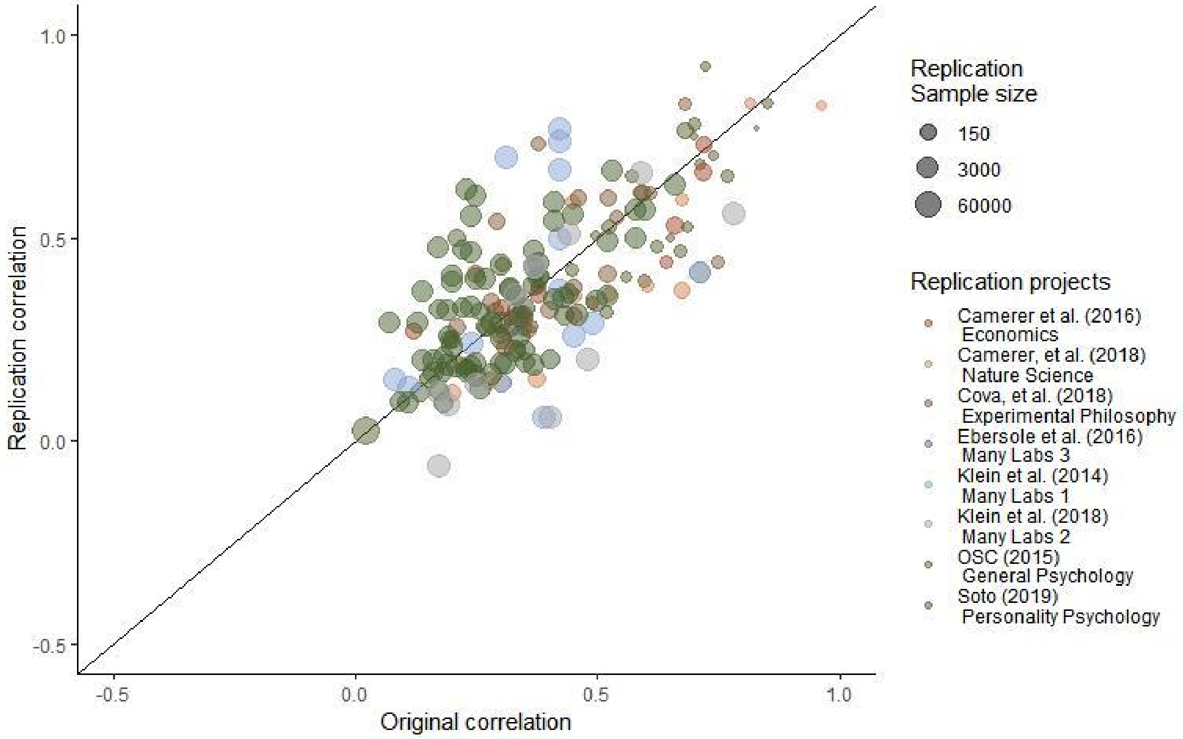


Figure SM8. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with $BF_{rep0} > 3$

Table SM10. Multilevel meta-analysis model estimates and random effects for studies with $BF_{0rep} < 3$

Estimate	95% CI LB	95% CI UB	SE	p	Random effects
-0.06	-0.12	0	0.03	0.05	
					Project variance = 0.005, n = 8
					Article variance = 0.016, n = 159
					QE(219) = 2532.44, p < .001

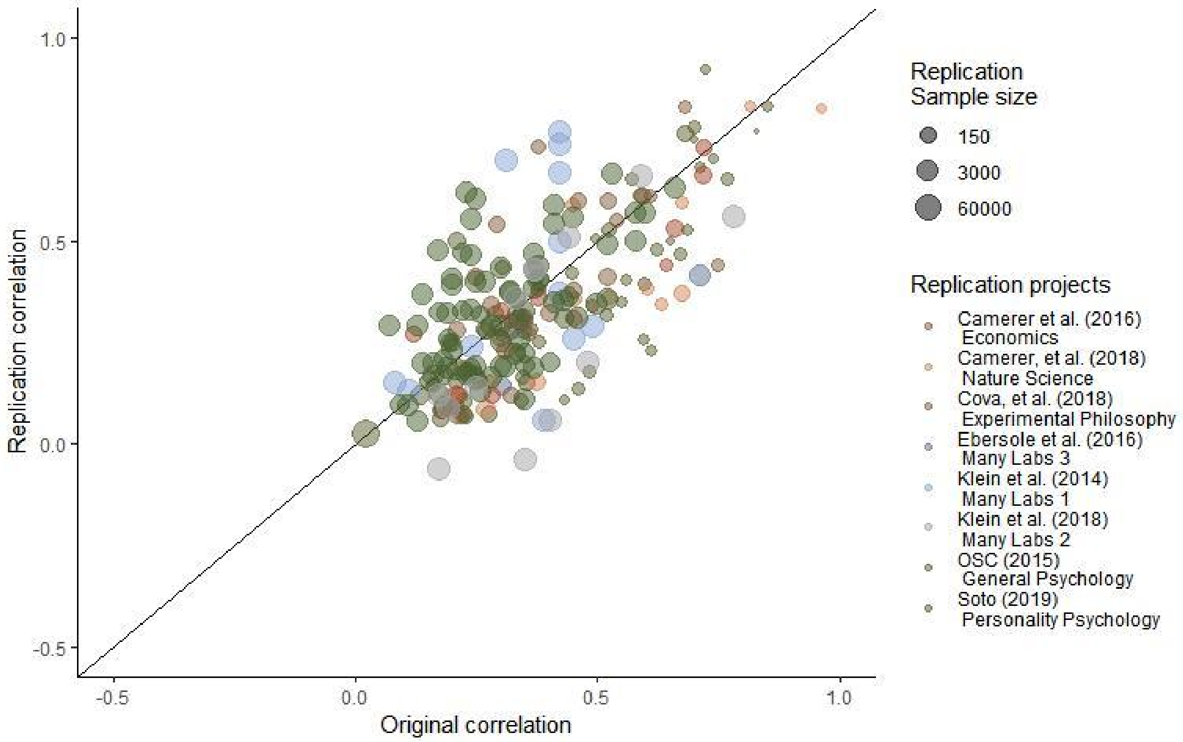


Figure SM9. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with $BF_{0rep} < 3$

Simulation of exclusion method accuracy

In order to assess whether the methods that were used to estimate the proportion change in studies excluding null results develop reasonable estimates, a series of simulations were performed. Simulations took as a starting point the observed effects in the original studies, estimating a true effect from these original results based on the Fisher Transformed ES standard error (i.e., estimating the true effect of each original study assuming a normal distribution with a mean of the original effect and a standard deviation of the standard error), and applying an attenuation factor (i.e., the proportion by which the true effect is reduced between initial and replication studies). Simulations were performed on attenuation factors from 0 to 1 in steps of .1. Simulation studies also varied the number of true effects, also varying between 0 and 1 in steps of .1, setting some studies to have true effect sizes of 0 randomly.

These simulations assumed that the probability of each study being a true null results was unrelated to the original effect size, sample size, source or original paper. See Table [all estimates output] for a table of how each method functions under each set of parameter values, along with the number of simulations that make up each value. See Plots [simulation] - [simulation] for heat maps of the root mean square error (RMSE), the mean absolute error (MAE) and average error are reported below in tables for all models. See table [simulation output] for a table of each method's root mean square error (RMSE), the mean absolute error (MAE) and average error using each exclusion rule.

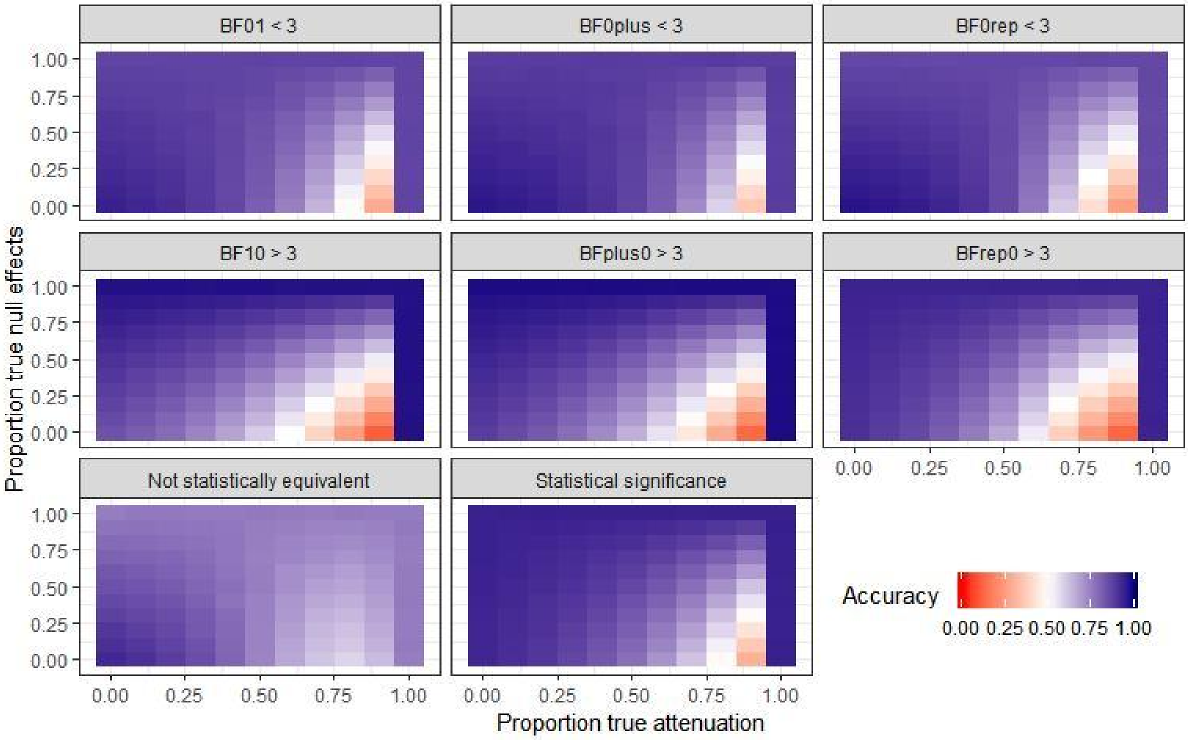


Figure [accuracy of cut scores]. The proportion of studies correctly classified in 11958 simulations of the accuracy of cut scores under varied true proportions of attenuation and proportion of effects which are true nulls.

Table [accuracy plot]

Data inclusion rule	Accuracy	Accuracy SD
Not statistically equivalent	0.75	0.074
Statistical significance	0.84	0.135
BF01 < 3	0.81	0.127
BF10 > 3	0.78	0.127
BFplus0 > 3	0.81	0.188
BF0plus < 3	0.84	0.113
BFrep0 > 3	0.78	0.190
BF0rep < 3	0.81	0.144

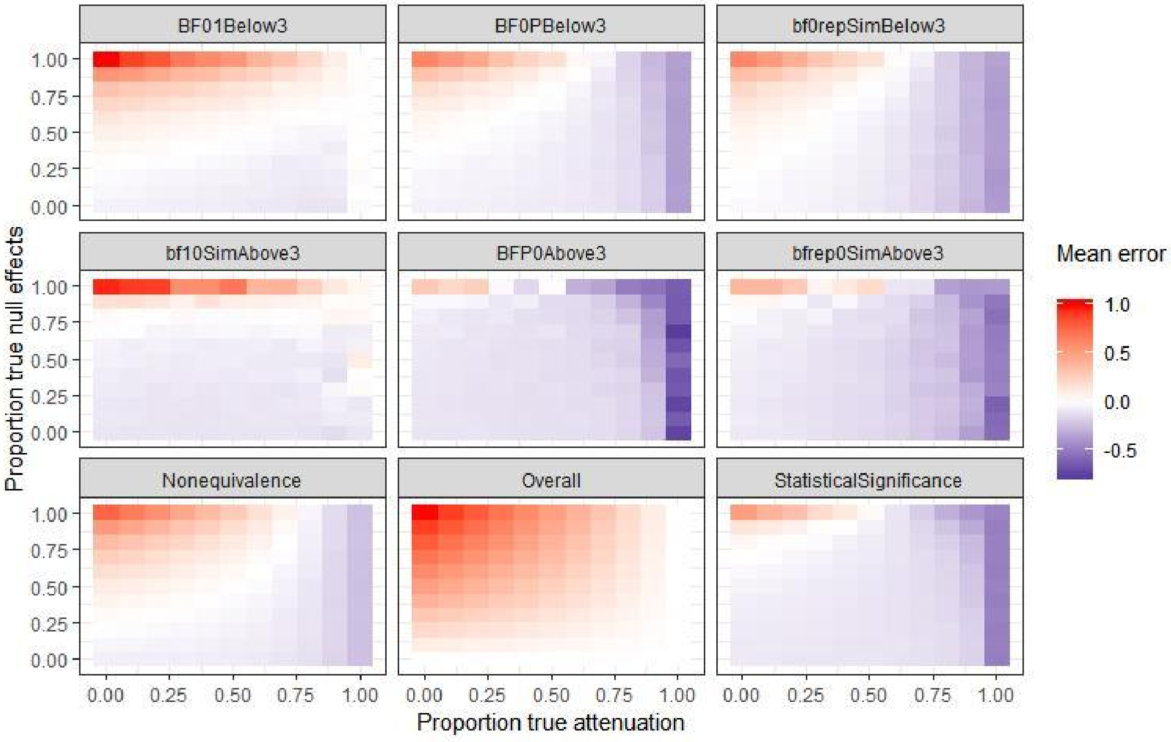
Additionally, simulations were performed using the same data-generation method to estimate the accuracy of the raw methods of estimating the simulated true proportion decrease under these different scenarios. Looking the mean proportion of effect size attenuation in the study, the results of the simulation study suggest that none of these methods for removing effect sizes lead to particularly accurate estimates of the true mean proportion error or the true average reduction in effect sizes in extreme circumstances. The simulation studies show Mean Absolute Errors (MAE) of between 0.13 and 0.25 for estimates of the proportion of attenuation seen, with error standard deviations of between 0.18 and 0.35, compared to a MAE of 0.25 when not removing any studies (error SD = 0.24). However, at reasonable levels of attenuation and proportion of null effects being correct, the simulations suggest that these methods are more accurate. For example, excluding simulations with a proportion of null results or attrition of .8 or greater, these methods have a MAE range of between 0.06 and 0.12, error SDs of 0.04 to 0.08, compared to MAE of 0.23 when not excluding any studies (error SD = 0.18).

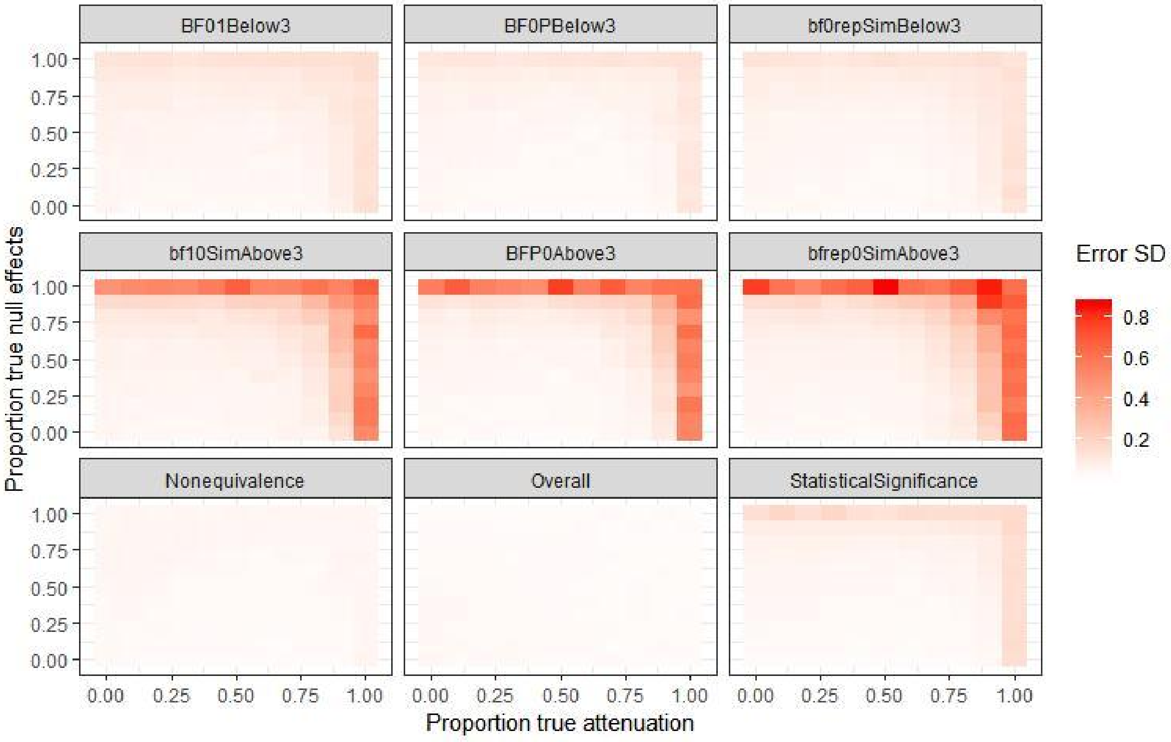
Note that these values are only valid under a specific data generation process, where there is a consistent factor effect size decrease, and where the studies which are null are random and independent of the original effect and sample sizes. See Table [simulation output] for the mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean error (i.e., the average difference between the estimated proportion of effect size attenuation and the simulated amount of effect size attenuation), the error standard deviation, (i.e., the SD of the error scores for each simulation) across the parameter space, and figures 10 to 13 for heat-plots of the simulation mean error, mean absolute error and error SD across simulation conditions. The code used in these simulations is available from [OSFOSF.io].

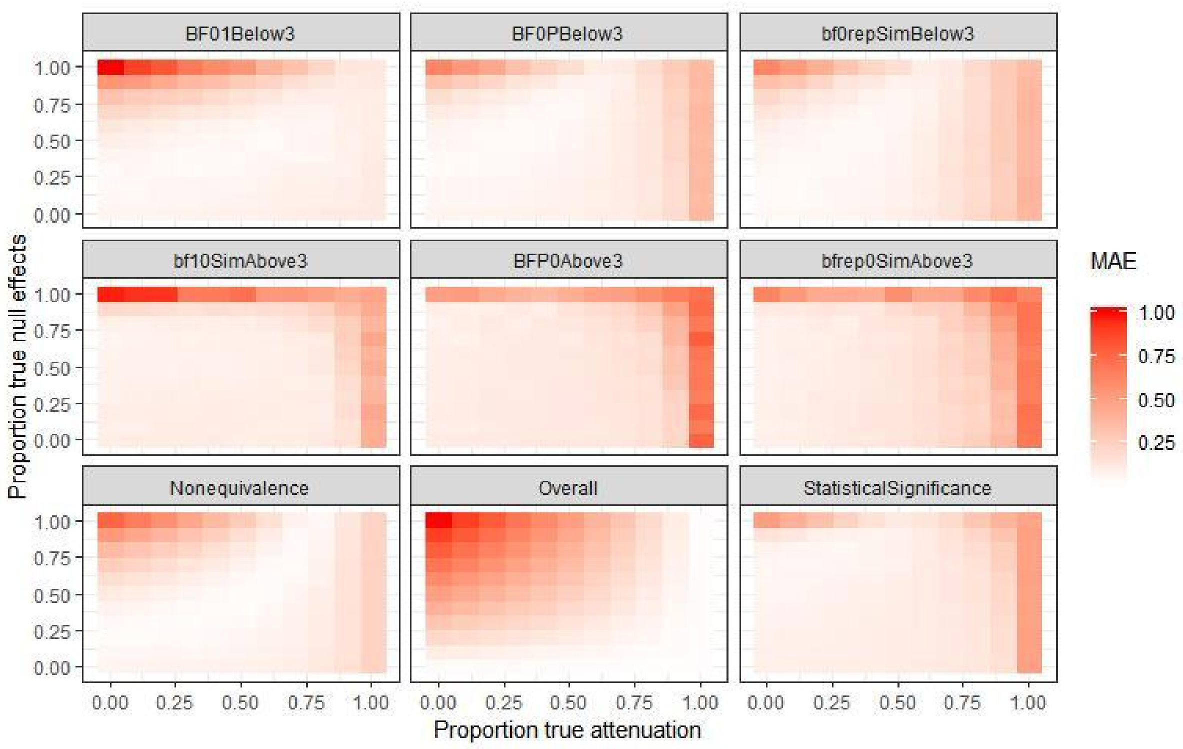
heat maps of the mean absolute error at each benchmark and full simulation output tables. Table [simulation output]. The number of simulations for each subsample, the mean squared error (MSE), root mean square error (RMSE), mean absolute error (MAE), mean error (i.e., the average difference between the estimated proportion of effect size attenuation and the simulated amount of effect size attenuation), the error standard deviation, (i.e., the SD of the error scores for each simulation).

Subsample	nSims	MSE	Mean Error	RMSE	MAE	Error SD
BF01Below3	10101	0.05	0.08	0.23	0.14	0.22
BF0PBelow3	10101	0.04	-0.06	0.19	0.13	0.18
bf0repSimBelow3	10101	0.04	-0.06	0.21	0.15	0.20
bf10SimAbove3	10101	0.11	0.00	0.32	0.19	0.32

BFP0Above3	10101	0.13	-0.18	0.36	0.22	0.32
bfrep0SimAbove3	10101	0.15	-0.17	0.39	0.25	0.35
Nonequivalence	14232	0.04	0.02	0.19	0.13	0.19
Overall	14232	0.12	0.25	0.35	0.25	0.24
StatisticalSignificance	14232	0.05	-0.12	0.21	0.15	0.18







LOO Cross validation output

Table *LOO cross validation output*.

0th, 25th, 50th, 75th and 100th percentiles from leave one out cross validation for each multilevel model, excluding one original article at a time, including only the sample indicated in “subsample”.

Subsample	Proportion significant	Minimum estimate	25th percentile	Median	75th percentile	Maximum estimate
bf01<3	1.00	-0.09	-0.09	-0.09	-0.09	-0.08
bf0plus<3	1.00	-0.10	-0.09	-0.09	-0.09	-0.08
bf0Rep<3	1.00	-0.10	-0.09	-0.09	-0.09	-0.08
bf10>3	0.00	-0.06	-0.05	-0.05	-0.05	-0.04
bfplus0>3	0.00	-0.06	-0.05	-0.05	-0.05	-0.04
bfRep0>3	0.98	-0.07	-0.07	-0.07	-0.07	-0.05
Significant in same direction	0.01	-0.06	-0.05	-0.05	-0.05	-0.04
All studies included	1.00	-0.15	-0.14	-0.14	-0.14	-0.13

Table *LOO cross validation output*.

0th, 25th, 50th, 75th and 100th percentiles from leave one out cross validation for each multilevel model, excluding one replication project at a time, including only the sample indicated in “subsample”.

Subsample	Proportion significant	Minimum estimate	25th percentile	Median	75th percentile	Maximum estimate
bf01<3	0.56	-0.11	-0.10	-0.08	-0.08	-0.07
bf0plus<3	0.89	-0.11	-0.10	-0.09	-0.08	-0.07
bf0Rep<3	1.00	-0.11	-0.10	-0.08	-0.08	-0.06
bf10>3	0.00	-0.07	-0.06	-0.04	-0.04	-0.03
bfplus0>3	0.11	-0.07	-0.06	-0.04	-0.04	-0.03
bfRep0>3	0.33	-0.09	-0.08	-0.06	-0.06	-0.04
Significant in same direction	0.11	-0.07	-0.06	-0.05	-0.04	-0.03
All studies included	1.00	-0.16	-0.15	-0.14	-0.13	-0.13

Bayesian Mixture Model

The mixture model results presented in text presents the model developed by Camerer et al,. (2018; see *https://osf.io/xhj4d/* for their detailed description of this model). All priors were chosen to be uninformative or vague. The mixture model assumes that the observed replication effect sizes either come from the null hypothesis, a true effect sampled from a normal distribution with a mean of zero and a estimated precision (tau). This model uses an errors-in-variables approach to account for possible attenuation of effect sizes due to measurement error and estimation uncertainty following (Matzke et al., 2017), which means the effect size attenuation factor is the factor change between the estimated true effect of the original and replication study effect size. Although this may be reasonable in that the true effect size of the effect may not be the true effect size of a particular study and analysis set up, this poses an interpretative problem in that alpha now represents the difference between the estimated original effect and the replication effect.

Box SM1. The original model reported in (Camerer et al., 2018) and reported on in the main text of the current article.

```
model{
# Mixture Model Priors:
alpha ~ dunif(0,1) # flat prior on slope for predicted effect size under H1
tau ~ dgamma(0.001,0.001) # vague prior on study precision
phi ~ dbeta(1, 1) # flat prior on the true effect rate
# prior on true effect size of original studies:
for (i in 1:n){
trueOrgEffect[i] ~ dnorm(0, 1)
}
# Mixture Model Likelihood:
for(i in 1:n){
clust[i] ~ dbern(phi)
# extract errors in variables (FT stands for Fisher-transformed):
orgEffect_FT[i] ~ dnorm(trueOrgEffect[i], orgTau[i])
repEffect_FT[i] ~ dnorm(trueRepEffect[i], repTau[i])
trueRepEffect[i] ~ dnorm(mu[i], tau)
# if clust[i] = 0 then H0 is true; if clust[i] = 1 then H1 is true and
# the replication effect is a function of the original effect:
mu[i] <- alpha * trueOrgEffect[i] * equals(clust[i], 1)
# when clust[i] = 0, then mu[i] = 0;
# when clust[i] = 1, then mu[i] = alpha * trueOrgEffect[i]
  }
}
```

Supplementary materials [meta-moderaters]

Two methods were used to normalize the distribution of the p values, the Tukey-Freeman double Arcsine transform (Miller, 1978), and the The Ordered Quantile normalization transformation (Peterson, 2018). Residual normality appeared to approximately hold in all cases, and as the results for all methods were functionally identical to those derived from those without any transformation only the raw results are presented in the main text. See tables SM11 [meta-moderators] for the results of the model with normalized predictors.

Table SM11.

Table SM12.