## Chapter 9 - Estimating the effect of publication bias in Behavioural research

## Abstract

{THIS VERSION IS BEHIND THE \_NHB VERSION, UPDATE}

This paper uses the results of over 300 replication studies conducted as a part of eight large scale replication projects – projects which have performed close or direct replications of a set of original articles in order to assess the reproducibility of bodies of behavioural research – to estimate the effect of publication and reporting biases on effect sizes in the behavioural sciences research literature. Using a mixed effects meta-analytic framework to analyse the results of 306 replication studies, we estimate that effect sizes in the published literature are, on average, exaggerated by r = 0.14 (95% CI [0.07, 0.2]), equivalent to a Cohen’s d of 0.28 (95% CI [0.14, 0.42]), or an effect equal to 34% (95% CI [17%, 51%]) of the mean effect size of the original studies (a Fisher’s z transformed correlation equivalent to r = 0.39). This paper also estimates the effect size decrease conditional on the true effect size being non-zero in order to account for the presence of null results. The results from a Bayesian mixture model (Camerer et al., 2018) suggest an average decrease in effect sizes of 19% (with a 95% highest probability density interval of [11%, 28%]). Researchers should be aware that effect sizes in the published literature are exaggerated and account for this when planning, reading and interpreting 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 effects is essential to accurately interpreting and understanding an effect and for planning future research effectively. As such the discovery and precise estimation of effects is essential to developing a coherent and reliable scientific literature. 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 true outcome effect sizes (Hedges, 1992). The degree to which publication bias inflates effect sizes in the behavioural sciences literature is currently unknown. Recent large-scale replication projects – projects which have systematically replicated bodies of research – provide a new resource with which to estimate the extent of effect size inflation in the behavioural sciences literature.

All of the included replication projects were primarily conducted in order to assess the degree to which their particular area of research contains results which are irreplicable, or to estimate variability in effects among subpopulations. Most of these studies suggest that the replicability of the published literature 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, the percentages of replication attempts with a statistically significant result in the same direction as the original study, and the number of studies from each project included in the current analysis.

*Table 1.* A list of each included replication project, the number of replication studies performed as a part of each replication project, the percentage of replication studies which were “successful” (defined here as replication studies which found statistically significant in the same direction as the original study), the number of studies for which are included in the current study, and the percentage of each project’s studies which are included in the current analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Replication projects | Number of replication studies performed | Replication results statistically significant in the same direction as original (%) |  | Included studies | Percent of performed replication studies included in the current project |
| 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% |  | 18 | 100% |
| 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% |  | 21 | 100% |
| 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. | 37 | 78% |  | 33 | 89% |
| 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> | 9 | 33% |  | 8 | 89% |
| Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., BahnÃ­k, Å ., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating Variation in Replicability. Social Psychology, 45(3), 142-152. <doi:10.1027/1864-9335/a000178> a | 16 (13 effects) | 88% (85%) |  | 15 | 94% (92%) |
| Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances In Methods and Practices in Psychological Science, 1(4), 443-490. <doi:10.1177/2515245918810225> | 28 | 54% |  | 22 | 79% |
| Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. <doi:10.1126/science.aac4716> | 97 | 36% |  | 89 | 92% |
| Soto, C. J. (in press). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes Of Personality Replication Project. *Psychological Science*. | 121 | 86% |  | 100 | 83% |
| All projects | 347 | 65% |  | 306 | 88% |

Note: a Klein et al. (2014) includes 4 operationalisations of a single effect which were input separately for analysis in the current study, the bracketed values in the table refer to the number of results at the effect level. aSoto’s (in press) replication rate was recalculated on the “study” level (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. Soto (in press) uses as its primary unit of analysis trait-outcome associations, and found that 87% of the 78 trait-outcome associations were supported by a significant result in the same direction.

### The effect of publication and reporting bias on reported effect sizes

Publication bias is the process by which studies that report finding results supporting their hypotheses, usually by showing statistically significant results, are more likely to be published than those that do not (Lane & Dunlap, 1978; Mahoney, 1977){Sterling, 1959 #1043}. 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). The extent to which effect sizes are exaggerated in a literature depends, primarily, on the degree to which non-significant results are excluded, the true statistical power of studies (a product of the effect size and sample size of studies given the experimental design and analysis strategy) and the proportion of true nulls being investigated (Hedges, 1992; Oakes, 1986). If anything, publication bias towards statistically significant results appears to be particularly acute in the behavioural sciences research literature, with an estimated 92% of papers reporting a significant finding (Fanelli, 2012) and around 75% of all reported p values being below .05 (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016), despite the low estimated average power of psychological research (Szucs & Ioannidis, 2017){Cohen, 1962 #487}.

Selective reporting among measures and Questionable Research Practices (QRPs) like p-hacking and Hypothesising After the Results are Known (HARKing) on the basis of statistical significance or the size of effects (Kerr, 1998) can also lead to effect sizes being exaggerated and increase the proportion of false positives in a scientific literature (Bakker, van Dijk, & Wicherts, 2012; Murphy & Aguinis, 2017; Simmons, Nelson, & Simonsohn, 2011). Recent surveys in the behavioural research literature 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). The current paper provides an estimate of the cumulative effect of publication bias and QRPs on published effect sizes, an essential piece of information for researchers to accurately reading and interpreting the scientific literature.

### The effect of publication and reporting bias in the behavioural sciences literature

We have identified only a single previous study which attempts to empirically estimate the effect of publication bias seen in the psychology literature. Stanley, Carter, and Doucouliagos (2018) used several meta-analytic bias adjustment methods (specifically using three different estimators; WLS, WAAP-WLS, and PET-PEESE) in a reanalysis of 200 meta-analyses published in Psychological Bulletin. They report finding a median effect size exaggeration of 8 to 15%, with the range depending on the meta-analytic bias reduction method used. As Stanley et al. (2018) points out, this study estimates the effect of publication bias using a likely unrepresentative literature, and it is unclear how much this estimate is likely to generalize outside of the pages of Psychological Bulletin.

In order to estimate the effect of publication and reporting bias on effect sizes published in the behavioural sciences literature, the current study presents an exploratory analysis of recent replication projects using three analytic approaches. The first approach reported here uses a multilevel meta-analytic framework to estimate the expected effect size change between original replication studies. As this database is likely to include effects which are true null effects (or effects which are so close to true null effects to be practically dismissible), the remaining analyses are designed to estimate the degree to which effect sizes reported in the literature are exaggerated assuming that the effect under study is non-null, arguably of more use for people attempting to plan studies based on the published literature. As such, the second approach reported here reperforms the random effects meta-analyses on subsets of studies; excluding studies on the basis of (a) statistical significance and (b) equivalence testing. A final approach estimates the expected effect size attenuation from original to replication study and the true null rate using the Bayesian Mixture Model presented in Camerer et al. (2018).

## 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 {Klein, 2014 #988;Ebersole, 2016 #985}. In these cases, these values were manually extracted from the original articles. When sample sizes for original studies were not reported in the data provided by each replication project they were manually extracted from original articles where possible.

For analysis, the original and replication effect sizes were transformed to Fisher z Transformed correlation coefficients following the methods used in {Open Science Collaboration, 2015 #611}. Cases where the study-level results were not reported in correlation coefficients, Cohen’s d values, as t-tests, or as F statistics were excluded from analysis (e.g., cases when no effect size was reported in the original study or in the replication project data set). In cases where sample sizes were not reported per group, sample sizes among groups were assumed to be equal in these conversions. See supplementary materials 1 for a comprehensive account of exclusions and study specific extraction details for each replication project. See Table 1 for the number of valid studies extracted from each project. An original and replication effect size that could be converted to a Fisher z score along with sample sizes for original and replication studies was extracted for a total of 306 studies.

### Analysis

This paper uses three analytic approaches to examine the change in effect sizes from original to replication studies. All analysis was performed in R version 3.5.1 {R Development Core Team, 2018 #314} and meta-analyses were performed using the Metafor package version 2.1 {Viechtbauer, 2010 #796} using restricted maximum-likelihood estimation. All analyses and difference scores (i.e., proportion changes and mean differences) were calculated using Fisher Z transformed effect sizes, and effect sizes are back transformed to correlation coefficients for easy interpretation unless otherwise stated. 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 specifying that all analyses would be exploratory.

##### Analysis 1: Multilevel random effects meta-analysis

The first approach uses a random effects meta-analysis framework to estimate the expected effect size decrease between original and replication studies.

This analysis treats each pair of effects, the original and replicated sample size, as one “study” in a meta-analytic framework, and estimates the change from original to replication study (). In order to account for non-independence between the included difference scores, random effects for replication project () and original article () are included (as in some cases, multiple effects from an original article were replicated or multiple operationalisations of an original effect were tested), and random effects are also included for each individual effect (). Standard errors for each difference score were estimated as , with being the sample size in the original study and being the sample size in the replication study. This standard error is an approximation for the included F tests with a greater than 1 and chi square tests, and in order to check whether this was strongly impacting results all multilevel meta-analyses were re-performed excluding these studies. No differences in the substantive interpretation of results would follow from this change (i.e., did not change by more than 0.02, no statistical tests changed their statistical significance at the .05 alpha level, and variance partitioning did not change enough to alter interpretations). Empirical Bayes estimates and 95% credible intervals for random effects were calculated following {Robinson, 1991 #999}{Morris, 1983 #1000}.

#### Accounting for null results

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. The average effect size difference between original studies and their replications could be could be extremely high, and yet but this effect could be entirely driven by the presence of null effects (e.g., if 50% of studies examined scenarios where there was no between-group difference or association at the population level, and yet all non-zero effects are identical to those reported, the average attenuation would be 50%). In order to account for this issue, three methods are used.

##### Analysis 2 and 3: Multilevel random effects meta-analysis with exclusions

The first method is to re-perform the meta-analysis excluding studies using two exclusion criteria; (a) excluding studies in which the replication study was not statistically significant (using an alpha of .05 and a two-tailed test when applicable) and with an effect in the same direction as the original and (b) by removing statistically equivalent studies. Because both of these methods function by removing small replication effects, no significance testing was performed on the difference between the model estimates estimated after accounting for small or near-null effects.

Analysis 2, excluding studies in which the replication study was not significant, means that replication studies which have a low level of statistical power to detect the true effect size under study are likely to be excluded. Especially as in some of the replication projects the sample size in the replication study was chosen using a power analysis of the observed effect in the original study {Open Science Collaboration, 2015 #611}, this method is likely to underestimate the amount of effect size exaggeration due to the exclusion of under-powered replications.

In order avoid excluding under-powered studies erroneously analysis 3 excludes studies based on whether the results are statistically equivalent to the null hypothesis, or statistically significant in the opposite direction {Lakens, 2017 #214;Lakens, 2018 #951}. As a requirement for equivalence testing is that an equivalence bound is selected (defining the null hypothesis), we follow one suggestion in {Lakens, 2018 #951} and use the lowest effect size that would be statistically significant to the original study (assuming an alpha of .05). Equivalence tests were performed using Z tests of the Fisher Z transformed effect sizes, excluding studies where the observed replication effect is significantly smaller than the equivalence bound using a one tailed test at the 95% confidence level. Standard errors of each study were estimated as , except for studies from {Camerer, 2018 #967} which had more than a single replication attempt, where standard errors are those derived from the meta-analyses that produced the effect size estimate (see supplementary materials [exclusion rules] for details).

In interpreting results based on this exclusion criterion, it is important to note that the minimum detectable effect was occasionally quite high as original sample sizes were often very small (mean equivalence bound in correlation coefficient terms = 0.18, SD = 0.11, 0th, 25th, 50th, 75th and 100th quintiles = [0, 0.1, 0.15, 0.23, 0.63]). This means that original studies were sometimes under-powered to detect even large effects using the current analysis, and as such this method may exclude studies which have replication effects the original authors may have considered important {Thompson, 2002 #1039}. See supplementary materials 2 for scatter plots of the data-set using each exclusion rule.

##### Analysis 3: Bayesian mixture model

The final approach to estimating the amount of effect size attenuation expected given that the effect under study is non-zero was the Bayesian mixture model presented in {Camerer, 2018 #967}. This model assumes that each observed replication effect size comes from one of two components, either from the null-hypothesis or from the alternative-hypothesis component. If the replication effect size is drawn from the null-hypothesis component, it is assumed to have come from a normal distribution with a mean 0 and a standard deviation equal to the standard error of the replication study (estimated here as , n being the replication sample size). If the replication effect size is sampled from the alternative distribution, 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. The attenuation factor is constrained 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, 2018 #967}), the degree to which effect sizes are attenuated between original and replication studies, and the overall rate at which studies are assigned to have come from the null hypothesis (the “assignment rate”). This analysis was performed in JAGS version 4.3.0 {Depaoli, 2016 #1010} using the rjags interface (version 4.8.0; {Plummer, 2018 #1011}). See supplementary materials [“mixture model”] for model syntax and further analysis details.

### Descriptives

Looking at the 306 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 219 articles, 72% of the included studies. An exact binomial test shows that this is extremely unlikely under the assumption that replication effect sizes are equally likely to be smaller or larger in the replication study, p = < .001. The average effect size for original studies was r = 0.39, and the mean effect size for replication studies was r = 0.27, a mean decrease of r = 0.11. Notably, this represents an average decrease in effect sizes from the original to the replication study of 28%. 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.

Examining just the 198 replications in which the replication study was statistically significant, 65% of all studies, the average effect for original studies was 0.4, and the mean effect size for replication studies was 0.39. This represents a mean decrease of r = 0.02, a mean percentage increase in effect sizes of 4% and a median decrease of 7%. Using equivalence testing, a total of 237 studies were not statistically equivalent to the null, 77% of included studies. 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 = 0.06, a mean percentage decrease of 7%, and a median decrease of 17%.

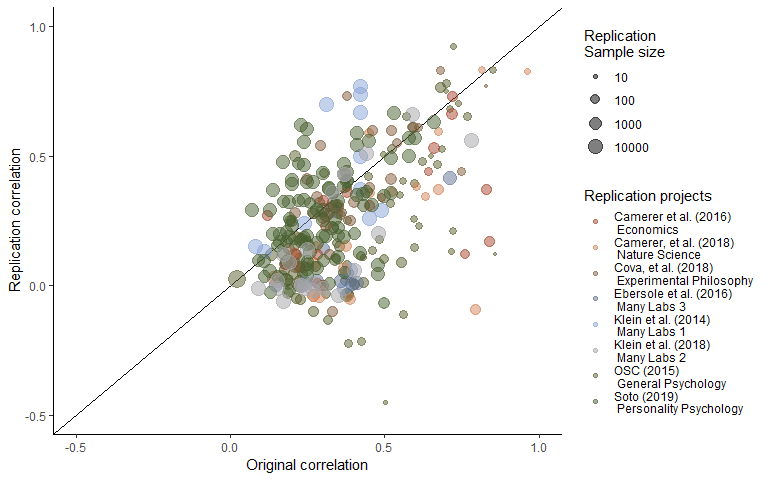


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of participants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

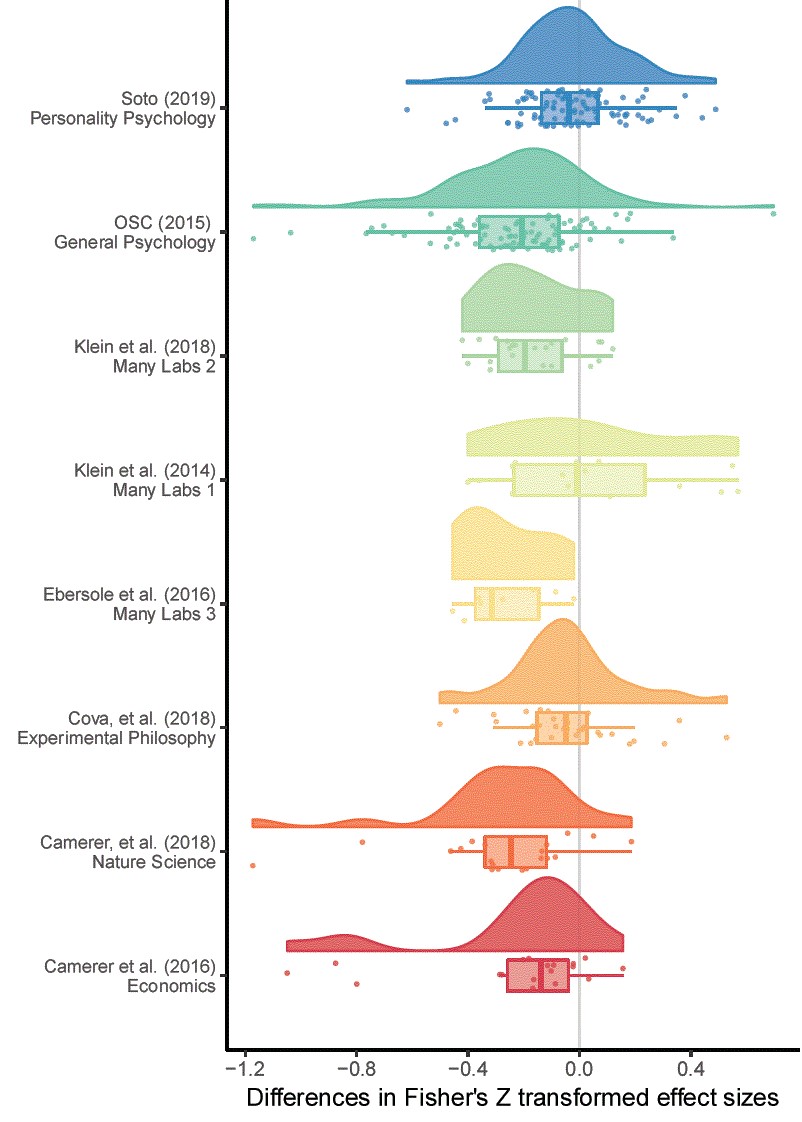


Figure 2. A raincloud plot of the change in effect sizes (here Fisher Z scores) from the original to the replication study by the replication project that each replication study was performed as a part of.

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | All studies | statistically significant replications | Nonequivalent studies |
| n included | 306.00 | 198.00 | 237.00 |
| Mean original ES | 0.39 | 0.40 | 0.41 |
| Median original ES | 0.33 | 0.35 | 0.35 |
| Mean replication ES | 0.27 | 0.39 | 0.35 |
| Median replicaiton ES | 0.22 | 0.33 | 0.30 |
| Mean ES difference | -0.13 | -0.02 | -0.07 |
| Median ES difference | -0.11 | -0.03 | -0.06 |
| SD difference | 0.25 | 0.20 | 0.24 |
| Mean proportion change | -0.28 | 0.04 | -0.07 |
| Median proportion change | -0.34 | -0.07 | -0.17 |

### Results

#### Analysis 1: Multilevel random effects meta-analysis results

The random effects meta-analysis including all data estimated a r = -0.14 (95% CI [-0.2, -0.07]) decrease in effect sizes from the original to replication studies. This represents a decrease equivalent to 34% (95% CI [51%, 17%]) of the mean effect size in the original studies (a Fisher Z transformed correlation coefficient equivalent to a correlation coefficient of 0.37).

More variance was attributable to the article and effect level than to the project ( = 0.016, = 0.012, compared to = 0.008), representing an intraclass correlation (ICC) for the project of 0.215. There was a large amount of unexplained heterogeneity, QE(305) = 3531.9, p < .001, = 92.585 (calculated following {Nakagawa, 2012 #1023}), suggesting that 93% of variance in effect sizes was due to heterogeneity (i.e., variance in the true effect size differences), while the remaining 7% was attributable to sampling variance.

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 |
| -0.137 | -0.206 | -0.068 | 0.035 | < .001 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.016, n = 229 |
|  |  |  |  |  | Effect variance = 0.012, n = 306 |
|  |  |  |  |  | QE(305) = 3531.9, p < .001 |

#### Analysis 2 - 3: Results from multilevel random effects meta-analysis with exclusions

Looking at the estimated effect size decrease only including replications with statistically significant findings, there was an estimated r = -0.05 (95% CI [-0.11, 0.01]) change in effect sizes from original to replication studies. Including only the statistically equivalent studies leads to a predicted r = -0.08 (95% CI [-0.15, -0.01]) decrease in effect sizes. The estimates of the proportion of variance attributable to the article or replication project level did not change considerably in either of these subsets. See table [all model output] for the model estimates from each model.

These values represent changes equivalent to a decrease of 12% to 20% of the average original effect size (a correlation coefficient of r = 0.39). However, there was considerable imprecision in these estiamtes, with 95% confidence intervals for both of these subsamples extending from a considerable decrease equivalent to 38% of the average original effect size, to a small increase equivalent to 2% of the average original effect size.

##### 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 of the mean original effect size (r = 0.37).

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | All studies | Statistically significant replications | Nonequivalent studies |
| Model N | 306.00 | 198.00 | 237.00 |
| Model Estimate | -0.14 | -0.05 | -0.08 |
| 95% CI lb | -0.20 | -0.11 | -0.15 |
| 95% CI ub | -0.07 | 0.01 | -0.01 |
| Estimated % attenuation | -33.67 | -12.41 | -20.20 |
| LB % attenuation | -50.59 | -27.25 | -38.01 |
| UB % attenuation | -16.74 | 2.43 | -2.40 |

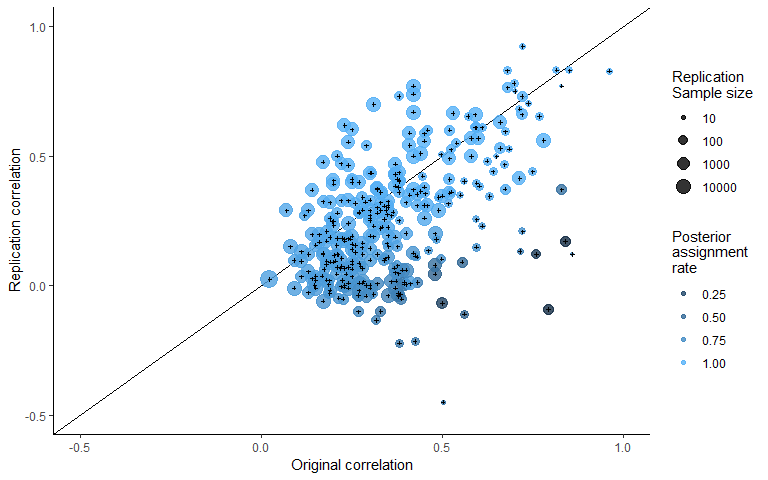
##### Leave one out cross validation of meta-analyses

To assess how sensitive the results of the multilevel models were to the inclusion of each of the replication projects, the included studies, and the individual replicated effects, all of the above multilevel models were rerun using leave one out cross validation, excluding both each effect, effects from each original study (i.e., in cases where multiple effects were tested from the same original source), and each replication project one at a time. None of these analyses led to model estimates (i.e., the expected decrease in effect size between original and replication study or equivalently the intercept estimate) that were further than 0.02 from those given above, suggesting that none of the individual projects, effects or studies included were overly influential. See supplementary material [LOO] for tables summarising the leave-one-out model output.

#### Analysis 4: Bayesian mixture model results

The Bayesian mixture model was estimated using four Markov chains from each of which 100,000 draws were taken (excluding an 11,000 draw burn-in period). Trace and density plots for the discussed parameters were assessed and the model appeared to have converged. The overall posterior assignment rate (i.e., the proportion of studies which are estimated to be from the non-null alternative hypothesis) is 89%, with a 95% highest probability density interval of [79%, 98%]. The overall attenuation factor (i.e., the estimated amount that effect sizes decreases between the original and replication studies) is 19% 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 color of each point indicates how often each effect was assigned to the alternative hypothesis.

As was seen and pointed out in the first use of this model in {Camerer, 2018 #967}, values close to the diagonal (i.e., cases in which the original and replication effect sizes are similar) are reliably assigned to the alternative hypothesis whereas effects far below the diagonal are more often assigned to the null hypothesis. The overall posterior assignment rate might be overly optimistic (i.e., assign studies to the non-null hypothesis at a high rate), likey in part due to the fact that this model allows for “true” effect sizes to be estimated as being extremely low or near zero and still assigned to the alternative hypothesis, with 29% of the estimated “true” replication effect sizes being smaller than a correlation coefficient 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 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

These results show that there was a substantial average decrease in effects sizes between the original and replication study and suggest that this is still the case even after accounting for the presence of non-zero effects. The results of the multilevel meta-analysis results show 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 [-0.42, -0.14]), or an estimated decrease of 34% (95% CI [51%, 17%]) of the mean effect size in the original studies (a equivalent to r = 0.39).

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 agreed, although the degree of precision differs. The Bayesian mixture model estimates that there is an average decrease of 19%, with a 95% highest probability density interval of [11%, 28%]. The multilevel models estimated excluding non-significant results and studies in which the replications studies were statistically equivalent to the null lead to similar conclusions, although they highlight the degree of uncertainty in this result. All models estimated using data-exclusions showed a lower effect size decrease when attempting to exclude null (or effectively null) effects with the confidence intervals over both of these models extending from a decrease of r = -0.15 to an increase of r = 0.01.



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

In interpreting these results it is important to note several limiations. Firstly the current study does not attempt to distinguish between effect size heterogeneity (i.e., effect sizes that are different under due to subtle unobserved moderators {Kenny, 2019 #1041} and effect size exaggeration. However, in so far as effect size heterogeneity tends to lead to smaller effects in replications studies, it seems reasonable to term this effect size exaggeration for the purposes of researchers hoping to replicate or plan future similar studies of the same type of effects.

The Bayesian mixture model presented above assumes independence between effects and a uniform attenuation factor across all areas of psychological research, and allows for modeled true effect sizes to be negligibly small or even negative and still assumed to be sampled from the alternative distribution. Future research could help develop a more nuanced account of the underlying data-generation process by, for example, building a model that allows for the attenuation rate to change across replication studies, or by including more components in order to allowing for studies with negligible or negative but non-null effects in addition to the true alternative and null components modeled.

Most importantly, none of the projects included in this analysis replicated true random selections from the literature, and the sampling strategies of the replication projects included vary widely (e.g., Soto, in press, examines studies included in a previous overview of trait-outcome associations, cf. Camerer et al., 2018 which only included studies published between 2010 and 2015 in the journals Nature and Science). It is possible that the effect size decreases seen here are systematically different from what would be seen across the behavioural sciences literature. While these results should be considered preliminary, this analysis nonetheless provides an initial estiamte of the amount of effect size attenuation that should be expected when planning research, and suggests that even accounting for the presence of null effects the amount of effect size attenuation between the published literature and replicaiton studies is still noteworthy.

### Conclusion

Researchers should be aware that their experiments are likely to be underpowered if they plan their sample sizes using the effect size reported in a previous experiment. As a conservative heuristic for researchers performing formal sample size planning such as power analysis on the basis of previous research, researchers could follow the advice given in {Camerer, 2018 #967} and plan their experiments assuming that the true effect size is 50% of the reported effect size, a value matched by the more extreme 95% confidence interval of the estimated amount of effect size exaggeration across studies in this sample. Alternatively, it may be preferable to use methods of sample size planning that do not rely on precise a priori estimation of the effect size under study, such as planning studies to reliably detect the smallest effect size of interest {Lakens, 2018 #951}, using sequential analysis strategies {Pocock, 1977 #553;Lakens, 2014 #169}, or planning for adequate precision in parameter estimates across a range of possible effect sizes {Maxwell, 2008 #559;Kelley, 2017 #727}.

This research also emphasizes the importance of efforts to reduce publication and reporting biases, many of which individual researchers can voluntarily and easily take part in. Firstly, data-sharing platforms such as figshare (figshare.com) and the Open Science Framework (osf.io) make it possible for researchers to easily share the results of research whether or not a study is published in a traditional journal. Secondly, pre-prints allow researchers to report and publicize reports and data that may otherwise remain in the file draw, making it easier to ensure that non-significant results are accessible to other researchers and meta-analysts. Finally, projects like registered reports, in which papers are reviewed before data-collection on the basis of the design and analysis strategy as opposed to the results, also show promise in helping to develop a body of literature which is not affected by reporting and publication bias {Nosek, 2014 #202}. However, until large bodies of research free of publication bias become available, researchers should be aware that effect sizes in published studies are on average, considerably overstated.

### References

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., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433. Retrieved from <http://science.sciencemag.org/content/351/6280/1433.abstract>

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

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.

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

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:<https://doi.org/10.1016/j.jesp.2015.10.012>

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

Frank, M. C. (2017, February 2). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. <https://doi.org/10.31234/osf.io/27b43>.

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.

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

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2017). Accuracy in Parameter Estimation for a General Class of Effect Sizes: A Sequential Approach. *Psychological Methods*. doi:10.1037/met0000127

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol Methods*. doi:10.1037/met0000209

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

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., . . . Brumbaugh, C. (2014). Data from investigating variation in replicability: A “many labs” replication project. *Journal of Open Psychology Data, 2*(1).

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances In Methods and Practices in Psychological Science, 1*(4), 443-490. doi:10.1177/2515245918810225

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

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>

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537-563. doi:doi:10.1146/annurev.psych.59.103006.093735

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

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances In Methods and Practices in Psychological Science, 1*(4), 501-515. doi:10.1177/2515245918797607

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

Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253-1274. doi:10.1007/s10682-012-9555-5

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>

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

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191-199. doi:10.1093/biomet/64.2.191

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>

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*.

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

Thompson, B. (2002). What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes. *Educational Researcher, 31*(3), 25-32. doi:10.3102/0013189X031003025

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

## Supplementary material

## SM1

### Replication project Extraction and exclusion details

##### Open Science Collaboration (2015)

Three original studies which did not report that their findings were indicative of a non-zero effect were excluded from those studies extracted from Open Science Collaboration (2015). Three studies for which z transformed correlation coefficients could not be extracted due to missing data in the downloaded data set were also excluded from analysis. Effect sizes for original and replication studies are included for 94 out of 97 studies replicated studies from Open Science Collaboration (2015) which reported having found a non-zero effect.

##### Camerer et al. (2018)

Original and replication effect sizes were extracted for all 21 studies included in Camerer et al. (2018). In some cases in the Nature Science reproducibility project (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’ criterion than was originally reported in Camerer et al. (2018), where they used the p value from the largest performed study instead of a pooled estimate.

##### Soto (in press)

Effect sizes were extracted for original and replication studies for 101 out of 121 included studies, and one original study’s sample size was not available. In Soto (in press) Effect sizes which were only reported in this dataset as beta coefficients were not converted to Fisher z scores as not enough information was available in the data set. A total of 100 of 121 effects were included in the current analysis. As some replication studies used shorter form versions of the original data collection instruments, 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 the original and replication study effects were switched if the original effect was negative.

##### Cova et al. (2018)

Cova et al. (2018) included three replications of original studies which were non-significant (and which did not claim to provide evidence for the effects under test), these were removed from analysis. Effect sizes were reported by Cova et al. (2018) and are included in the current study for 33, original and replication studies, out of an original 37 replicated studies with significant original results. The four studies for which no effect sizes were reported performed analyses for which Cova et al. (2018) could not develop reasonable effect size estiamtes (e.g., a Sobel test, GEE analysis).

##### Many labs 1 R. Klein et al. (2014)

Many labs 1 (R. Klein et al., 2014) examined whether effects from 13 original papers replicated, one of which did not report an effect size or test statistic so is not included in the current sample. No effect size was extractable for one original study, and this effect was excluded for the purposes of the current analysis. Four different operationalisations of anchoring effects were tested, all of which are included in the current analysis, leading to a total of 15 paired data-points being included from this study. The multilevel models reported below accounts for non-independence between effects by including a random effect for study.

##### Many labs 2 R. A. Klein et al. (2018)

A total of 22 of 28 paired original and replication effects sizes are included for this analysis. Four studies from (R. A. Klein et al., 2018) were removed because 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 making it difficult to derive an appropriate effect size. Two additional studies were excluded because their effect sizes were only available as Cohen’s q.

##### Ebersole et al. (2016)

Original and replication effect sizes were extracted for all 9 original and replication studies from (Ebersole et al., 2016), excluding a study they term a “conceptual replication”. Most effects (6/9) were converted to correlation coefficients from the Cohen’s d values reported in this replication project. The results of three additional studies reported as partial Eta squared were converted to correlation coefficients from F statistics using the formula:

##### Camerer et al. (2016)

The economics replication project (Camerer et al., 2016). Original and replication effect sizes for all 18 studies were reported in correlation coefficients and all are included in this analysis.

### SM2

### 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 1. 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 | 0.04 | < .001 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.016, n = 229 |
|  |  |  |  |  | Effect variance = 0.012, n = 306 |
|  |  |  |  |  | QE(305) = 3531.9, p < .001 |

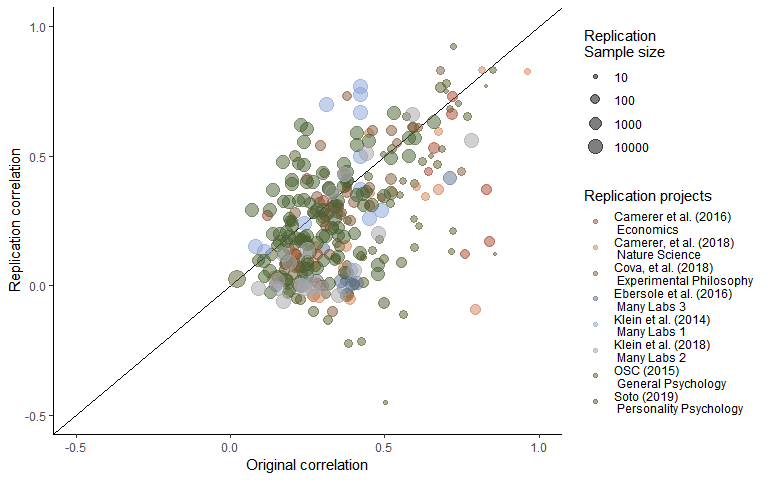


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

Table SM2. 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 | -0.11 | 0.01 | 0.03 | 0.1 |  |
|  |  |  |  |  | Project variance = 0.005, n = 8 |
|  |  |  |  |  | Article variance = 0.014, n = 132 |
|  |  |  |  |  | Effect variance = 0.009, n = 198 |
|  |  |  |  |  | QE(197) = 2715.24, p < .001 |

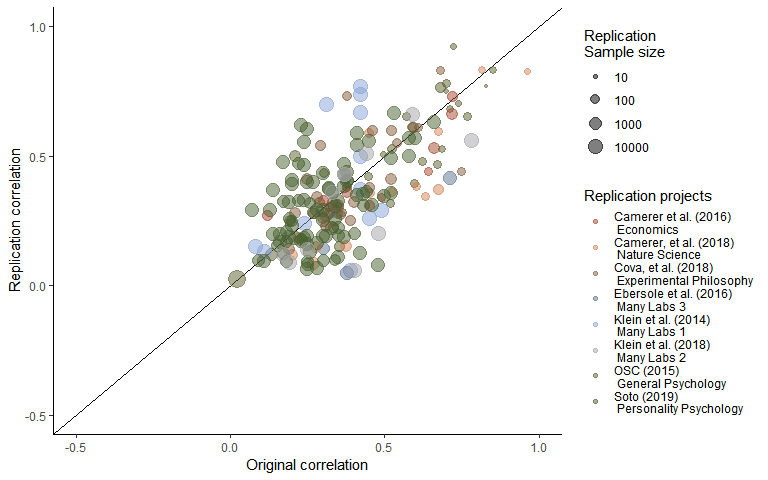


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

Table SM3. 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 | -0.15 | -0.01 | 0.04 | 0.03 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.018, n = 169 |
|  |  |  |  |  | Effect variance = 0.009, n = 237 |
|  |  |  |  |  | QE(236) = 3031.58, p < .001 |

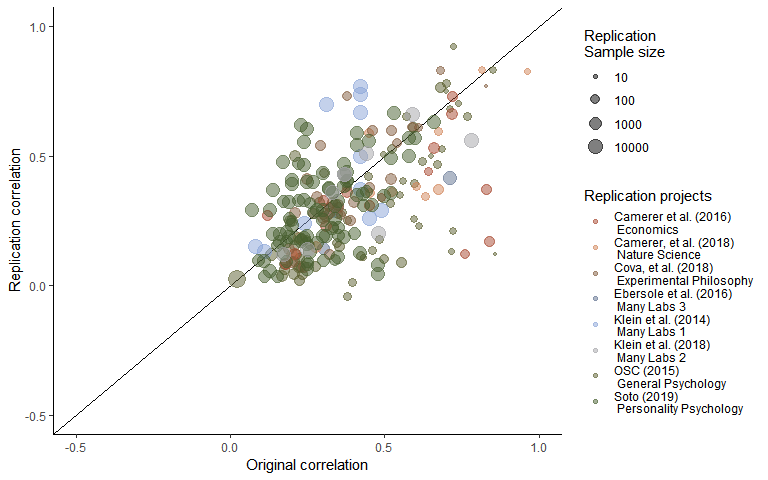


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original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

### SM3

#### LOO Cross validation output

#### Table [LOO cross validation output](file:///C:\Users\fsingletonthorn\Documents\PhD\Effect%20size%20adjustment%20testing%20paper\SimplifiedEffectSizeAdjustment%2029%20Jan%202018.docx#loo-cross-validation-output).

0th, 25th, 50th, 75th and 100th percentiles from leave one out cross validation for each multilevel model, for each exclusion method an, including only the sample indicated in “LOO exclusions”.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LOO exclusions | Subsample | Proportion significant | Minimum estimate | 25th percentile | Median | 75th percentile | Maximum estimate |
| Replication project | Only non-equivalent replications | 0.38 | -0.10 | -0.09 | -0.08 | -0.07 | -0.06 |
| Replication project | Only significant replications | 0.12 | -0.07 | -0.06 | -0.06 | -0.04 | -0.03 |
| Replication project | P value as Moderator | 1.00 | -0.10 | -0.09 | -0.07 | -0.07 | -0.06 |
| Replication project | All data | 1.00 | -0.16 | -0.15 | -0.13 | -0.13 | -0.12 |
| Study | All data | 1.00 | -0.14 | -0.14 | -0.14 | -0.14 | -0.13 |
| Study | Only significant replications | 0.01 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 |
| Study | Only non-equivalent replications | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Study | P value as Moderator | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | Only significant replications | 0.00 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 |
| Effect | Only non-equivalent replications | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | P value as Moderator | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.07 |
| Effect | All data | 1.00 | -0.14 | -0.14 | -0.14 | -0.14 | -0.13 |

### 

### SM4

#### 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]  
 }  
}