EstimatingPublicationBias

## Introduction

See word document

## 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 or when results were only reported in Cohen’s q values were excluded). 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 {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 approximation is ab approximation for certain effect sizes, such as those converted from F tests with a greater than 1 or chi square tests. All multilevel meta-analyses were re-performed excluding these studies and no differences in substantive interpretation would result (i.e., none of the beta coefficients changed by more than 0.02, no statistical tests changed their statistical significance at the .05 level, and variance partitioning did not change enough to alter their interpretations), the results reported below therefore use this approximation. 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 attenuation could be extremely high, and yet this effect be almost entirely driven by the presence of effectively-null effects. In order to account for this issue, three main 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 significant and (b) 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 decreases after accounting for small or near-null effects.

Method (a), 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.

Method (b) was used to avoid excluding under-powered studies erroneously. Method (b) excludes studies based on whether the results are statistically equivalant to the null, or statistically significant in the opposite direction {Lakens, 2017 #214;Lakens, 2018 #951}. As a requirement for equivalence testing is that an equivilance bound is selected, 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. 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 effects the original authors may have considered important. See supplementary materials 2 for scatter plots of the data-set using each exclusion rule.

##### Analysis 4: 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 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 , 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 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, 2018 #967}), 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 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

#### Raw changes in effect sizes

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 articles. The average effect size for original studies was r = 0.39, and the mean effect size for replication studies was r = 0.27. There was an mean decrease of r = -0.13. 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.

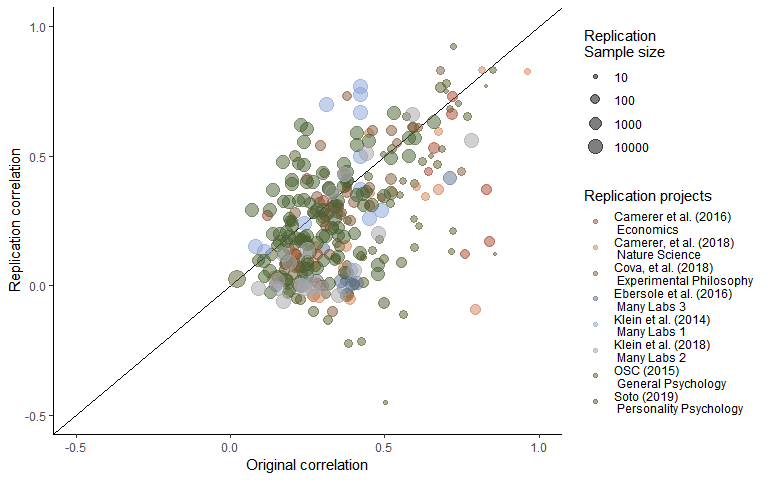


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of participants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

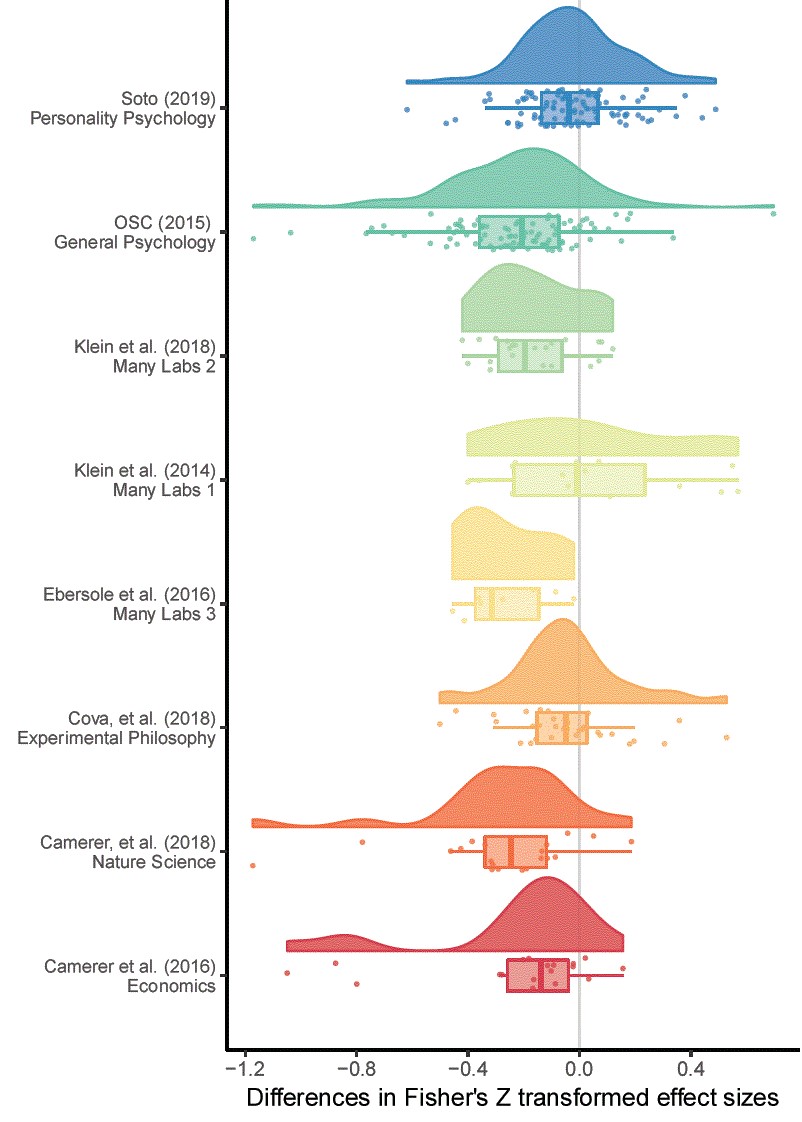


Figure 2. A raincloud plot of the change in effect sizes (here Fisher Z scores) from the original to the replication study by the replication project that each replication study was performed as a part of.

##### Descriptives of samples after exclusions

Examining 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. There was an average decrease of r = -0.02 (the slight discrepancy between this value and those above is because these values are being back transformed from Fisher Z statistics), 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 = -0.07 (again the slight discrepancy between this value and those above is because these values are being back transformed from Fisher Z statistics), a mean percentage decrease of 7%, and a median decrease of 17%.

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 | Non-equivalent |
| 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 estimates 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 is 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. QE tests of heterogeneity suggest that there is a large amount of unexplained heterogeneity, QE(305) = 3531.9, p < .001, and an of 92.585 (calculated following {Nakagawa, 2012 #1023}), suggesting that 93% of variance in effect sizes is due to heterogeneity (i.e., variance in the true effect size differences), while the remaining 7% is attributable to sampling variance.

Empirical Bayes estiamtes for the projects show that there is not much

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 |

Table [BLUP]. Empirical Bayes estimates and 95% credible intervals for random effects (i.e., estimates of the difference between the replication project’s mean effect size difference and the overall estimated mean effect size difference). These values 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 | -0.059 | 0.051 | -0.160 | 0.041 |
| Camerer, et al. (2018) Nature Science | -0.075 | 0.052 | -0.177 | 0.027 |
| Cova, et al. (2018) Experimental Philosophy | 0.079 | 0.046 | -0.011 | 0.169 |
| Ebersole et al. (2016) Many Labs 3 | -0.064 | 0.060 | -0.181 | 0.052 |
| Klein et al. (2014) Many Labs 1 | 0.118 | 0.054 | 0.013 | 0.223 |
| Klein et al. (2018) Many Labs 2 | -0.025 | 0.048 | -0.119 | 0.068 |
| OSC (2015) General Psychology | -0.051 | 0.041 | -0.132 | 0.030 |
| Soto (2019) Personality Psychology | 0.078 | 0.042 | -0.004 | 0.160 |

#### 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 is 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 predicts an 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). Looking at the range of the 95% confidence intervals for both of these subsamples, the coverage extends from a considerable decrease equivalent to -38% of the average original correlation, to a small increase in effect sizes 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).

|  |  |  |  |
| --- | --- | --- | --- |
|  | All studies | Statistically significant | Non-equivalent |
| 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 |

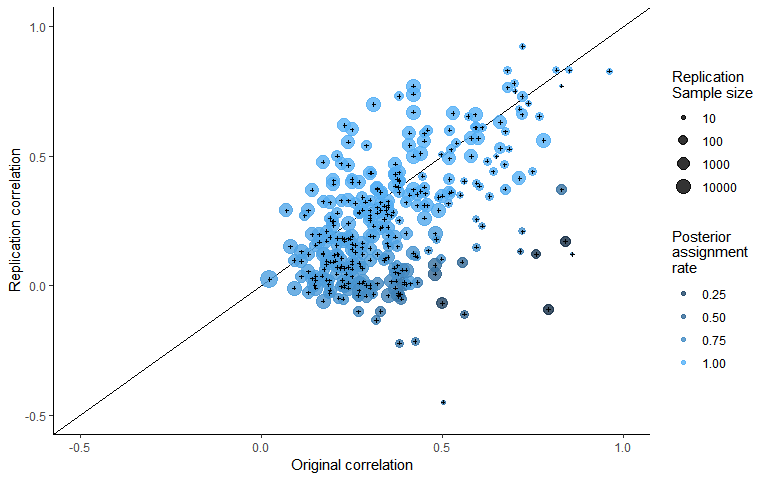
##### Leave one out cross validation of meta-analyses

To assess how sensitive the results of the multilevel models are 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 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. 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). In part, the high posterior assignment rate may be due to the fact that this model allows for “true” effect sizes to be estimated as being extremely low or near zero and 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. Across the whole sample, replication sample sizes fell by a mean of r = -0.13, and were 28% smaller in replication 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 [-0.42, -0.14]), or an estimated decrease of 34% (95% CI [51%, 17%]) of the mean effect size in the original studies (a equal to r = 0.39). All of these estimates are higher than that suggested by {Stanley, 2018 #1002} of 8 to 15%, possibly because of the different samples used (i.e., studies that were the focus of meta-analyses in Stanley et al. compared the various sampling strategies used in the large scale replication projects).

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 [-28%. The average effect size differences 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 -0.02 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. All models estimated using data-exclusions show a lower effect size decrease when attempting to exclude null (or effectively null) results, the confidence intervals over both of these models extend from a decrease of -41.96% of the average original correlation coefficient, to an increase of 2.71%.

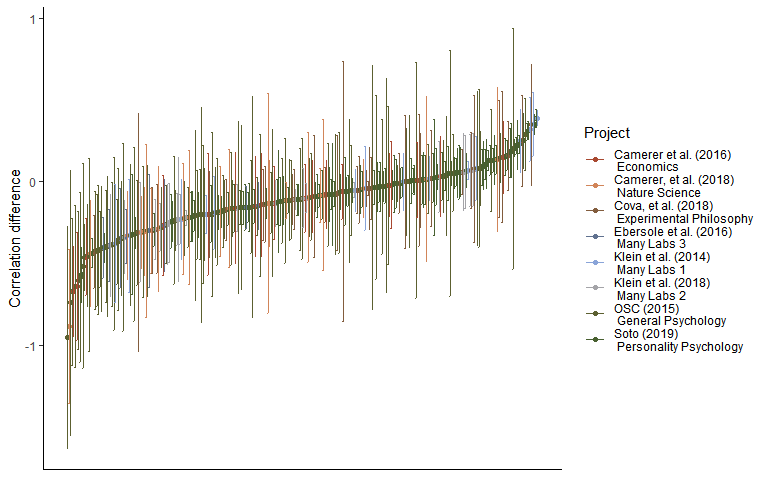


Figure x. A caterpillar plot of the effect size difference between original and replication study effect sizes ordered by magnitude, error bars are 95% confidence intervals around effect size differences.

#### 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 than 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 may still be 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 type of effects.

The Bayesian mixture model used here has two major issues. Firstly, it assumes independence between effects and 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. Future reserach could help develop a more nuanced account of the data-generation process that leads to this data-set by for example building a model that allows for the attenuation rate to change across replication studies, possibly includoing 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 research.

### 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. Changes of this magnitude could make a considerable difference in many research scenarios, and researchers reading the literature should be aware that the effect sizes reported in the scientific literature are likely exaggerated. Researchers should also 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. As a 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 size, 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 {Schönbrodt, 2017 #672;Albers, 2018 #942;Lakens, 2014 #169} when practical constraints mean this approach is feasible.

This research also 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, 2016 #475}. 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. Additionally, 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, 2014 #202}.

## Supplementary material

## SM1

### Replication project Extraction and exclusion details

##### {Open Science Collaboration, 2015 #611}

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 #611}. 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 #611} which reported having found a non-zero effect.

##### {Camerer, 2018 #967}

Original and replication effect sizes were extracted for all 21 studies included in {Camerer, 2018 #967}. In some cases in the Nature Science reproducibility project {Camerer, 2018 #967} 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 #796} 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, 2018 #967}, where they used the p value from the largest performed study instead of a pooled estimate.

##### Soto et al., (2019)

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 [LOOPR study CITATION] 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, 2018 #984}

{Cova, 2018 #984} 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 {Klein, 2014 #988}

Many labs 1 {Klein, 2014 #988} 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 {Klein, 2018 #1021}

A total of 22 of 28 paired original and replication effects sizes are included for this analysis. Four studies from {Klein, 2018 #1021} 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, 2016 #985}

Original and replication effect sizes were extracted for all 9 original and replication studies from {Ebersole, 2016 #985}, 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, 2016 #983}

The economics replication project {Camerer, 2016 #983}. 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 |

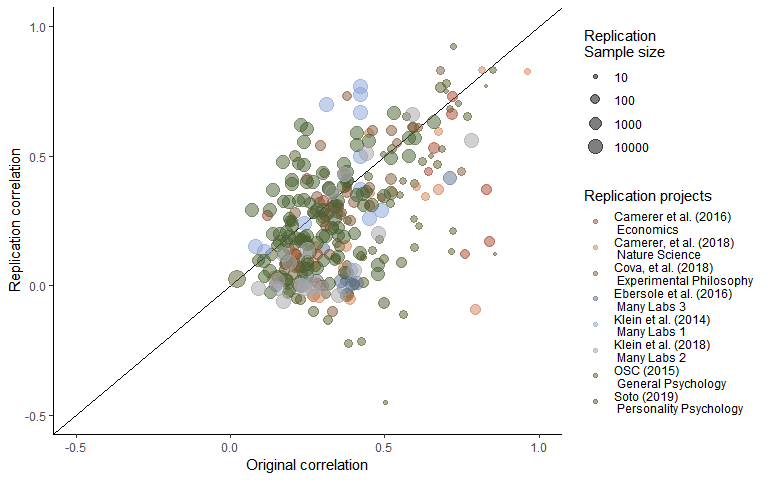


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

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 |

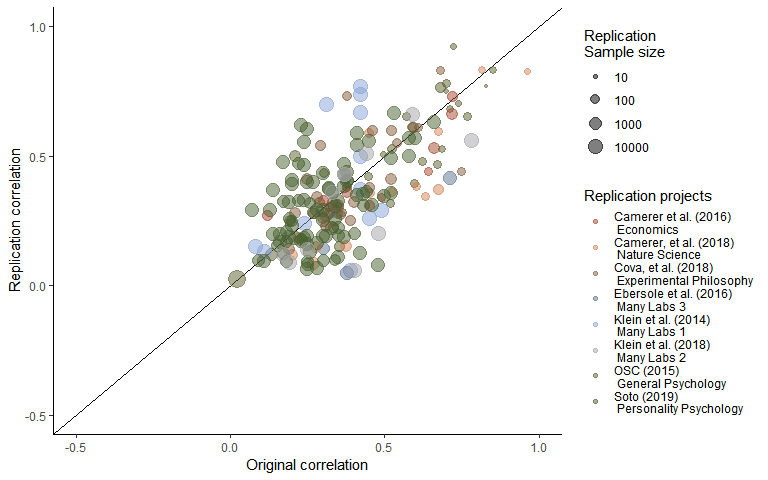


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

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 |

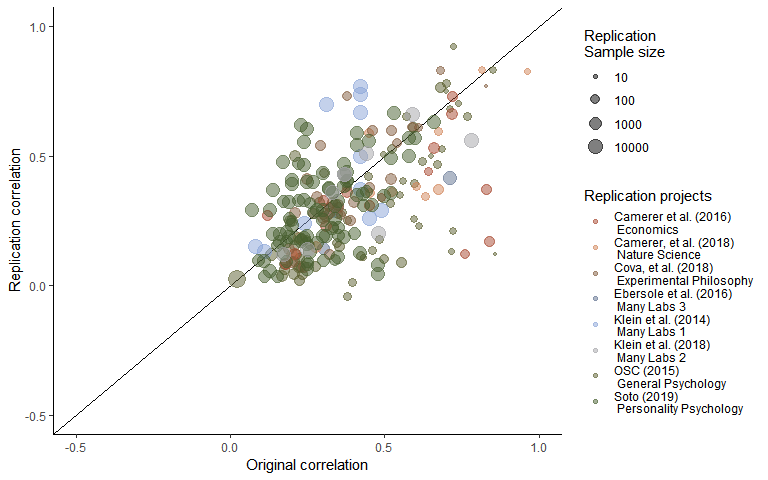


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original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

### SM3

#### LOO Cross validation output

#### Table [LOO cross validation output](#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

#### Logit transformed p value multilevel model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Estimate | -0.26 | -0.32 | -0.21 | 0.03 | < .001 |  |
| p value | -0.06 | -0.07 | -0.04 | 0.01 | < .001 |  |
|  |  |  |  |  |  | Project variance = 0.004, n = 8 |
|  |  |  |  |  |  | Article variance = 0.013, n = 229 |
|  |  |  |  |  |  | Effect variance = 0.013, n = 306 |
|  |  |  |  |  |  | QE(305) = 3058.54, p < .001 |

Predicted output at examining the estimated amount of effect size decrease at a replicaiton p value of the logit of the the smallest positive floating-point number possible using 32-bit integers and use IEC 60559 floating-point (double precision) arithmetic (2.220446e-16):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | pred | se | ci.lb | ci.ub | cr.lb | cr.ub |
|  | -0.06 | 0.03 | -0.11 | 0 | -0.38 | 0.26 |

### SM5

#### 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, 2017 #1012}, 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, 2018 #967} 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]  
 }  
}