EstimatingPublicationBias

## Introduction

See manuscript

## 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). 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), see supplementary materials 5 for details regarding these conversions. This conversion used data from the replication project whenever possible (i.e., whenever effect sizes were reported in correlation coefficients in a summary table or in a project’s online data this was directly converted to Fisher z values). If the study-level results were not reported as correlation coefficients, Cohen’s d values, as t-tests, or as F statistics in the original or replication project we excluded the result from this 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. For each of the Many Labs projects the top level result was used (i.e., the results of the analysis which collapased the data across the multiple labs). 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. 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 pairs of studies, excluding a total of 41 study pairs.

### Analysis

All analyses were 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, however all reported analyses should be considered exploratory. All of the data and analysis code used in this study and an RMarkdown document to allow the current document to be easily reproduced are available from <https://osf.io/7qvna/>.

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

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

This analysis treats each pair of effects, the original and replicated effect sizes, as one “study” in a meta-analytic framework. This model estimates the change from original to replication study effect sizes () with a fixed intercept (), a random effect for replication project (), a random effect for each original article (), and a random effect for each individual replication (). Random effects at the project level are included to account for non-independence between replications from each replication project. Random effects at the original article level are included to account for cases when multiple effects from an original article were replicated or multiple operationalisations of an original effect were tested. 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., the intercept coefficient and random effects varaince estimates changed by less than 0.003).

Using the aggregate summary statistics from the replication projects where a set of labs conducted replications (e.g., the Many Labs Projects) may underestimate the standard error of the difference scores (as their standard error is also a function of the unknown heterogeneity across labs). As a sensitivity analysis, we also ran all multilevel models using a conservative estimate of their sampling variance - calculating their standard errors using the mean sample size included in each replication study as opposed to the total sample size. Again, no differences in the substantive interpretation of results would follow from this change, with the coefficient estimates and estimates of the variance of the random effects changing by less than 0.005.

#### 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 replication studies where the null hypothesis is true. The average effect size difference between original studies and their replications could be extremely high, and yet this effect could be entirely driven by the presence of null effects. 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 in the original article, the average attenuation would be 50%. Analyses 2 to 5 were performed in order to account for this issue.

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

Analyses 2 and 3 reperform the above meta-analysis excluding studies using two exclusion criteria. Analysis 2 excludes studies in which the replication study was not statistically significant with an effect in the same direction as the original (using the *p* value reported in the replication projects’ datasets, at an alpha of .05, and using two-tailed tests where applicable). Analysis 3 removes effects in which the replication study effect is “statistically equivalent” to the null according to an equivalence test.

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 to 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}. A requirement for equivalence testing is that an equivalence bound is selected (i.e., an effect size below which the effect size is said to be for all practical purposes equal to zero). For this, we use the lowest effect size that would have been statistically significant in the original study (assuming an alpha of .05), following a suggestion in {Lakens, 2018 #951}. 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 1 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 dataset using each exclusion rule.

##### Analysis 4: Bayesian mixture model

Analyses 2 and 3 both rely on excluding studies using exclusion rules that will, respectively, exclude or retain studies due to low statistical power in the replication study. In part in order to avoid this issue the final approach to estimating the amount of effect size attenuation conditional on the effect under study being 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 true replication effect size is drawn from the null-hypothesis component, it is assumed to be drawn from a normal distribution with a mean of 0, and the observed replication effect size is assumed to be drawn from a normal distribution centered on the true replication effect size with 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 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 4 for model syntax and further analysis details.

### Descriptives

Looking at the 306 included original-replication study pairs included in this analysis, 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 in this sample and Figure 1 for a raincloud plot of the Fisher Z score change in effect sizes by replication project.

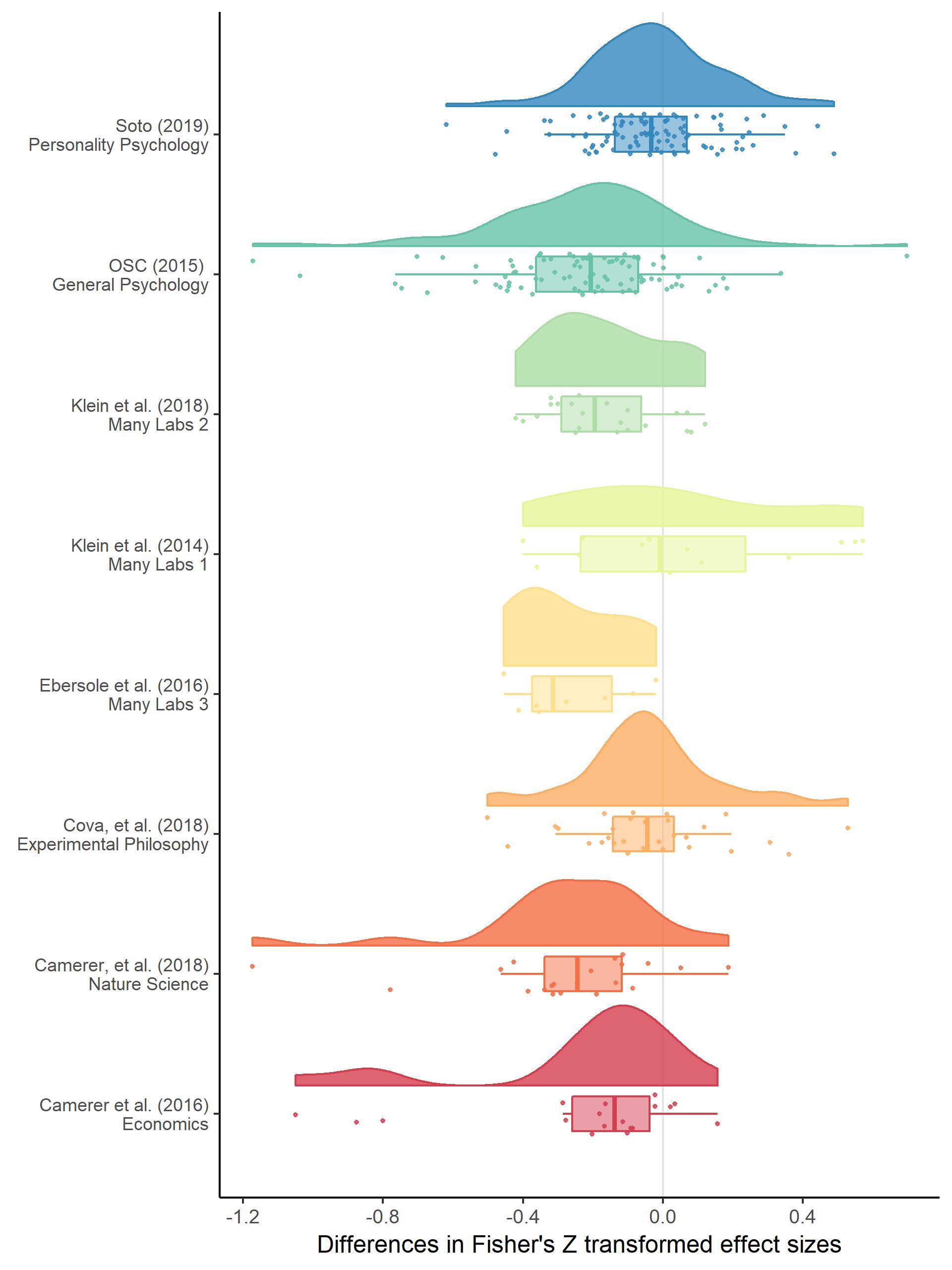


Figure 1. A raincloud plot (density, box and scatter plot) of the change in effect sizes (here Fisher Z scores) from the original to the replication study by the replication project that each replication study was performed as a part of.

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 ( = 0.128, = 0.11, compared to = 0.088), 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 3. 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

Examining just the 198 cases in which the replication study was statistically significant (65% of all studies), the average effect for original studies was 0.404, and the mean effect size for replication studies was 0.387. This represents a mean decrease of r = 0.017, a mean percentage increase in effect sizes of 4% and a median percentage decrease of 7%. Using equivalence testing 77% of replication studies were not statistically equivalent to the null (n= 237). The average effect size in the original non-equivalent studies was 0.406, compared to a mean effect size for replication studies of r = 0.348. This is a mean decrease of r = 0.058, a mean percentage decrease of 7%, and a median percentage decrease of 17%.

Reperforming the meta-analysis only including studies for which the replication was statistically significant and had an effect in same direction as the original produced an estimated r = -0.051 (95% CI [-0.111, 0.01]) change in effect sizes from original to replication studies. Including only the studies which were not statistically equivalent to the null leads to a predicted r = -0.082 (95% CI [-0.154, -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 = 0.387). However, there was considerable imprecision in these estimates, 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 = 0.366).

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

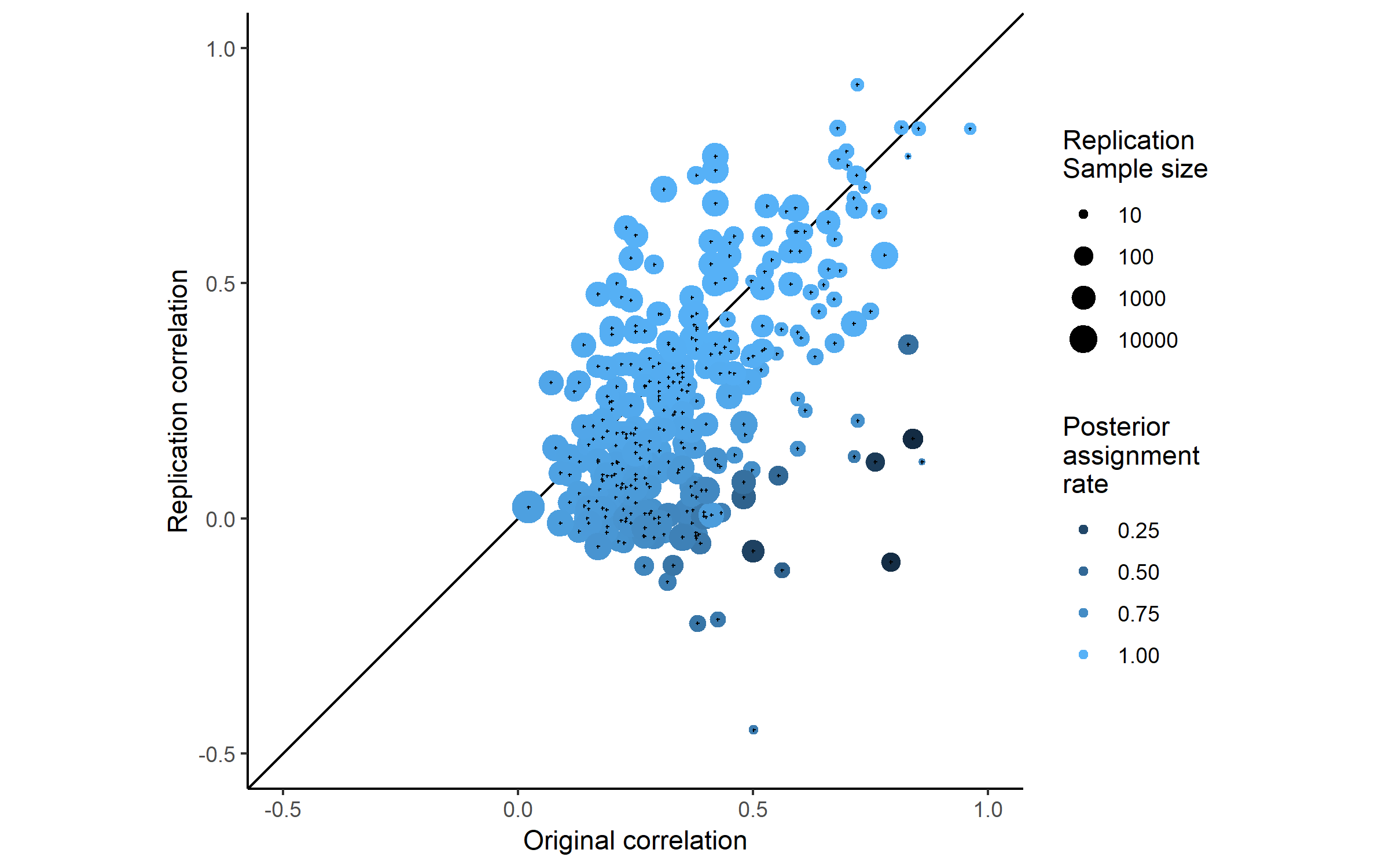
##### Leave one out cross validation of meta-analyses

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 3 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 2 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 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), likely 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 2.

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 null 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 Fisher Z 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 in their estimates differs. The Bayesian mixture model suggests 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 give less precise estimates, highlighting the uncertainty in these estimates. For example, the confidence intervals over both of the models estimated excluding data extend from a meaningful decrease of r = -0.15 to a slight increase of r = 0.01.

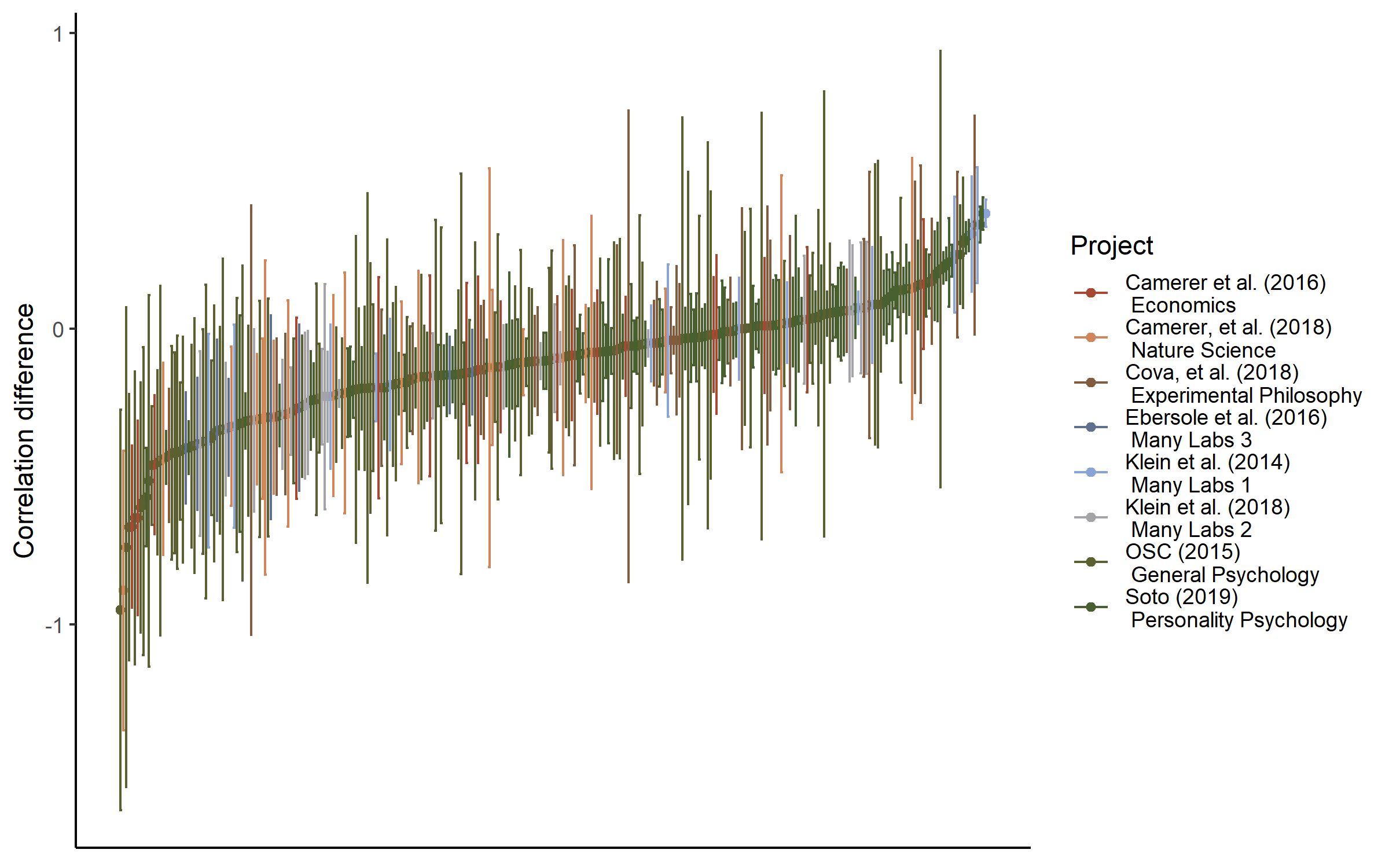


Figure 3. A caterpillar plot of the effect size difference between original and replication study effect sizes ordered by magnitude, error bars are 95% confidence intervals around effect size differences.

In using the results of this analysis to inform future research (e.g., in sample size planning) and to interpret the literature, it is essential to take note of the level of heterogeneity in the amount of effect size attenuation across not just replicated effects but also across replication projects. The sampling decisions and replication methods used by each of the included projects appears to have had a considerable effect on the amount of effect size attenuation seen (e.g., the estimated standard deviation of the mean project effect size attenuation from analysis 1 is 0.13, 95% CI [0.04, 0.18] in analysis 1). The degree of heterogeneity in the amount of effect size attenuation across studies and projects means that using any single estimate of the amount of effect size decrease is likely to be misleading in the case of any individual replication study.

#### Limitations and future directions

In interpreting these results it is important to note several limitations. Firstly, the current study does not attempt to distinguish between effect size heterogeneity (i.e., effect sizes that are different 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 replication 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. It also cannot be ruled out on the basis of this data that the effect size differences seen in these large-scale replication projects are larger than would be seen for individuals attempting to replicate particular effects (e.g., if researchers in these large-scale replications have less access to the tacit knowledge that would normally facilitate replicators efforts).

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 data-generation process underlying this dataset 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 replication 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 whereas Camerer et al., 2018 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. As stated in the introduction, this analysis therefore should be seen as producing an estimate of the effect size differences we would expect to see in future large-scale replication projects, under the assumption the effects of the different sampling decisions used by these studies lead to normally distributed effects around some global mean level of effect size attenuation. While these results should be considered preliminary, this analysis nonetheless provides an initial estimate of the mean effect size attenuation that should be expected in replication studies, and suggests that even after accounting for the presence of null effects the amount of effect size attenuation between the published literature and replication studies is still noteworthy.

### Conclusion

The findings of this study reinforce the importance of recent efforts to reduce psychology literature’s reliance on underpowered original research designs, to circumvent publication bias, and to avoid QRPs like p-hacking and HARKing {Bakker, 2012 #38}. Efforts to avoid the impact of any of these issues would likely reduce the degree to which effect sizes are attenuated in replications of the primary research literature. There are several recent efforts to reduce the impact of publication and reporting biases that readers should be aware of, many of which individual researchers can voluntarily and easily take part in.

In order to avoid performing future underpowered reserach, 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 a 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}, or planning for adequate precision in parameter estimates across a range of possible effect sizes {Maxwell, 2008 #559;Kelley, 2017 #727}. Recent large scale multinational data collection efforts like the Many labs Projects or the Psychological Science Accelerator {Moshontz, 2018 #1025} also help to avoid the negative impacts of low statistical power by allowing for extremely high powered studies of even very small effects.

Careful preregistration of analysis plans offers one method that reserachers can use to avoid biases in their data-analysis which may otherwise lead to inflated effect sizes {Wicherts, 2016 #475}. 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. Similarly, pre-prints (e.g., <https://psyarxiv.com>) allow researchers to report and publicize reports and data that may otherwise remain in the file draw. Both preprints and data repositories make it easier to ensure that non-significant results are accessible to other researchers and meta-analysts. Finally, registered reports, in which papers are reviewed before data-collection on the basis of the research 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 likely to be considerably overstated.

## 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 (these included 1 study which used multiple statistical tests in the original and replication studies, one study in which the replication and original study used different statistical tests, and one study for which the effect sizes were reported as beta coefficients wihtout test statistics or degrees of freedom). 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, in press #1032}

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 #1032} 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 to do so. 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 the 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 inverted for analysis 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 estimates (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 |

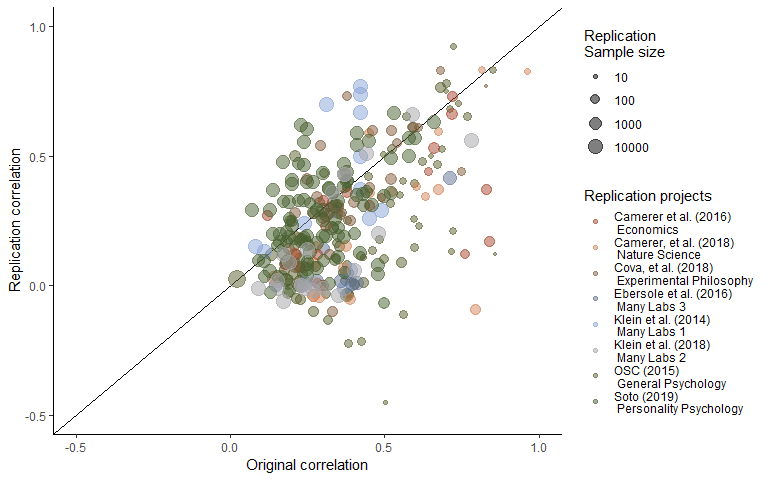


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

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 |

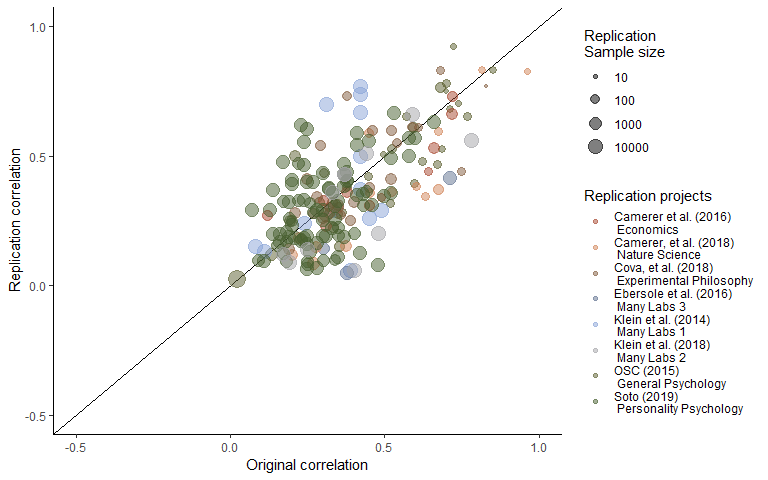


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

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 |

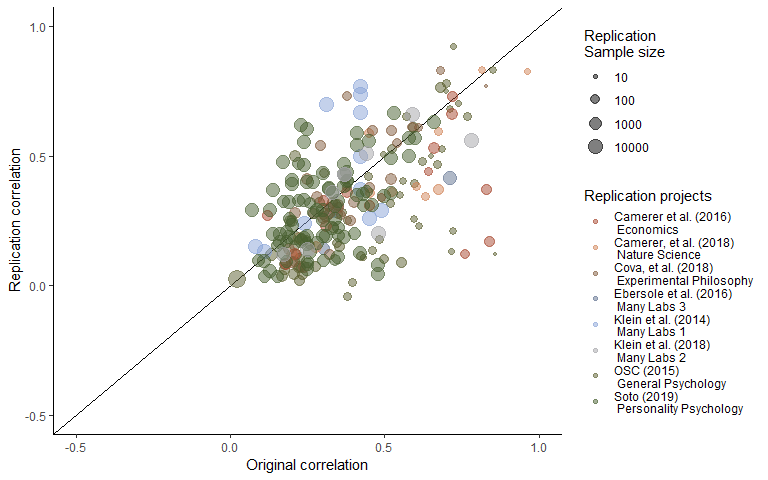


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original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

#### 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]  
 }  
}

### SM5

#### Conversions

All statistical tests extracted were transformed into correlation coefficients as follows, using the methods reported in {Open Science Collaboration, 2015 #611}.

t statistics:

Where is the observed t statistic and is the degrees of freedom of the t test.

F statistics:

Where is the observed F statistic and is the degrees of freedom of the numerator and is degrees of freedom of the denominator.

Chi square statistics:

Where is the observed statistic and is the associated degrees of freedom.

All values were then transformed into fisher Z transformed correlation coefficients using: