Estimating the effect of publication and reporting bias

## Introduction

See word document

## Methods

### Data extraction

All of the large scale replication projects that have been performed in behavioral science research were collected. The original source of each study, test statistics, effect sizes, sample sizes, standard errors, p-values were extracted for each original and replication study. Several of the large scale replication projects did not present the original test statistics and p values (e.g., Many labs 1 and 3). In these cases, these values were manually extracted from the original articles. When sample sizes for original studies were not available they were manually extracted from original articles. When the original and replication effect sizes were not reported as Fisher Z transformed correlation coefficients, effect sizes were converted from test statistics or effect sizes for analysis. In cases where sample sizes were not reported per group, equal sample sizes among groups were assumed to be equal in these estimates. See table one for the number of valid studies extracted from each project. All results are reported in correlation coefficients following {Open Science Collaboration, 2015 #611} in order to present results in a common metric which is likely intuitively understandable and familiar to most psychologists and behavioral researchers.

Three studies which did not report that their findings were indicative of a true effect were excluded from {Open Science Collaboration, 2015 #611}. For the Nature Science reproducibility projects {Camerer, 2018 #967}, when multiple replication studies were run, a fixed effects meta-analysis was performed using the metafor package {Viechtbauer, 2010 #796} for each study to estimate the true effect. P values, standard errors and sample sizes reflect this pooled estimate. This method leads to one study more “replicating” according to the ‘statistical significance in the same direction of the original study’ than was originally reported in the nature science project, where they using the largest performed study instead of a pooled estimate.

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

INSERT TABLE 1 HERE

### Analysis

All analysis was performed in R {R Development Core Team, 2018 #314}. Mean raw differences along with Wald-type 95% confidence intervals around the mean difference, median effect size differences, and raw proportion decreases in effect sizes (i.e., ) were calulcated on the Fisher-Z transformed effect sizes. The reported Wald-type confidence intervals do not account for non-independece between effects taken from the same paper, or between studies from the same replicaiton projects. In order to account for this non-independence, multilevel-meta-analysis framework was used. Any studies with missing data (e.g., missing effect sizes or sample sizes for the initial or replicaiton studies) were excluded, and sample sizes are reported alongside each analysis in tables.

#### Multilevel meta-analysis

Meta-analyses were performed using the Metafor package {Viechtbauer, 2010 #796}. In order to obtain a reasonable estimate of the change in effect size between original and replication studies, a multilevel random effects meta-analysis was performed on the difference in Fisher Z transformed correlations between original and replication studies. Standard errors were estimated as , with being the sample size in the original study and being the sample size in the replication study. Empirical Bayes estimates and 95% credible intervals for random effects were calculated following {Robinson, 1991 #999}{Morris, 1983 #1000}.

Confidence intervals around binomial proportions are 95% Wilson Score intervals. All difference scores (i.e., proportion changes and mean differences) were calculated using Fisher Z transformed effect sizes. 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 a preregistration of this project.

##### Leave one out cross validation

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

#### Accounting for null effects

An important question in assessing the degree to which effects are attenuated in this literature is how much this effect is driven by the presence of null effects (or effects so small as to be effectively null). The average disattenuation could be extremely high, and yet this effect be almost entirely driven by the presence of effectively-null effects. This aspect becomes especially important as the sampling of the literature is non-random, meaning it is plausible that some effects were chosen for replication to a greater or lesser extent as it was expected that they may not replicate. In order to account for this issue, the average effect size attenuation was calculated and multilevel models were estimated exluding original studies based on multiple excliusion rules; using the statistical significance of the replication study, equivalence testing, and approximate Bayes Factors. Because all of these methods function by removing small or near-null effects, no significance testing was performed on the difference between the model estimates estimated decreases after accounting for small or near-null effects. It is certain that, at a population level, all of these actions would lower the size of the observed effect size decrease.

##### Statistical significance of the replication study

The first method used to attempt to exclude likely null effects is to only look at effects that reached statistical significance in the replication study in the same direction as the original effect. This has the issue of meaning that studies which were under-powered to detect a non-null but true effect are likely to be excluded from this analysis. Especially as in some of the replication projects the sample size in the second study was chosen using a power analysis of the observed effect in the original study {Open Science Collaboration, 2015 #611}, this method is likely to underestimate the amount of effect size exaggeration due to the exclusion of underpowered replications. Original studies which found large effects lead to follow up studies which have smaller sample sizes, and are therefore unlikely to reach statistical significance given a true, non-zero, but smaller effect size.

##### Equivilence tests

A second method we use is to exclude studies from estimates of the amount of effect size decrease based on whether the results of the replication study were statistically equivalent to the null {Lakens, 2017 #214;Lakens, 2018 #951}, or significant in the opposite direction. As a requirement for equivalence testing is that a minimum effect size of interest is selected, we follow one suggestion in {Lakens, 2018 #951} and use the lowest effect size that would be statistically significant to the original study as the smallest effect of interest (assuming an alpha of .05). Equivalence tests were performed used the Fisher Z transformed effect sizes, and approximated the standard errors of each study as , except for studies from {Camerer, 2018 #967} which had more than a single replication attempts, where standard errors are those derived from the meta-analyses that produced the effect size estimate. Equivalence tests were performed using z tests, i.e., assuming a normal sampling distribution. Ideally, a full reanalysis would be performed for each original study using the original statistical test and full access to the original and replicaiton data. However, it was not feasible to extract and reperform full analyses for the over 600 total original and replication studies. As a method of testing how closely this method of approximating standard errors matches the original replication projects results, significance tests for the replication and original studies were performed using this approximation. The results matched the significance or non-significance as reported in the replication projects in every case.

However, the minimum detectable effect was occasionally quite high as original sample sizes were often very small (mean = 0.17, SD = 0.12, 0th, 25th, 50th, 75th and 100th quintiles = [0, 0.1, 0.15, 0.23, 0.74]). This means that original original studies were sometimes under-powered to detect even large effects, meaning that this method may exclude studies which have effects the original authors may have considered important, but either would have been able to detect, or have used an experimental design and statistical tests that were more sensative than the this analysis suggests.

##### Approximate bayes factors

Three different types of Bayes factors were developed for each study using default priors following {Wagenmakers, 2016 #994}. Bayes Factors express the relative evidence for the null hypothesis compared to an alternative model, or equivalently the degree to which a Bayesian observer should update their prior beliefs in response to the receipt of new data in favour of one model or another. If a Bayes factor is greater than one the data is more likely under the alternative hypothesis than under the null hypothesis, and the opposite is true when a Bayes factor is below one. Conventional labels have been proposed, suggesting that Bayes factors between 1 and 3 provide little to no evidence (or ‘anecdotal’ evidence) and Bayes factors from 3-10 provide “substantial” evidence {Jeﬀreys, 1961 #1001}{Wagenmakers, 2016 #994}.

Two of the developed Bayes Factors ignore the original study and express the relative evidence for and against the point null entirely based on results of the replication study, using a one () and and two tailed () default alternative hypothesis (for details see {Wagenmakers, 2016 #994}). Replication Bayes Factors () were also developed, in which the prior for the replication correlation coefficient is the posterior based on the original research and a flat prior, for details see {Wagenmakers, 2016 #994} and {Verhagen, 2014 #217}. This paper follows the typical notation where the order of the subscripts indicate whether a Bayes Factor represent evidence for the null (, , ) or for the alternative hypothesis (, , ).

The bayes factors presented here were developed using the effect sizes as reported in correlation coefficients, regardless of the original effect size measure and experimental design. Importantly, these Bayes factors differ from those that would normally be developed using the closest Bayesian equivalents to each original replicated study’s analysis, and should be viewed only as a coarse estimate of the degree of evidence provided for and against the null model. See table [bayesFactors] in supplementary materials [Bayes] for a table showing the differences between the values returned by this method compared to those reported in the Bayesian supplement to which were more appropriately calculated {Camerer, 2018 #967}, which demonstrates that the difference can be considerable when the original analysis was unusual. Normally, One of the benefits of Bayes Factors is the continueous and interpretable scale, however in this case these approximate Bayes Factors are used as a heuristic to discard the studies which appear to likely be true (or effectively) null effects. Two different cut scores were used for each type of Bayes factor, discarding studies when Bayes factors suggested that the null model is either more than three times more likely than the alternative model (i.e., when there is more than ‘anecdotal’ evidece that the null is true), or when the alternative model is not at least three times more likely than the null model.

#### Simulations to assess exclusion criteria

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

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

## Results

### Decrease in effect sizes

Looking at the 314 replications for which both original and replication effect sizes were available, the effect size seen in the replication study fell in 227 articles, (72%) . The average effect size for original studies was 0.38, and the mean effect size for replication studies was 0.27. There was an average decrease of r = -0.13 (Wald-type 95% CI [-0.16, -0.1]). Notably, this represents an average decrease in effect sizes from the original to the replication study of -29.44%. See Table 2 for a more comprehensive list of descriptives on the effect size differences seen, and figure 1 for a scatterplot of the replication effect sizes plotted against the original studies’.

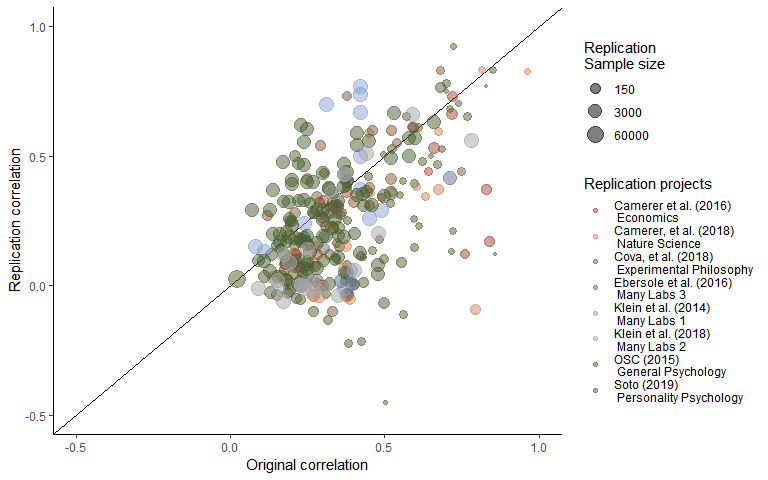
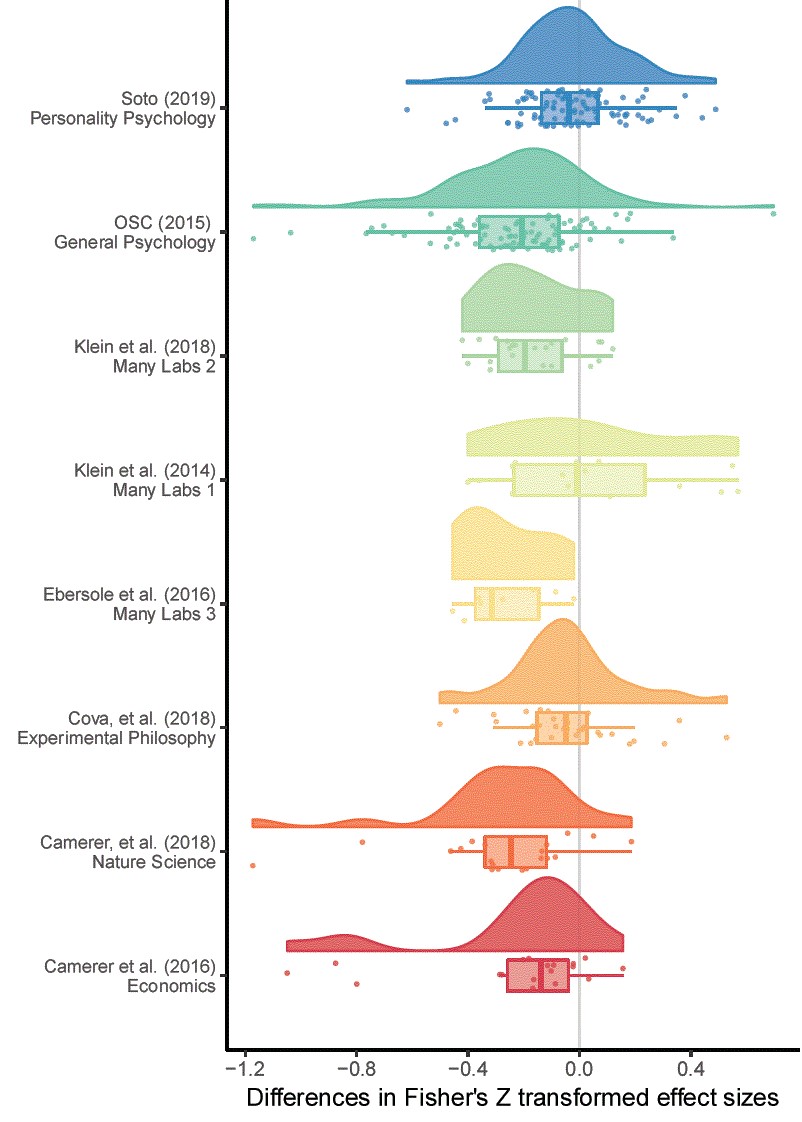


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of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

 Figure 2. A raincloud plot of the change in effect sizes (here Fisher Z scores) from the original to the replication study, panneled by the original study.

### Excluding null results

#### Examining only the statistically significant replicaiton studies

Looking at the 219 replications in which the replication study was statistically significant, the average effect for original studies was 0.41, and the mean effect size for replication studies was 0.38. There was an average decrease of r = -0.02 (naive 95% CI [-0.05, 0, an average decrease of 3.47%.

#### Examining only studies which were not statistically equivilent to the null

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

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

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n included | n criteria calculable for | Mean original ES | Median original ES | Mean replication ES | Median replicaiton ES | Mean ES difference | 95% CI LB Mean ES Change | 95% CI UB Mean ES Change | Median ES difference | SD difference | Mean proportion change | Median proportion change |
| Overall | 314 | 314 | 0.38 | 0.33 | 0.27 | 0.20 | -0.13 | -0.16 | -0.10 | -0.11 | 0.25 | -0.29 | -0.35 |
| StatisticalSignificance | 219 | 314 | 0.41 | 0.35 | 0.38 | 0.32 | -0.02 | -0.05 | 0.00 | -0.04 | 0.20 | 0.03 | -0.07 |
| Nonequivalence | 237 | 305 | 0.41 | 0.35 | 0.35 | 0.30 | -0.07 | -0.10 | -0.04 | -0.06 | 0.24 | -0.07 | -0.16 |
| BF0RepBelow3 | 220 | 303 | 0.40 | 0.34 | 0.37 | 0.32 | -0.04 | -0.06 | -0.01 | -0.05 | 0.20 | -0.01 | -0.13 |
| BFRep0Above3 | 186 | 303 | 0.41 | 0.35 | 0.40 | 0.36 | -0.01 | -0.04 | 0.02 | -0.01 | 0.20 | 0.09 | -0.05 |
| BF01Below3 | 221 | 304 | 0.42 | 0.36 | 0.36 | 0.32 | -0.06 | -0.10 | -0.03 | -0.05 | 0.25 | -0.04 | -0.13 |
| BF10Above3 | 177 | 304 | 0.41 | 0.35 | 0.40 | 0.35 | -0.01 | -0.05 | 0.02 | -0.01 | 0.20 | 0.08 | -0.04 |
| BF0PBelow3 | 232 | 304 | 0.42 | 0.35 | 0.36 | 0.31 | -0.07 | -0.10 | -0.03 | -0.05 | 0.24 | -0.04 | -0.14 |
| BFP0Above3 | 186 | 304 | 0.41 | 0.35 | 0.40 | 0.35 | -0.01 | -0.04 | 0.02 | -0.01 | 0.21 | 0.08 | -0.05 |

### Multilevel model

The model including all data estiamtes a -0.14 (95% CI [-0.2, -0.07]) point decrease in effect sizes from the orignal to replication studies in correlation coefficent terms. This is reresnets a change equvilant to -38.08 (95% CI [-55.91%, -19.92%]) of the mean effect size in the original studies (r = 0.36).

Looking at there was more variance attributable to the article (i.e., the original article) than too the project ( = 0.02 compared to = 0.01). QE tests of hetrogenetiy suggest that there is a large amount of unexplained hetrogeneity, QE(304) = 3527.86, p < .001.

The model was re-estimated using each the subsets of studies, excluding studies based on the exclusion criteria detailed above. See table [all model output] for the model estiamtes from each model. The estimates of the proportion of variance attributable to the article or replicaiton project level did not change considerably in any of these models. There is a notable reduction in the estimated effect sizes under these different selection criteria, with estimates of the amount of effect size decrease from r = -0.04 to -0.09, representing 9.98 to 24.83 of the average effect in the original studies. See supplemtary materials [all exclusion crtiera output] for full model output and scatterplots of the dataset using each exclusion rule.

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 replicaiton project the replicaiton was a part of) and the original (i.e., replicated) article or effect.

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Standard Error | 95% PI lower bound | 95% PI upper bound |
| Camerer et al. (2016) |  |  |  |  |
| Economics -0.06 | 0.05 | -0.15 | 0.04 |  |
| Camerer, et al. (2018) |  |  |  |  |
| Nature Science -0.07 0 | .05 | -0.17 | 0.03 |  |
| Cova, et al. (2018) |  |  |  |  |
| Experimental Philosophy 0.08 | 0.04 | -0.01 | 0.17 |  |
| Ebersole et al. (2016) |  |  |  |  |
| Many Labs 3 -0.06 0 | .06 | -0.17 | 0.05 |  |
| Klein et al. (2014) |  |  |  |  |
| Many Labs 1 0.11 | 0.05 | 0.01 | 0.21 |  |
| Klein et al. (2018) |  |  |  |  |
| Many Labs 2 -0.03 | 0.05 | -0.12 | 0.06 |  |
| OSC (2015) |  |  |  |  |
| General Psychology -0.05 | 0 | .04 | -0.12 | 0.03 |
| Soto (2019) |  |  |  |  |
| Personality Psychology 0.07 | 0 | .04 | -0.01 | 0.15 |

##### Table [all model output]

The number of studies included in each model, and the estimated correlation coefficent decrease from each model. Models were estiamted using Fisher Z transformed correlation coefficents and back transformed for interpretability.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Inclusion rule | Model N | Model Estimate | 95% CI lb | 95% CI ub |
| All studies | 305 | -0.14 | -0.20 | -0.07 |
| Statistically significant | 194 | -0.05 | -0.12 | 0.01 |
| BF0P < 3 | 228 | -0.09 | -0.16 | -0.02 |
| BFP0 > 3 | 182 | -0.05 | -0.11 | 0.02 |
| BF01 < 3 | 217 | -0.09 | -0.16 | -0.01 |
| BF10 > 3 | 173 | -0.05 | -0.11 | 0.02 |
| BF0Rep < 3 | 220 | -0.06 | -0.12 | 0.00 |
| BFRep0 > 3 | 186 | -0.04 | -0.09 | 0.02 |
| Non-equivalent | 235 | -0.08 | -0.15 | -0.01 |

## Discussion

Overall, there was an average decrease in correlation coefficent terms of -0.13, approximately equal to a Cohen’s d difference of -0.26. On average, replicaiton sample sizes were 30% smaller in replicaiton studies, than they were in original studies, a considerable decrease. In looking at the results of the multilevel meta-analysis including all data, there is an etimated -0.14 correlation coefficent decrease (95% CI [-0.2, -0.07]), equivalent to a -0.28 point Cohen’s d decrease (95% CI [-0.41, -0.14]). However, there is evidence to suggest that this decrease may be largely driven by the presence of null (or effectivly null) effects.

Averaging across exclusion methods, there is an -0.04 point average change in correlation coefficent terms from the original to replicaiton effect size, or an average decrease of just 1.5%. The results from the varied multilevel models provide support for this idea, showing that there is a much lower effect size decrease when attempting to exclude effectivly null results. The highest effect size decrease using any of these rules, estimating the decrease using only the results of the 228 experiments which did not provide have a of greater than 3 (i.e., which did not have at more than “anecdotal” evidence for the null hypothesis compared to the one sided alternative hypthesis), showed that an estimated decrease of modSumariesR -0.09, 95% CI [-0.16, -0.02].

#### Limitiations:

None of the projects included in this analysis were true random selections from the literature, and it is possible that the pattern in the selected sample may be different that that which would be seen in the literature overall.

All of the methods that were used to the replication studies which were null-or-effectively were bound to decrease the amount of effect size decrease that is seen. At worst, they could be seen as just removing the studies which happened to find low effects as opposed to removing all the true null hypotheses. However, this preliminary analysis does provide suggestive evidence that the degree of effect size attenuation that is seen may be largely attributed to the presence of effectively-null results, and that the overall. The current study also cannot distinguish between heterogeneity (i.e., effect sizes that are different under different scenarios) and effect size scenario. However, it seems reasonable to except in that effect size heterogeneity should lead to symmetrical effect size differences, leading to a mean unbaised estimate of the degree of effect size deflation that can be expected.

## Supplementary material

### Approximate bayes factors comparison

A comparison of the results of the Bayes Factors as estiamted here and as reported in {Camerer, 2018 #967} shows that they do, for the most part, agree with each other, although there are some notable discrepancies. See Table SM1 for the Bayes factors reported in {Camerer, 2018 #967} and those reported in the current paper. The only large discrepancy included is seen in Balafoutas and Sutter (2012) in which the Bayes Factor reported in {Camerer, 2018 #967} was based on a hypothesis test of orderd binomial probabilities, making it difficult to appropraitely convert into a correlation coefficient, and likely accounting for the large difference.

Table SM1. One-sided and () and replication () Bayes Factors for as reported in {Camerer, 2018 #967} and as estimated in the current paper, along with the reported correlation coefficients and sample sizes from the original and replication studies.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Article | Original\_r | Original\_N | Replication\_r | Replication\_n | Camerer\_et\_al.\_BFP0 | Camerer\_et\_al.\_BFRep0 | BFrep0 | BF0plus | BF01 |
| Ackerman et al. (2010), Science | 0.27 | 54 | 0.09 | 858 | 5.4e-01 | 3.1e-01 | 2.6e+00 | 2.1e+00 | 1.0e+00 |
| Aviezer et al. (2012), Science | 0.96 | 15 | 0.83 | 14 | 4.5e+02 | 5.7e+01 | 2.3e+02 | 2.7e+02 | 1.4e+02 |
| Balafoutas and Sutter (2012), Science | 0.28 | 72 | 0.15 | 243 | 4.2e+00 | 4.3e+00 | 4.1e+00 | 2.1e+00 | 1.1e+00 |
| Derex et al. (2013), Nature | 0.52 | 51 | 0.36 | 65 | 3.1e+03 | 3.7e+03 | 3.3e+01 | 2.2e+01 | 1.1e+01 |
| Duncan et al. (2012), Science | 0.67 | 15 | 0.37 | 128 | 2.7e+03 | 2.5e+03 | 2.0e+03 | 2.3e+03 | 1.2e+03 |
| Gervais and Norenzayan (2012), Science | 0.29 | 57 | -0.04 | 755 | 6.0e-02 | 3.0e-02 | 3.0e-02 | 2.0e-02 | 9.0e-02 |
| Gneezy et al. (2014), Science | 0.22 | 178 | 0.18 | 407 | 2.3e+02 | 4.9e+02 | 4.7e+02 | 1.1e+02 | 5.7e+01 |
| Hauser et al. (2014), Nature | 0.82 | 40 | 0.83 | 22 | 2.9e+03 | 1.0e+04 | 1.0e+05 | 2.5e+04 | 1.2e+04 |
| Janssen et al. (2010), Science | 0.63 | 63 | 0.34 | 42 | 5.9e+00 | 0.0e+00 | 1.9e+00 | 4.2e+00 | 2.1e+00 |
| Karpicke and Blunt (2011), Science | 0.60 | 40 | 0.38 | 49 | 1.5e+01 | 1.2e+01 | 1.4e+01 | 1.3e+01 | 6.5e+00 |
| Kidd and Castano (2013), Science | 0.27 | 86 | -0.04 | 999 | 5.0e-02 | 1.0e-02 | 1.0e-02 | 2.0e-02 | 8.0e-02 |
| Kovacs et al. (2010), Science | 0.45 | 24 | 0.59 | 95 | 5.6e+07 | 1.3e+08 | 8.7e+07 | 5.4e+07 | 2.7e+07 |
| Lee and Schwarz (2010), Science | 0.39 | 40 | -0.05 | 409 | 8.0e-02 | 1.0e-02 | 2.0e-02 | 3.0e-02 | 1.1e-01 |
| Morewedge et al. (2010), Science | 0.45 | 32 | 0.35 | 89 | 8.7e+01 | 1.6e+02 | 1.6e+02 | 8.1e+01 | 4.0e+01 |
| Nishi et al. (2015), Nature | 0.20 | 200 | 0.12 | 480 | 7.0e+00 | 7.8e+00 | 8.4e+00 | 2.9e+00 | 1.4e+00 |
| Pyc and Rawson (2010), Science | 0.38 | 36 | 0.15 | 438 | 6.8e+00 | 4.0e+00 | 1.7e+01 | 1.6e+01 | 8.0e+00 |
| Ramirez and Beilock (2011), Science | 0.79 | 20 | -0.09 | 105 | 1.4e-01 | 0.0e+00 | 0.0e+00 | 7.0e-02 | 1.9e-01 |
| Rand et al. (2012), Nature | 0.14 | 343 | 0.03 | 3150 | 1.4e-01 | 1.0e-01 | 1.3e-01 | 1.3e-01 | 7.0e-02 |
| Shah et al. (2012), Science | 0.27 | 56 | -0.04 | 897 | 7.0e-02 | 4.0e-02 | 4.0e-02 | 2.0e-02 | 8.0e-02 |
| Sparrow et al. (2011), Science | 0.37 | 69 | 0.07 | 338 | 1.5e-01 | 3.0e-02 | 6.0e-02 | 2.6e-01 | 1.5e-01 |
| Wilson et al. (2014), Science, | 0.67 | 30 | 0.59 | 39 | 6.0e+02 | 1.9e+03 | 1.9e+03 | 8.3e+02 | 4.2e+02 |

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

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

Table SM2. Multilevel meta-analysis model estimates and random effects for all data.

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

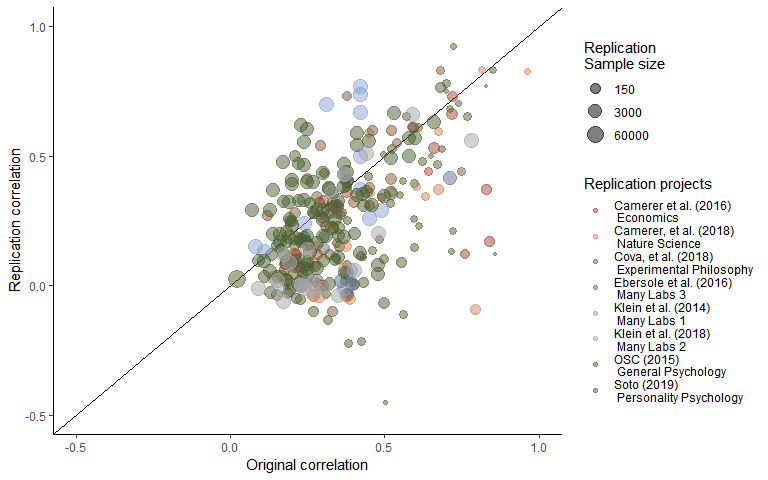


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including all data.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.05 | -0.12 | 0.01 | 0.03 | 0.1 |  |
|  |  |  |  |  | Project variance = 0.006, n = 8 |
|  |  |  |  |  | Article variance = 0.02, n = 128 |
|  |  |  |  |  | QE(193) = 2626.93, p < .001 |

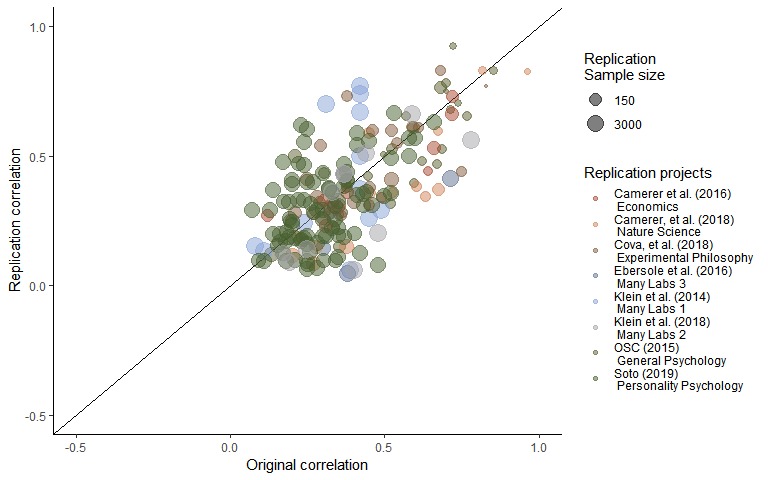


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including only statistically significant replications.

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

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

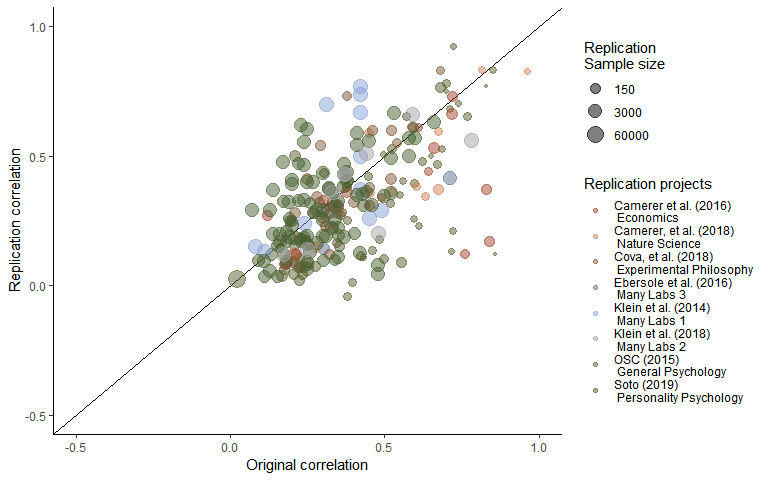


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original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

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

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

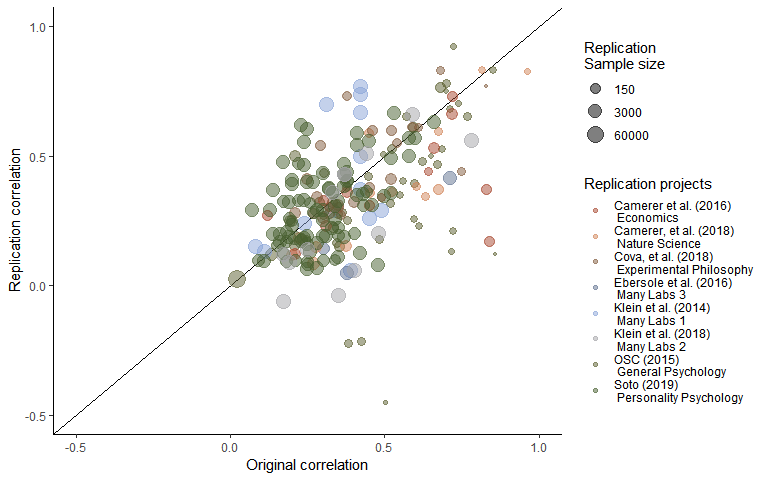


Figure SM4. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including only studies with < 3.

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

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

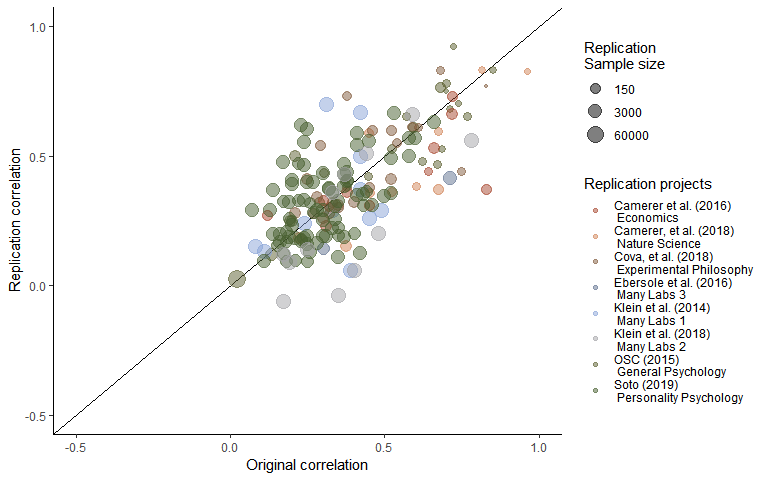


Figure SM5. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including only studies with > 3.

Table SM7. Multilevel meta-analysis model estimates and random effects for studies with < 3.

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

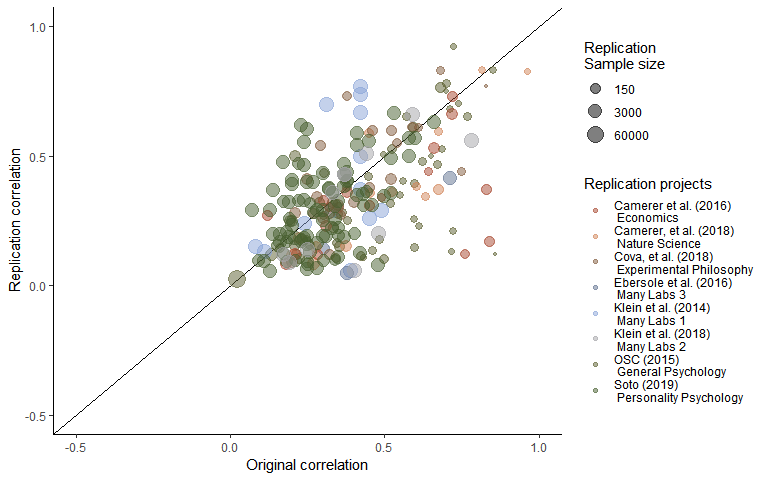


Figure SM6. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including only studies with < 3.

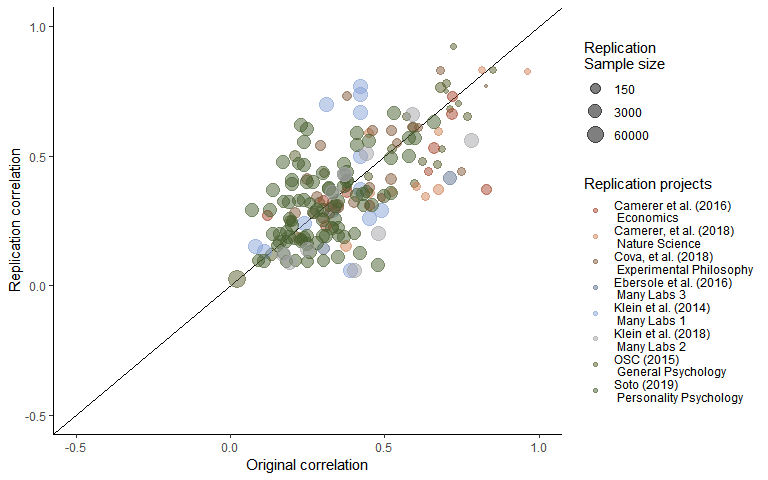


Figure SM7. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including [!]

#### FIX LABS FROM HERE !!!!!! !!! !

Table SM8. Multilevel meta-analysis model estimates and random effects for [!]

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

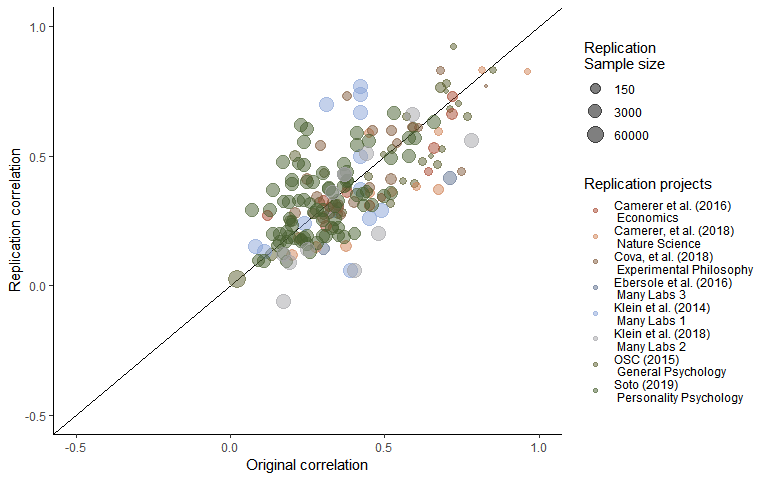
 Figure SM8. Scatter plot of replication effect sizes (in correlation coefficents) plotted against original effects including [!]

Table SM9. Multilevel meta-analysis model estimates and random effects for [!]

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

Table SM10. Multilevel meta-analysis model estimates and random effects for [!]

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

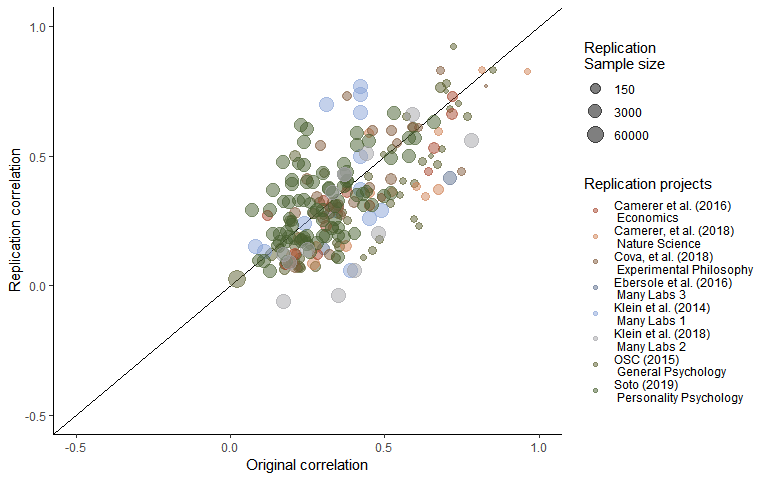


Figure SM

### Simulation of removal methods

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

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

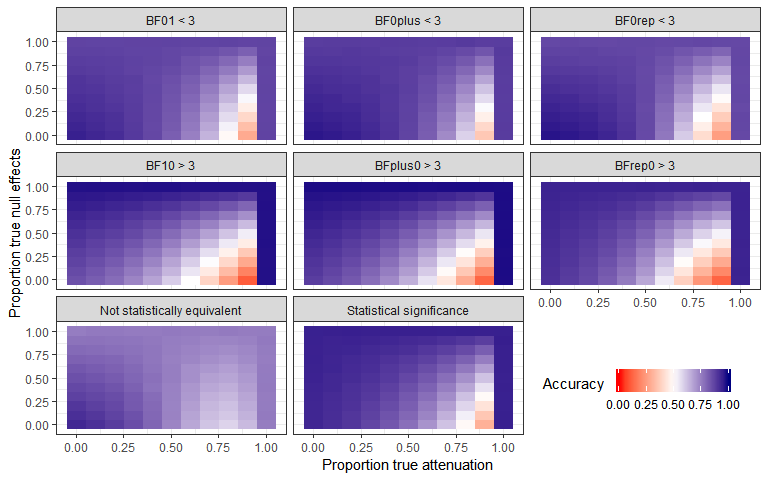


Figure [accuracy of cut scores]. The proportion of studies correctly classified in 11958 simulations of the accuracy of cut scores under varied true proportions of attenuation and proportion of effects which are true nulls.

Table [accuracy plot]

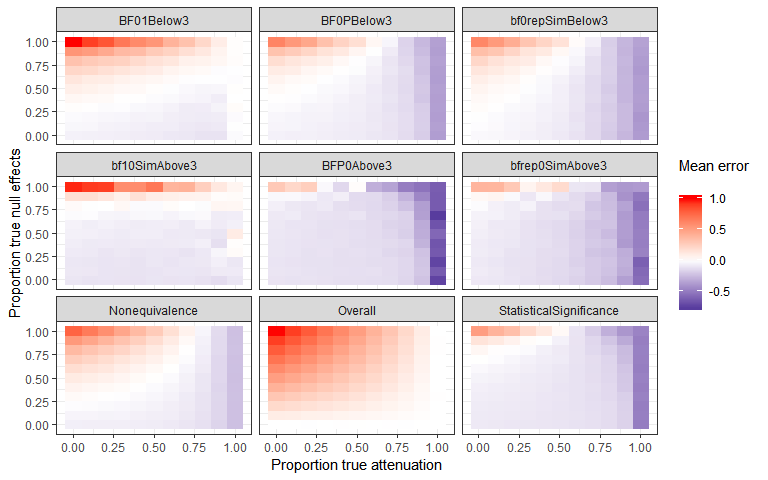
|  |  |  |
| --- | --- | --- |
| Data inclusion rule | Accuracy | Accuracy SD |
| Not statistically equivalent | 0.75 | 0.074 |
| Statistical significance | 0.84 | 0.135 |
| BF01 < 3 | 0.81 | 0.127 |
| BF10 > 3 | 0.78 | 0.127 |
| BFplus0 > 3 | 0.81 | 0.188 |
| BF0plus < 3 | 0.84 | 0.113 |
| BFrep0 >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

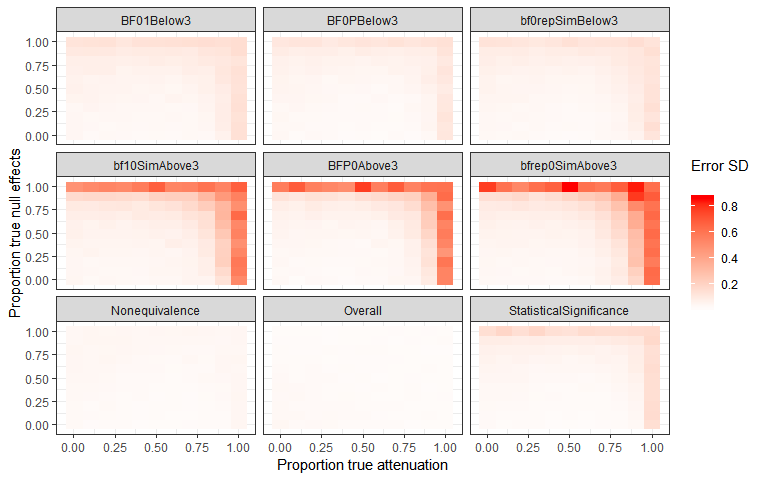
Additonally, simulations were performed using the same data-generation method to estimate the accruacy of the raw methods of estimating the simulated true proportion decrease under these different scenarios. Looking the mean proportion of effect size attenuation in the study, the results of the simulation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

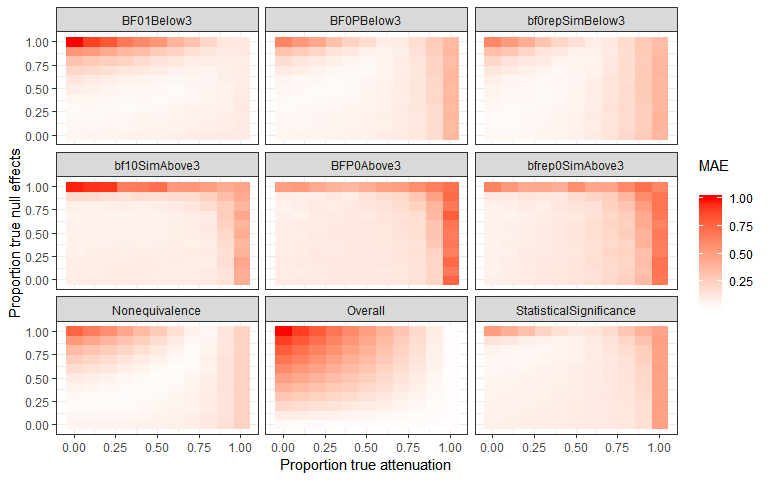
Note that these values are only valid under a specific data generation process, where there is a consistent factor effect size decrease, and where the studies which are null are random and independent of the original effect and sample sizes. See Table [simulation output] for the mean squared error (MSE),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

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subsample | nSims | MSE | Mean Error | RMSE | MAE | Error SD |
| BF01Below3 | 10101 | 0.05 | 0.08 | 0.23 | 0.14 | 0.22 |
| BF0PBelow3 | 10101 | 0.04 | -0.06 | 0.19 | 0.13 | 0.18 |
| bf0repSimBelow3 | 10101 | 0.04 | -0.06 | 0.21 | 0.15 | 0.20 |
| bf10SimAbove3 | 10101 | 0.11 | 0.00 | 0.32 | 0.19 | 0.32 |
| BFP0Above3 | 10101 | 0.13 | -0.18 | 0.36 | 0.22 | 0.32 |
| bfrep0SimAbove3 | 10101 | 0.15 | -0.17 | 0.39 | 0.25 | 0.35 |
| Nonequivalence | 14232 | 0.04 | 0.02 | 0.19 | 0.13 | 0.19 |
| Overall | 14232 | 0.12 | 0.25 | 0.35 | 0.25 | 0.24 |
| StatisticalSignificance | 14232 | 0.05 | -0.12 | 0.21 | 0.15 | 0.18 |







### LOO Cross validation output

# Table [LOO cross validation output](#loo-cross-validation-output).

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

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

# Table [LOO cross validation output](#loo-cross-validation-output).

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

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