Estimating the effect of publication and reporting bias

## Introduction

See word document

## Methods

### Included data extraction and imputation

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 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 available they were manually extracted from original articles where possible.

For analyses, all study level results were transformed to Fisher z Transformed correlation coefficents following the methods used in {Open Science Collaboration, 2015 #611}. All reported results are back-transformed from Fisher Z transformed correlation coefficients for ease of interprtation to correlation coefficients unless it is otherwise stated.

When the original and replication effect sizes were not reported as Fisher Z transformed correlation coefficients, effect sizes were converted from test statistics or effect sizes for analysis if the original results, those reported in the replication project were reported in correlation coefficients or Cohen’s d or where the orignal test result was reported as a t-test or F statistic, but not otherwise. In cases where sample sizes were not reported per group, sample sizes among groups were assumed to be equal in these conversions. See Table 1 for the number of valid studies extracted from each project. This means that the current analysis follows the assumptions for conversion between effects following each replication project, and that some studies have been left out of the current analysis (e.g., those which used Chi Square tests in {Open Science Collaboration, 2015 #611} and those reported only as Beta coefficents in [loopr citation]). See supplementary materials 1 for a comprehensive account of exclusions and study specific extraction details for each replication project. An original and replication effect size that could be converted to a Fisher z score was extracted for 312 replications, and this inforamtion along with sample sizes for original and replication studies was extracted for a total of 303 studies.

Table 1. The reference of each replication project, the number of replication studies performed as a part of each replication project, and the number of studies for which Fisher Z transformed correlation coefficents could be extracted for both the original and replication studies.

|  |  |  |  |
| --- | --- | --- | --- |
| Replication projects | Number of replication studies performed | Reported replication rate (statistically significant results in the same direction) | Included studies |
| Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. Science, 351(6280), 1433. DOI: 10.1126/science.aaf0918 | 18 | 61% | 18 |
| Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2(9), 637-644. <doi:10.1038/s41562-018-0399-z> | 21 | 62% | 21 |
| Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Colombo, M. (2018). Estimating the reproducibility of experimental philosophy. Review of Philosophy and Psychology, 1-36. doi: 10.1007/s13164-018-0407-2. | 37 | 78% | 33 |
| Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67, 68-82. <doi:10.1016/j.jesp.2015.10.012> | 9 | 33% | 8 |
| Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating Variation in Replicability. Social Psychology, 45(3), 142-152. <doi:10.1027/1864-9335/a000178> | 16 (13 effects) | 88% (85%) | 15 |
| Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances In Methods and Practices in Psychological Science, 1(4), 443-490. <doi:10.1177/2515245918810225> | 28 | 54% | 22 |
| Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. <doi:10.1126/science.aac4716> | 97 | 36% | 94 |
| Soto et al (2018) | 121 | 86% | 101 |

Note: Soto et al’s (2018)’s replication rate and was recalculated on the “study” (i.e., using the number of replicated effects not the number of trait-outcome associations as is reported in the paper) using results disattenuated using the Spearman-Brown prediction formula and Spearman disattenuation formula (Lord & Novick, 1968) to account for less reliable shorter form measures used in the replication studies. Klein et al. (2014) includes 4 operationalisations of a single effect which were input seperatly for analysis in the current study.

### Analysis

This paper uses three approaches to examining the change in effect sizes from original to replication studies. The first approach is descriptive, giving the mean, median and mean raw effect size decrease, along with the mean and median proportion decrease in effect sizes seen across the sample. The reported Wald-type confidence intervals do not account for non-independence between effects taken from the same paper, or between studies from the same replication projects. The second approach is designed to account for this non-indepdence using a random effects meta-analysis framework. The final appraoch developed below uses the Bayesian Mixture Model developed and presented in {Camerer, 2018 #967} to estimate the ammount of effect size decrease.

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. 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}. 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 coefficents for easy interpretation unless otherwise stated. Studies are excluded case-wise, and as exclusion methods sometimes required data that was not avaliable (such as sample sizes for original or replication studies), sample sizes are reported alongside each analysis in tables.

#### Accounting for null effects

An important question in assessing the degree to which effects are attenuated in this literature is how much this effect is driven by the presence of a subset of null effects (or effects so small as to be effectively null). The average attenuation could be extremely high, and yet this effect be almost entirely driven by the presence of effectively-null effects. In order to account for this issue, three main approaches were taken.

Firstly, original studies were simply excluded using various exclusion criteria (detailed below), and raw effect size differences calculated and multilevel meta-analysis models re-estimated. Because all of these methods function by removing small effects, no significance testing was performed on the difference between the model estimates estimated decreases after accounting for small or near-null effects. It is certain that, at a population level, all of these exclusion rules would lower the size of the observed effect size decrease. The second method of accounting for the presence of null or effectively null effects was to include the p value of the replication studies as a moderator in the meta-analysis of the effect size differences. In this model the meta-analytic mean is the predicted mean effect size decrease assuming a replication p value of 0. Finally, the Bayesian mixture model adapted from {Camerer, 2018 #967} is designed to estimate the propbability that each study is sampled from the null distribution as well as the expected amount of effect size decrease when studies are sampled from a non null distribution.

Multiple exclusion rules were used to attempt to account for null results in the raw descriptives and in the multilelvel model; excluding studies in which the replication study was not significant, removing statistically equivalent studies using equivalence testing, and using a cut score from approximate Bayes Factors estimated from the reported correlation coefficient effect sizes. Plots of all of the subsets are provided in supplementary materials [exlusions].

###### Statistical significance of the replication study

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

###### Equivalence tests

In order to avoid excluding under-powered studies erroneously, we also excluded studies based on whether the results of the replication study were statistically equivalent to the null {Lakens, 2017 #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 of the Fisher Z transformed correlation coefficients. As a method of testing how closely this method of approximating standard errors matches the original replication projects results, significance tests for the replication and original studies were performed using this approximation. The results matched the significance or non-significance as reported in the replication projects in every case.

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

###### Bayes factors

Three 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 favor of one model or another. If a Bayes factor is greater than one the data is more likely under the alternative hypothesis than under the null hypothesis, and the opposite is true when a Bayes factor is below one. Conventional labels have been proposed, suggesting that Bayes factors between 1 and 3 provide little to no evidence (or ‘anecdotal’ evidence) and Bayes factors from 3-10 provide “substantial” evidence {Jeﬀreys, 1961 #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 study’s experimental design. Importantly, these Bayes factors differ from those that would normally be developed using the closest Bayesian equivalents to each original replicated study’s analysis, and should be viewed only as a coarse estimate of the degree of evidence provided for and against the null model. See table [bayesFactors] in supplementary materials [Bayes] for a table showing the differences between the values returned by this method compared to those reported in the Bayesian supplement to which were more appropriately calculated {Camerer, 2018 #967}, which demonstrates that the difference can be considerable, especially when the original analysis was idiosyncratic. Normally, One of the benefits of Bayes Factors is their continuous and interpretable scale, however in this case these approximate Bayes Factors are used as a heuristic to discard the studies which appear to likely be true (or effectively) null effects. Two different cut scores were used for each type of Bayes factor, discarding studies when Bayes factors suggested that the null model is either more than three times more likely than the alternative model (i.e., when there is more than ‘anecdotal’ evidence that the null is true), or when the alternative model is not at least three times more likely than the null model. A full Bayesian treatment of this issue is presented in the Bayesian Mixture Model below.

#### Simulations to assess exclusion criteria

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

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

### Approach 1: descriptives

#### Raw changes in effect sizes

Looking at the 312 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 225 articles, 72% of the included articles. The average effect size for original studies was 0.39, and the mean effect size for replication studies was 0.27. There was an average decrease of r = -0.13 (Wald-type 95% CI [-0.16, -0.1], the slight discrepancy between values occurs because Fisher Z scores have been back transformed to correlation coefficients). Notably, this represents an average decrease in effect sizes from the original to the replication study of -29.19%. 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.

(i.e., )

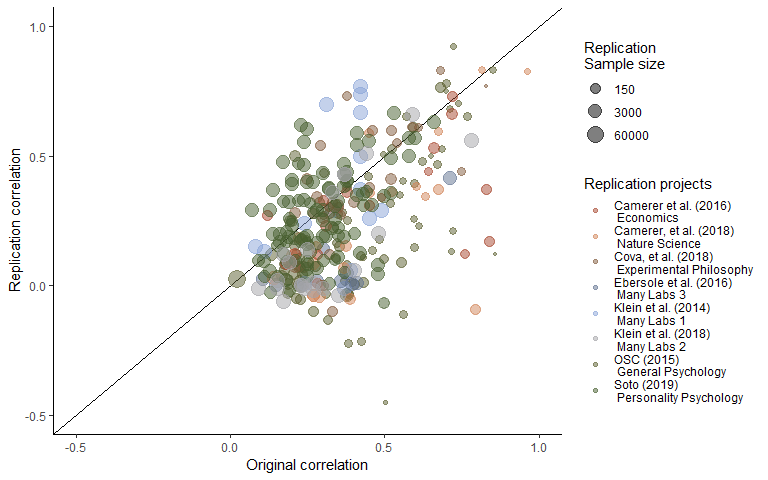


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of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

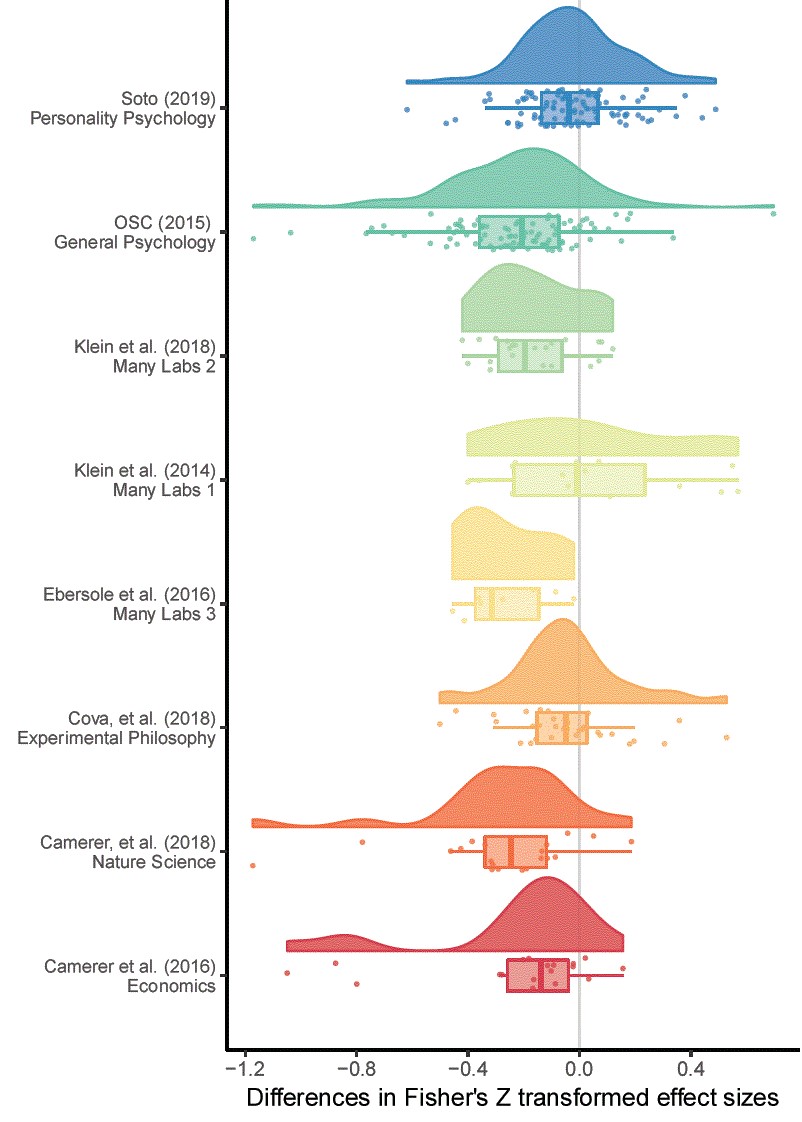


Figure 2. A raincloud plot of the change in effect sizes (here Fisher Z scores) from the original to the replication study by the replication project that each replication study was performed as a part of.

#### Descriptives accounting for null effects

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

Excluding studies which are not statistically significant is likely to lead to an underestimate of the degree of effect size attenuation, as this exclusion rule will lead to the exclusion of under-powered replication studies as well as studies which are likely to be true null effects. In order to avoid this issue, equivalence tests were performed, meaning that the studies which are not-statistically equivalent to the null are included (using a bound of equivalence equal to the minimum detectable effect in the original study). This method is an attempt to not exclude the non-diagnostic replication studies, studies which are not statistically significant but which do not suggest that that the result is statistically equivalent to the null. Using these method 235 replications were not statistically equivalent to the null, 77.56% of studies for which equivalence tests could be performed. The average effect size in the original non-equivalent studies was 0.41, compared to a mean effect size for replication studies of r = 0.35. This is a mean decrease of r = -0.07, Wald-type 95% CI [-0.1, -0.04], an average decrease of -6.63%.

The results of the various Bayes Factors analyses generally support the results of the analysis removing statistically equivalent studies. Using this method, 176 to 230 replications were included, 57.89 to 75.66% 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 | 312 | 312 | 0.39 | 0.33 | 0.27 | 0.21 | -0.13 | -0.16 | -0.10 | -0.11 | 0.25 | -0.29 | -0.35 |
| StatisticalSignificance | 201 | 312 | 0.40 | 0.35 | 0.38 | 0.32 | -0.02 | -0.05 | 0.00 | -0.04 | 0.20 | 0.03 | -0.07 |
| Nonequivalence | 235 | 303 | 0.41 | 0.35 | 0.35 | 0.30 | -0.07 | -0.10 | -0.04 | -0.06 | 0.24 | -0.07 | -0.16 |
| BF0RepBelow3 | 219 | 303 | 0.40 | 0.35 | 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 | 220 | 304 | 0.42 | 0.36 | 0.36 | 0.32 | -0.06 | -0.10 | -0.03 | -0.05 | 0.25 | -0.04 | -0.13 |
| BF10Above3 | 176 | 304 | 0.41 | 0.35 | 0.40 | 0.35 | -0.01 | -0.05 | 0.02 | -0.01 | 0.20 | 0.08 | -0.04 |
| BF0PBelow3 | 230 | 304 | 0.42 | 0.35 | 0.36 | 0.31 | -0.07 | -0.10 | -0.03 | -0.05 | 0.24 | -0.04 | -0.15 |
| BFP0Above3 | 185 | 304 | 0.41 | 0.35 | 0.40 | 0.35 | -0.01 | -0.04 | 0.02 | -0.01 | 0.21 | 0.08 | -0.05 |

### Approach 2: Multilevel meta-analysis

#### Analysis details

In order to obtain a reasonable estimate of the change in effect size between original and replication studies, a multilevel random effects meta-analysis was performed on the difference in Fisher Z transformed correlations between original and replication studies.

This analysis treats each pair of effects, the original and replicated sample size, as one “study” in a meta-analytic framework, and estiamtes the change from origianl to replication study (). In order to account for non-independence between studies random effects for journal ( and study () are included. 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. Empirical Bayes estimates and 95% credible intervals for random effects were calculated following {Robinson, 1991 #999}{Morris, 1983 #1000}.

#### Multilevel model

The model including all data estimates a r = -0.14 (95% CI [-0.2, -0.07]) decrease in effect sizes from the original to replication studies. This is represents a change equivalent to -34.17% (95% CI [-50.76%, -17.59%]) of the mean effect size in the original studies (r = 0.36).

More variance attributable to the article than too the project ( = 0.02 compared to = 0.01), representing an intraclass correlation (ICC) for the article of 0.23, meaning there is only a low correlation between the amounts of effect size change seen within projects. QE tests of heterogeneity suggest that there is a large amount of unexplained heterogeneity, QE(302) = 3525.8, p < .001, unsurprisingly given the heterogeneous sample included in the current sample.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.14 | -0.21 | -0.07 | 0.03 | < .001 |  |
|  |  |  |  |  | Project variance = 0.007, n = 8 |
|  |  |  |  |  | Article variance = 0.025, n = 226 |
|  |  |  |  |  | QE(302) = 3525.8, p < .001 |

Table [BLUP]. Empirical Bayes estimates and 95% credible intervals for random effects (i.e., estiamtes 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 | -0.06 | 0.05 | -0.16 | 0.04 |
| Camerer, et al. (2018) Nature Science | -0.07 | 0.05 | -0.17 | 0.03 |
| Cova, et al. (2018) Experimental Philosophy | 0.08 | 0.05 | -0.01 | 0.17 |
| Ebersole et al. (2016) Many Labs 3 | -0.06 | 0.06 | -0.18 | 0.05 |
| Klein et al. (2014) Many Labs 1 | 0.11 | 0.05 | 0.01 | 0.22 |
| Klein et al. (2018) Many Labs 2 | -0.03 | 0.05 | -0.12 | 0.06 |
| OSC (2015) General Psychology | -0.05 | 0.04 | -0.13 | 0.03 |
| Soto (2019) Personality Psychology | 0.07 | 0.04 | -0.01 | 0.15 |

#### Accounting for null results

The model was re-estimated using each the subsets of studies, excluding studies based on the exclusion criteria detailed above. See table [all model output] for the model estimates from each model. The estimates of the proportion of variance attributable to the article or replication project level did not change considerably in any of these models. There is a notable reduction in the estimated effect sizes under these different selection criteria, with estimates of the amount of effect size decrease from r = -0.04 to -0.09, representing 8.91% to 22.31% of the average effect in the original studies.

It is important to emphasize the degree of uncertainty in these results. For example, taking the highest effect size decrease using any of the exclusion criteria, estimating the decrease using only the results of the 227 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 hypothesis), showed that an estimated decrease of summarize r = -0.09, 95% CI [-0.16, -0.02]. Looking at the smallest estimated effect size difference under any exclusion criteria on the other hand, running the multilevel meta-analysis on just the results of the 186 experiments which had a of three or more (i.e., which had a replication Bayes factor which showed more than ‘anecdotal’ evidence for the alternative hypothesis), showed an estimated decrease of -0.04, 95% CI [-0.09, 0.02]. This is equivalent to a decrease of -8.91% of the average original effect size, with a 95% confidence interval that extends from a noticeable decrease to a moderate increase; 95% CI [-23.38, 5.55].

See supplementary materials [exclusions] for full model output and scatter plots of the data-set using each exclusion rule.

##### Table [all model output]

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Inclusion rule | Model N | Model Estimate | 95% CI lb | 95% CI ub | Estimated % attenuation | LB % attenuation | UB % attenuation |
| All studies | 303 | -0.14 | -0.20 | -0.07 | -34.2 | -51 | -17.59 |
| Statistically significant | 195 | -0.05 | -0.12 | 0.01 | -13.1 | -29 | 2.66 |
| Non-equivalent | 234 | -0.08 | -0.15 | -0.01 | -20.5 | -38 | -2.61 |
| BF0P < 3 | 227 | -0.09 | -0.16 | -0.02 | -22.3 | -40 | -4.55 |
| BFP0 > 3 | 182 | -0.05 | -0.11 | 0.02 | -11.9 | -28 | 3.78 |
| BF01 < 3 | 217 | -0.09 | -0.16 | -0.01 | -21.4 | -39 | -3.65 |
| BF10 > 3 | 173 | -0.05 | -0.11 | 0.02 | -11.9 | -28 | 4.31 |
| BF0Rep < 3 | 219 | -0.06 | -0.12 | 0.00 | -14.5 | -29 | -0.03 |
| BFRep0 > 3 | 186 | -0.04 | -0.09 | 0.02 | -8.9 | -23 | 5.55 |

### Including replication p values as a moderator

Given that the above methods all rely on throwing away data, a preferable method of modeling these results may be including the p value of the replication study as a moderator. A total of 303 studies provided enough data to be included in this model (i.e., studies for which replication p values, as well as sample sizes and effect sizes for replication and original studies, were extracted). This analysis leads to similar conclusions to the models with exclusions, with an estimated effect size decease of r = -0.08, 95% CI [-0.14, -0.02]. This represents a decrease of -20.2% (95% CI [-34.27%, -6.12%]) of the average original effect size. The projects and article level differences are functionally identical to the model that does not include replication p values as a moderator.

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Estimate | -0.082 | -0.14 | -0.025 | 0.029 | 0.005 |  |
| p value | -0.326 | -0.40 | -0.255 | 0.036 | < .001 |  |
|  |  |  |  |  |  | Project variance = 0.005, n = 8 |
|  |  |  |  |  |  | Article variance = 0.021, n = 226 |
|  |  |  |  |  |  | QE(302) = 3143.71, p < .001 |

##### Leave one out cross validation

To assess how sensitive the results of the multilevel model are to the inclusion of each of the replication projects and individual replicated effects within each replication project, the models were rerun using leave one out cross validation, excluding both the individual replication attempts and the replication projects one at a time. When leaving out individual studies the range of point estimates (i.e., the difference between the smallest and largest estimate of the difference between original and replication studies) for each of the LOO cross validation models did not exceed more than a Fisher z sore of 0.02. When excluding one replication project at a time, model estimate ranges did not exceed 0.04. See supplementary material [LOO] for a table of the proportion of model intercept 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.

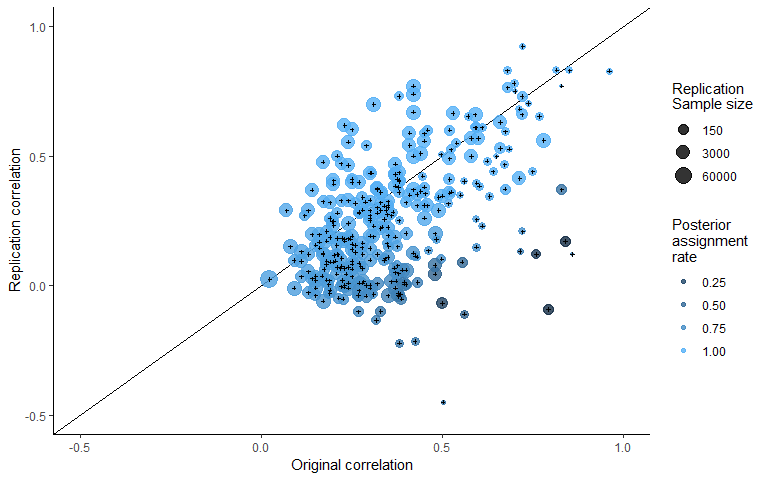
### Bayesian mixture model

The final approach to estimating the amount of effect size attenuation expected given that the effect under question is non-zero was the Bayesian mixture model presented in {Camerer, 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 model was estimated using four Markov chains from each of which 100,000 draws were taken (excluding a 11,000 draw burn-in period). Trace and density plots for the discussed parameters were assessed and the model appeared to have successfully converged. This analysis was performed in JAGS version 4.3.0 {Depaoli, 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.

#### Mixture model results

A total of 305 effects were included in the mixture model (i.e., all studies for which sample sizes and effect sizes for replication and original studies were extracted). The overall posterior assignment rate (i.e., the proportion of studies which are estimated to be from the non-null alternative hypothesis) is 88%, with a 95% highest probability density interval of [79%, 97%]. The overall attenuation factor is 19.3% with a 95% highest probability density interval of [11%, 28%]. Figure [mixture model], shows the original effect sizes plotted against replication effect sizes weighted by sample size, along with the posterior assignment rate. The color of each point indicates the proportion of times each effect was assigned to the alternative hypothesis. As was seen and pointed out in the first use of this model in {Camerer, 2018 #967}, values close to the diagonal are reliably assigned to the alternative hypothesis whereas effects far below the diagonal are more often assigned to the null hypothesis, although the overall posterior assignment rate might be overly optimistic (i.e., assign studies to the non-null hypothesis at a high rate). In part, this optimism may be due to the fact that this model allows for true effect sizes to be estimated as being extremely low or near zero due to sampling variability alone and still assigned to the alternative hypothesis not the null, with 29% of the estimated true replication effect sizes being smaller than an r of .1.



###### Figure [mixture model].

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

## Discussion

Overall, there was a substantial average decrease in effects sizes between the original and replication study. Taking the raw estimate, there was a moderate decrease of r = -0.13, approximately equal to a Cohen’s d difference of -0.26. On average, replication sample sizes were -13% smaller in replication studies than they were in original studies. The results of the multilevel meta-analysis results agree with this estimate, showing an estimated mean decrease of r = -0.14, (95% CI [-0.2, -0.07]), equivalent to a -0.28 point Cohen’s d decrease (95% CI [-0.42, -0.14]), or an estimated decrease of -34.17 (95% CI [-50.76%, -17.59%]) of the mean effect size in the original studies (a Fisher Z equal to r = 0.39). All of these estimates are higher than that suggested by {Stanley, 2018 #1002} of 8 - 15%, in fact being more similar to

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

Although excluding data is a less than ideal way of investigating this issue, the results raw results and multilevel models performed excluding data also support these results. The raw mean decreases seen across the exclusion methods range from an decrease of r = -0.07 to -0.01 from the original to replication effect size. The multilevel models estimated excluding data lead to similar conclusions, although they highlight the degree of uncertainty in this result. Although all models estimated excluding data estimates show a lower effect size decrease when attempting to exclude null (or effectively null) results, the confidence intervals for all results extend from a decrease of -44.26% of the average original correlation coefficient, to an increase of 6.19%.

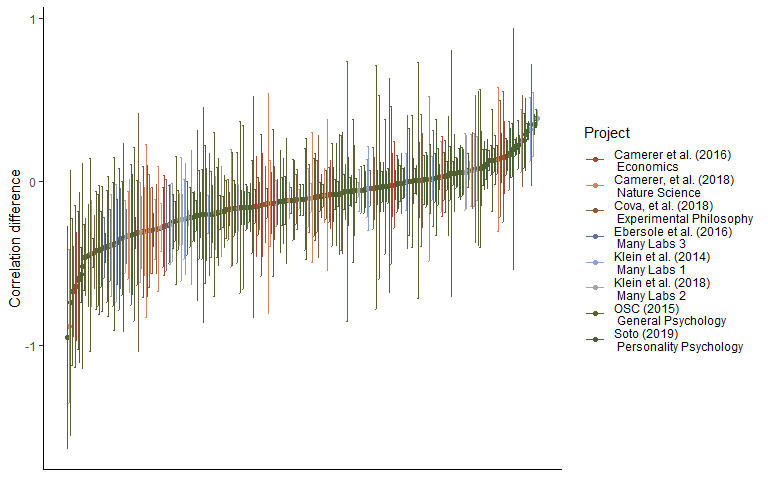


Figure x. A caterpillar plot of the effect size difference between original and replication study effect sizes ordered by magnitude, error bars are 95% confidence intervals around effect size differences.

#### Limitiations and future directions

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

Furthermore, although the two alternative methods of estimating the amount of effect size reduction seen in non-null models (i.e., the model including replication p values as a moderator and the Bayesian mixture model) provide convergent evidence supporting the estimated effect size decrease assuming a non-null effect, both have issues. The model including p values as a moderator is best considered a heuristic guide, and is not straightforwardly interpretable (i.e., the model provides an estimate of the expected effect size difference given a replication p value of 0, an impossibility, and really acts as a convenience method for adjusting for the probability of the data being observed under the null model in the replication study). The Bayesian mixture model used here has two major issues. Firstly, it assumes independence between effect and that there is a uniform attenuation factor across all areas of psychological research. Secondly, the modeled true effect size can be negligible or even negative and the replication effect size still assumed to be sampled from the alternative distribution not the null. This later issue occurs because the model estimates the relationship between original and replication effect sizes using a two step errors-in-variables approach following {Matzke, 2017 #1012}. Likely in large part due to this second issue, the model estimates the number of false positives at an extremely low rate (11.55% in contrast to other estimates of the number of false positive in this sample, which suggest a false positive rate of approximately 25%). For descriptive and interpretative purposes, this model does have a major benefit in that it estimates a single overall attenuation rate for non-null studies, the main goal of the current article. However, an important task in developing a more nuanced understanding of the data-generation process that leads to this data-set would be building a model that allows for the attenuation rate to change across replication studies, and possibly to allow it to include more components allowing for studies with negligible or negative but non-null effects in addition to the true alternative and null components modeled in the current studies.

### Conclusion

These results highlight some major issues in the psychological research literature. Looking at the raw average proportion decrease, -29%, or at the results of the multilevel model, an r =-0.14 (95% CI [-0.2, -0.07]) average decrease, differences which would make a considerable difference in most research scenarios. Researchers reading the literature should be aware of this large discrepancy, and plan their future experiments accordingly. Researchers who wish to ensure that they do not perform experiments that are unlikely to detect real effects should be aware that their experiments are likely to be under-powered if they plan their sample sizes using the effect size reported in a previous experiment. A researcher basing their intuitive estimates (or formal models) of the effect sizes that should be expected based on primary research findings are likely to be extremely optimistic. As a heuristic for researchers planning sample sizes, researchers could follow the advice given in {Camerer, 2018 #967} and plan their experiments assuming that the original effect size is 50% of its reported value, a value matched by the more extreme 95% confidence interval of the estimated amount of effect size decrease using the multilevel meta-analytic framework. Alternatively, it may be preferable to rely on estimates of the smallest effect size of interest or use flexible analysis strategies that do not rely on precise a priori specification of the sample size to be collected {Schönbrodt, 2017 #672;Albers, 2018 #942;Lakens, 2014 #169} when practical constraints mean this approach is feasible.

Finally, this project emphasizes the importance of efforts to reduce publication and reporting biases. In many cases pre-registration with a sufficiently detailed analytic strategy may help to combat reporting bias by making it easier for researchers to avoid engaging in questionable research practices {Wicherts, 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. 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 coefficents 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 projects {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.

##### {LOOPR}

Effect sizes were extracted for original and replication studies for 101 out of 121 included studies. 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 avaliable in the data set. 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 and are included in the current study for 33 original and replication studies, out of an original 37 replicated studies with significant original results.

##### 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 anlaysis. 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 original and replication effects sizes were extracted 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 derrive 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 extacted for all 9 original and replicaiton studies from {Ebersole, 2016 #985}, excluding a study they term a “conceptual replication”. Most effects (6/9) were converted to correlation coefficents from the Cohen’s d values reported in this replicaiton project. The results of three additional studies reported as partial Eta squared were converted to correlation coefficents 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 coefficents and all are included in this analysis.

### Approximate bayes factors comparison

A comparison of the results of the Bayes Factors as estimated here and as reported in {Camerer, 2018 #967} shows that they agree with each other in approximate magnitude and direction for the most part, although there are some notable discrepancies. See Table SM1 for the Bayes factors reported in {Camerer, 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 ordered binomial probabilities, making it difficult to appropriately convert into a correlation coefficient, and likely accounting for the large difference.

Table SM1. One-sided and () 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 |

### SM2

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

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

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.14 | -0.21 | -0.07 | 0.03 | < .001 |  |
|  |  |  |  |  | Project variance = 0.007, n = 8 |
|  |  |  |  |  | Article variance = 0.025, n = 226 |
|  |  |  |  |  | QE(302) = 3525.8, p < .001 |

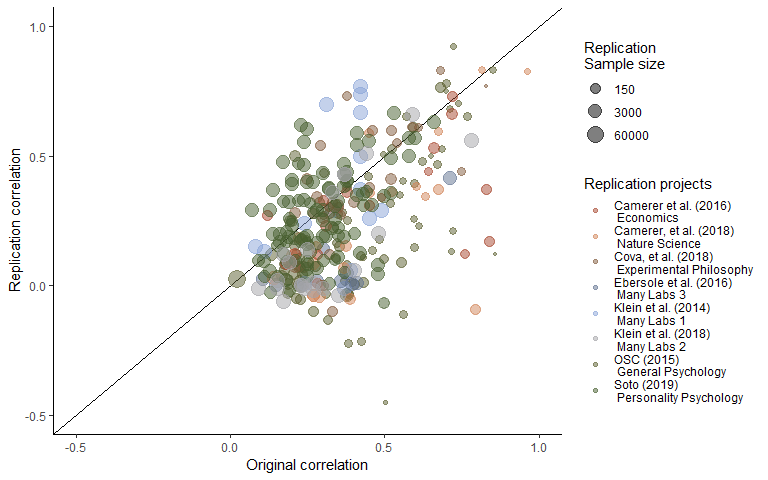


Figure SM1. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including all data.

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

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

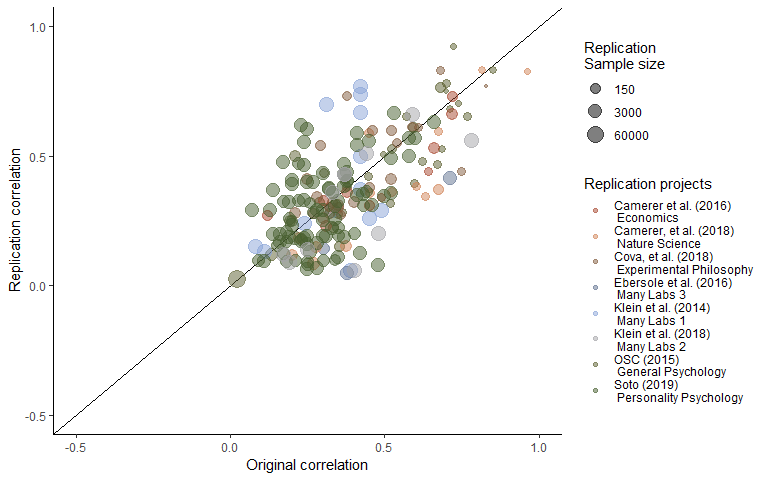


Figure SM2. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only statistically significant replications.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| -0.08 | -0.16 | -0.01 | 0.04 | 0.02 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.025, n = 166 |
|  |  |  |  |  | QE(233) = 3023.8, p < .001 |

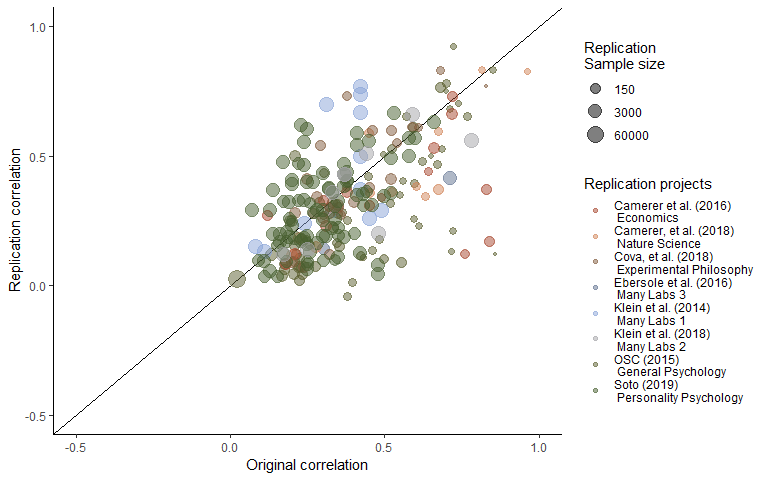


Figure SM3. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies which are not statistically equivalent to the null, using equivalence bounds set as the minimum effect size that would have been statistically significant in the original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

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

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

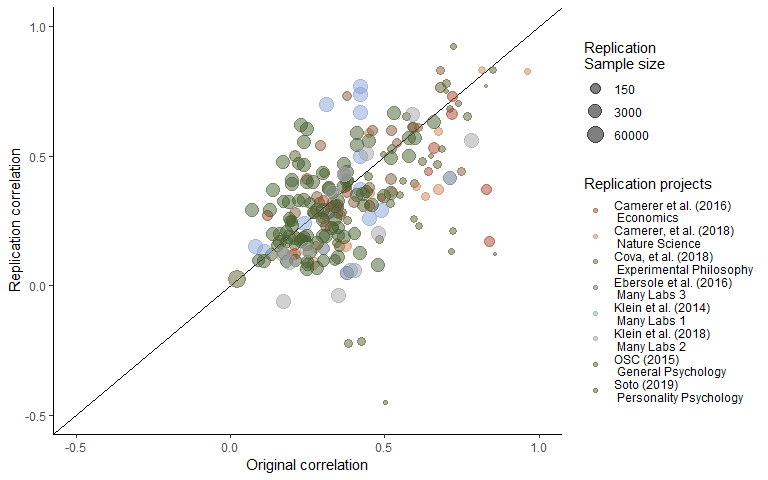


Figure SM4. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with < 3.

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 |

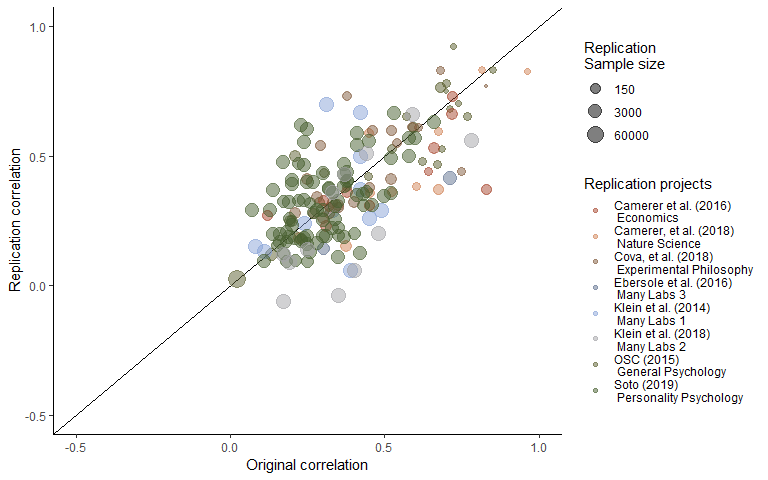


Figure SM5. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with > 3.

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 = 0.026, n = 160 |
|  |  |  |  |  | QE(226) = 2885.82, p < .001 |

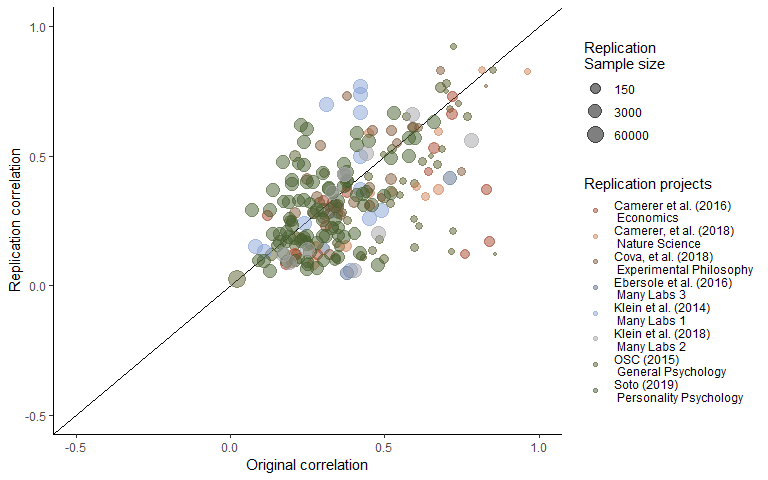


Figure SM6. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including only studies with < 3.

Table SM8. Multilevel meta-analysis model estimates and random effects for studies with

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

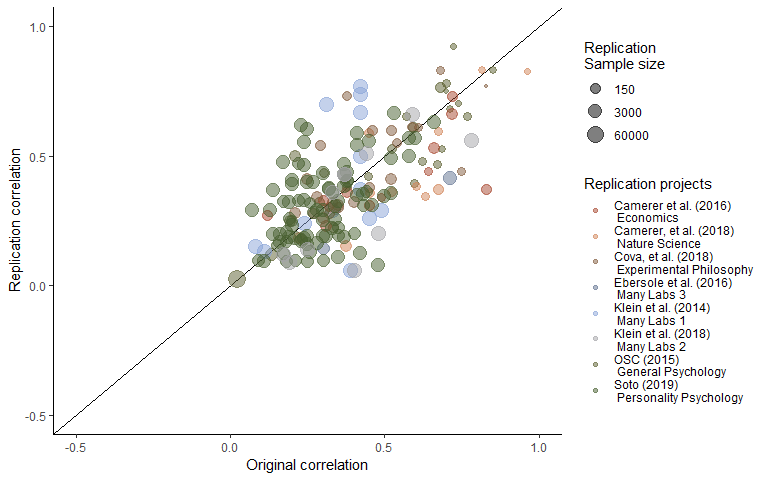


Figure SM7. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with

Table SM9. Multilevel meta-analysis model estimates and random effects for studies with

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

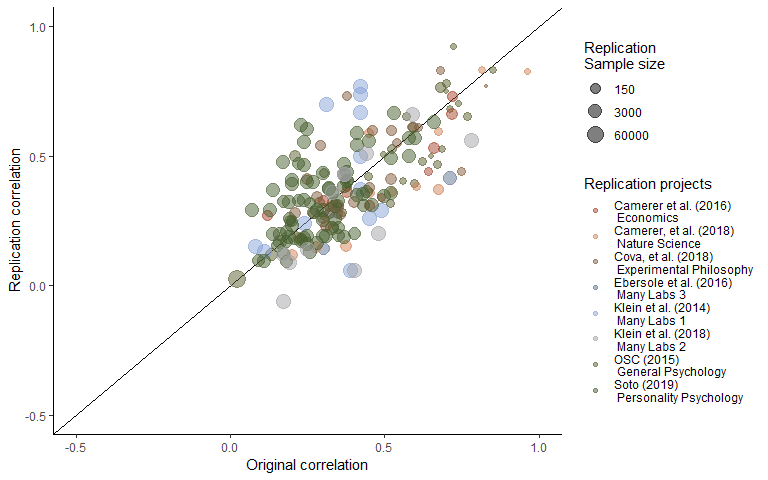


Figure SM8. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with

Table SM10. Multilevel meta-analysis model estimates and random effects for studies with

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 = 158 |
|  |  |  |  |  | QE(218) = 2532.39, p < .001 |

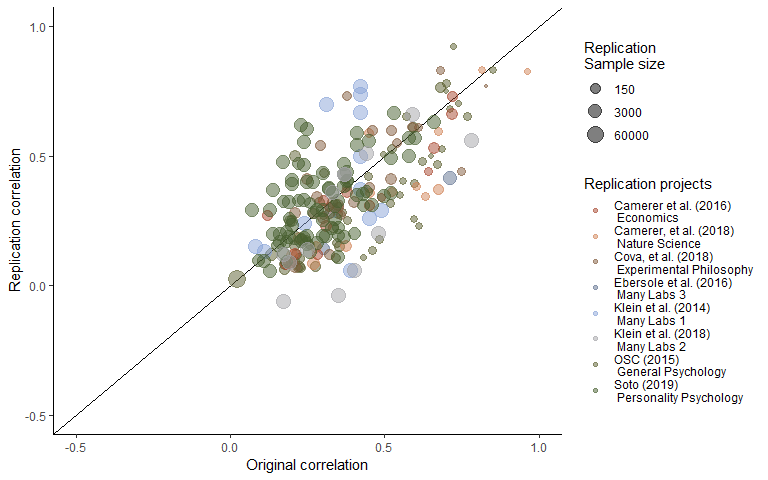


Figure SM9. Scatter plot of replication effect sizes (in correlation coefficients) plotted against original effects including studies with

## SM3

### Simulation of exclusion method accuracy

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

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

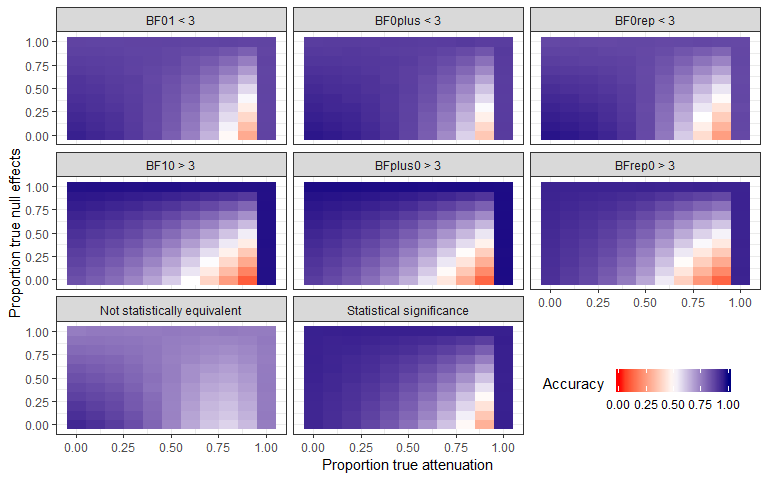


Figure [accuracy of cut scores]. The proportion of studies correctly classified in 11958 simulations of the accuracy of cut scores under varied true proportions of attenuation and proportion of effects which are true nulls.

Table [accuracy plot]

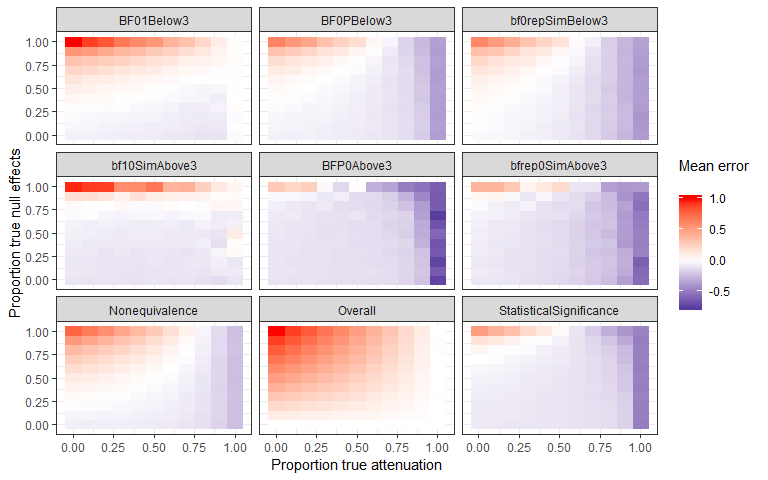
|  |  |  |
| --- | --- | --- |
| Data inclusion rule | Accuracy | Accuracy SD |
| Not statistically equivalent | 0.75 | 0.074 |
| Statistical significance | 0.84 | 0.135 |
| BF01 < 3 | 0.81 | 0.127 |
| BF10 > 3 | 0.78 | 0.127 |
| BFplus0 > 3 | 0.81 | 0.188 |
| BF0plus < 3 | 0.84 | 0.113 |
| BFrep0 >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

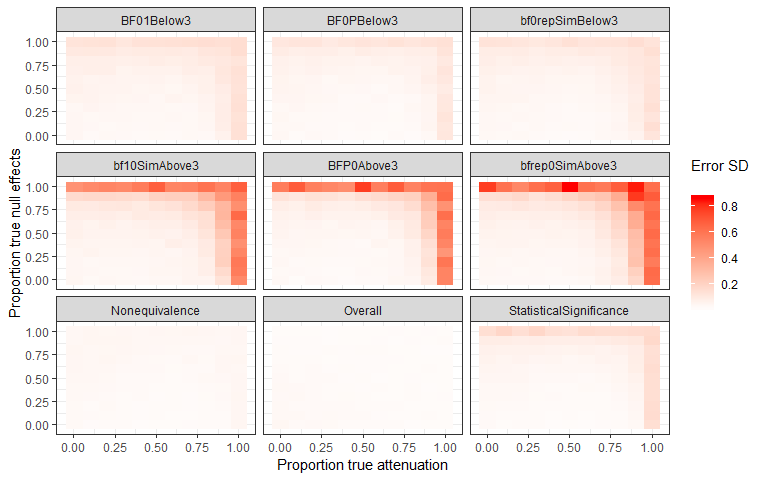
Additionally, simulations were performed using the same data-generation method to estimate the accuracy of the raw methods of estimating the simulated true proportion decrease under these different scenarios. Looking the mean proportion of effect size attenuation in the study, the results of the simulation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

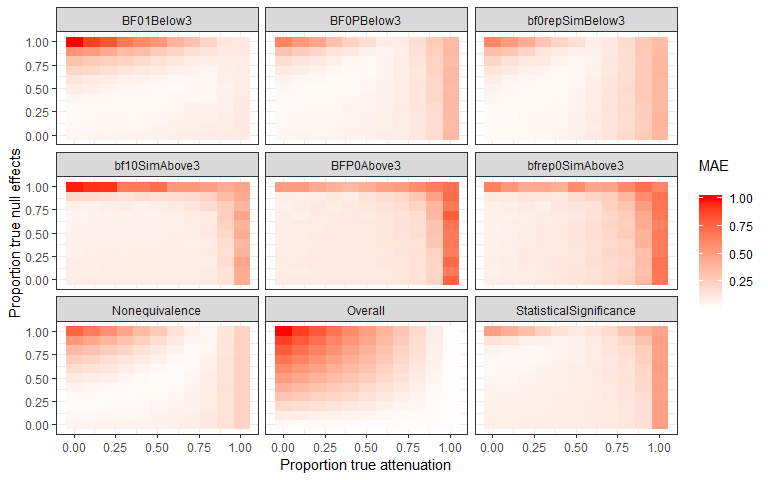
Note that these values are only valid under a specific data generation process, where there is a consistent factor effect size decrease, and where the studies which are null are random and independent of the original effect and sample sizes. See Table [simulation output] for the mean squared error (MSE),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

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

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







### SM4

### 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 | 0.69 | -0.07 | -0.06 | -0.06 | -0.06 | -0.05 |
| 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 | 0.00 | -0.04 | -0.04 | -0.04 | -0.04 | -0.03 |
| Significant in same direction | 0.00 | -0.06 | -0.05 | -0.05 | -0.05 | -0.04 |
| p value as Moderator | 1.00 | -0.09 | -0.08 | -0.08 | -0.08 | -0.08 |
| 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, excluding one replication project at a time, including only the sample indicated in “subsample”.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subsample | Proportion significant | Minimum estimate | 25th percentile | Median | 75th percentile | Maximum estimate |
| p value as Moderator | 1.00 | -0.10 | -0.09 | -0.08 | -0.07 | -0.07 |
| bf01<3 | 0.67 | -0.11 | -0.10 | -0.09 | -0.08 | -0.07 |
| bf0plus<3 | 0.89 | -0.11 | -0.10 | -0.09 | -0.08 | -0.07 |
| bf0Rep<3 | 0.44 | -0.08 | -0.07 | -0.06 | -0.05 | -0.04 |
| bf10>3 | 0.00 | -0.07 | -0.06 | -0.05 | -0.04 | -0.03 |
| bfplus0>3 | 0.11 | -0.07 | -0.06 | -0.05 | -0.04 | -0.03 |
| bfRep0>3 | 0.00 | -0.06 | -0.05 | -0.04 | -0.02 | -0.01 |
| 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 | -0.13 | -0.13 | -0.13 |

### Bayesian Mixture Model

The mixture model results presented in text presents the model developed by Camerer et al,. (2018; see <https://osf.io/xhj4d/> for their detailed description of this model). All priors were chosen to be uninformative or vague. The mixture model assumes that the observed replication effect sizes either come from the null hypothesis, a true effect sampled from a normal distribution with a mean of zero and a estimated precision (tau). This model uses an errors-in-variables approach to account for possible attenuation of effect sizes due to measurement error and estimation uncertainty following {Matzke, 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]  
 }  
}

### Supplementary materials [meta-moderaters]

Two methods were used to normalize the distribution of the p values, the Tukey-Freeman double Arcsine transform {Miller, 1978 #744}, and the The Ordered Quantile normalization transformation {Peterson, 2018 #1009}. Residual normality appeared to approximately hold in all cases, and as the results for all methods were functionally identical to those derived from those without any transformation only the raw results are presented in the main text. See tables SM11 [meta-moderators] for the results of the model with normalized predictors.

Table SM11.

Table SM12.