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 ns 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

#### Multilevel meta-analysis

All analysis was performed in R {R Development Core Team, 2018 #314} 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 of the random effects were obtained following {Robinson, 1991 #999}{Morris, 1983 #1000}.

Confidence intervals around binomial proportions are 95% Wilson Score intervals. Percentage change values were calculated using Fisher Z transformed effect sizes. All analyses were exploratory, and multiple models which were developed are not presented here, although I believe the presented results to be the best description of the data. 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.

#### 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 (or so small as to be effectivly null) effects. The average dissatenuation could be extremely high, and yet this effect be almost entierly driven by the presence of 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 caclulated using multiple methods of excluding original studies.

#### Statistical significance of the replication study

The first method 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 underpowered 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, this method is likely to underestimate the amount of effect size exaggeration. 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 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 or significant in the opposite direction {Lakens, 2017 #214;Lakens, 2018 #951}. 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. 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 single case. This method means that replication studies which found effects which were not statistically equivalent to the null were retained. However, as original sample sizes were often very small, the minimum detectable effect was occasionally quite high (mean = 0.17, SD = 0.12, quintiles = [0, 0.1, 0.15, 0.23, 0.74]).

##### Bayesian analysis

Three different types of Bayes factors were developed for each study following {Wetzels, 2012 #993} and {Wagenmakers, 2016 #994} using the correlation coefficients from each study. Bayes Factors express the relative evidence for one model compared to another, the degree to which a Bayesian observer should update their prior beliefs in response to the receipt of new data. 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 tailed () and and two tailed () alternative hypothesis. Replication Bayes Factors () were also developed, representing the relative evidence for or against the point null hypothesis as compared to the original study’s reported effect size. The order of the subscripts indicate whether the Bayes Factors represent evidence for the null (, , ) or for the alternative hypothesis (, , ). All of these Bayes factors were developed using only the transformed effect sizes and samples sizes reported in the Replication projects {Wagenmakers, 2016 #994}. 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 are not intended as anything more than a coarse estimate of the degree of evidence provided for and against the null model. See the supplementary material [!] table [bayesFactors] for a table showing the differences between the values returned by this method compared to those reported in the Bayesian supplement to {Camerer, 2018 #967}.

## Results

In the following I outline the raw decreases that are seen looking at all articles, the results of the methods of removing the null results from these raw figures, and finally the results of the multilevel models. The best estimates of the amount of effect size decrease that should be expected are developed from the multilevel models.

### Raw decreases

See Table 2 for a more comprehensive list of descriptives on the effect size differences seen.

Notably, there was an average decrease in effect sizes from the original to the replication study of 29%.

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 | 220 | 314 | 0.40 | 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 | 160 | 243 | 0.40 | 0.33 | 0.34 | 0.28 | -0.07 | -0.10 | -0.05 | -0.08 | 0.18 | -0.15 | -0.23 |
| BFRep0Above3 | 126 | 243 | 0.43 | 0.35 | 0.39 | 0.32 | -0.04 | -0.07 | -0.01 | -0.04 | 0.18 | -0.04 | -0.11 |
| BF01Below3 | 219 | 304 | 0.42 | 0.36 | 0.36 | 0.32 | -0.06 | -0.10 | -0.03 | -0.05 | 0.25 | -0.04 | -0.13 |
| BF10Above3 | 175 | 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 | -0.14 |
| BFP0Above3 | 184 | 304 | 0.41 | 0.35 | 0.40 | 0.35 | -0.01 | -0.04 | 0.02 | -0.01 | 0.21 | 0.08 | -0.05 |

### Statistical significant

### Equivilence tests

### BFs

#### 

#### 

#### 

### Multilevel models

## <https://osf.io/z7aux/>

## HAVE TO GO THROUGH AND REMOVE THOSE BASED ON

NOTE ! ! ! - it may be important to use dis-attenuated values from LOOPR because they used short form analyses

## Supplementary material

#### Table [BayesFactors]

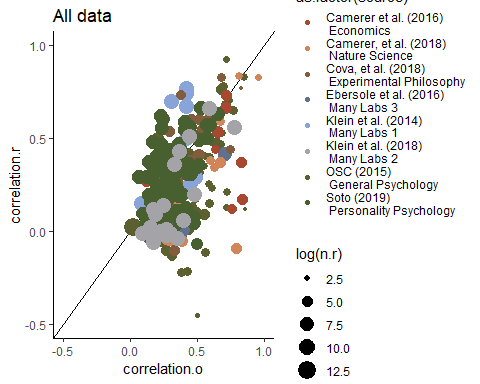
One-sided and () and replication () Bayes Factors for as reported in {Camerer, 2018 #967} and as estimated in the current paper, along with the reported correlation coefficients and sample sizes from the original and replicaiton studies.

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

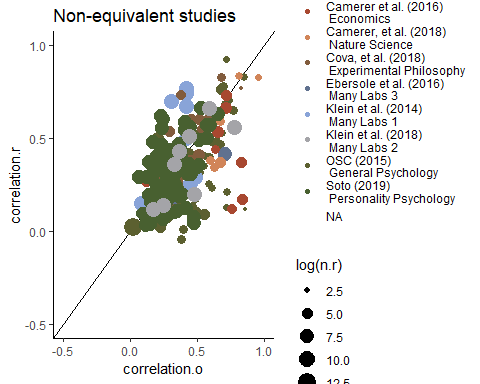
The only large discrepancy included is seen in Balafoutas and Sutter (2012) in which the Bayes Factor reported in {Camerer, 2018 #967} was based on a hypothesis test of orderd binomial probabilities, likely accounting for the large difference.

### Plots of the relationship between original and replication correlation coefficents, removing different sets of possibly null results

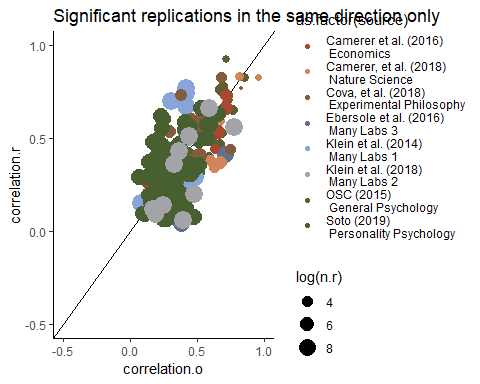
plotAllData



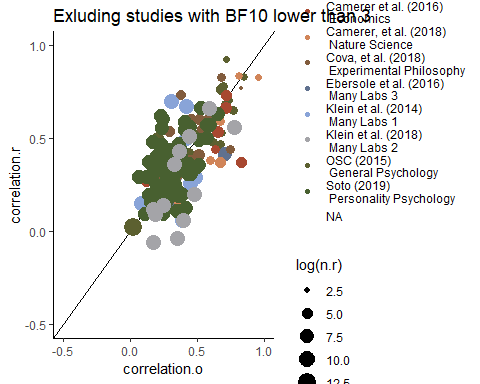
plotNonequiv



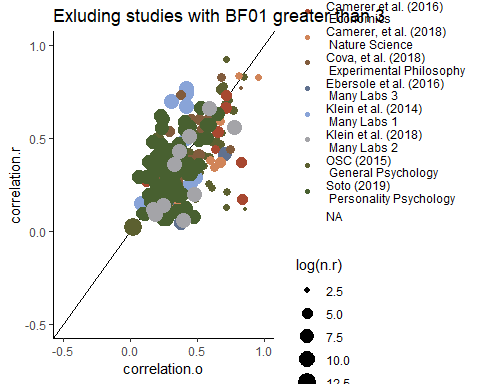
plotSigR



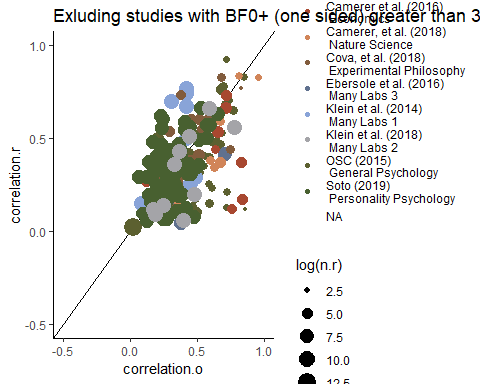
plotBF10Greater3



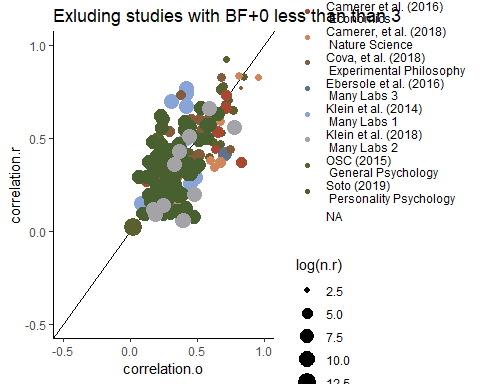
plotBF01Lesser3



plotBF0plusLesser3



plotBFPlus0Greater3



plotBFRep0Lesser3

