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 the 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 2 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 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.

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.

A second method we use is to exlude 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 equivilance 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. This test used the Fisher Z transformed scores and approximated the standard errors of each study as , except for studies from {Camerer, 2018 #967}, where meta-analyticly derrived standard errors were used. The large sample approximation was used assuming Fisher Z scores had a normal sampling distribuiton. As a method of testing how closely this method of approximating standard errors matches the origina replicaiton projects results, significance tests for the replication and original studies were performed using this approxiamtion. 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 equivilant to the null were retained. However, as original sample sizes were often very small, the minimum detectable effect was occasionally quite high mean = ()

Three different 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 new evidence. Two of the developed Bayes Factors ignore the original study and develop evidence entierly based on the replication study. Importantly, these bayes factors would differ from those that would normally be developed using the closest Bayesian equivilents to each original replicated study’s analysis, and are not intended as anything more than a corse estimate of the degree of evidence provided for and against the null model. One Bayes Factors and and replication Bayes Factors were developed using the transformed correlation coefficents from the original and replicaiton studies {Wagenmakers, 2016 #994}. One sided Bayes factors only used information from the replication study. 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}.

### What would be a reasonable blanket rule for an effect that no longer matters?

I.e., an effect that is .2 of the original effect?

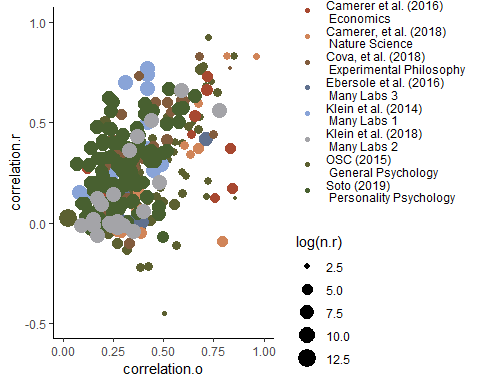
Bayesian test of r != 0 vs. r >/< 0, BF in which direction (assuming a flat or default prior)

All of the methods in the RPP study - original within bounds of 95% replication CI

EXCLUDE STUDIES - Find the smallest effect that would have been statistically significant in the original study and base the 95% CI on that. Lakens , Scheel and Isager 2018 P 262

## Testing

ggplot(allData, aes(correlation.o, correlation.r,size = log(n.r), colour = as.factor(source))) + geom\_point(alpha = 1, na.rm = T)+ ochRe::scale\_colour\_ochre(palette = "tasmania") + theme\_classic() + ylim(c(-.5, 1))+ xlim(c(-.0, 1)) + scale\_shape\_manual(values = c(8,16))



# summary(abs(loopr$correlation.o))   
mean((allData$correlation.r-allData$correlation.o)/allData$correlation.o, na.rm = T)

## [1] -0.3000678

mean((allData$fis.r -allData$fis.o)/allData$fis.o, na.rm = T)

## [1] -0.2944172

# cor.test(allData$correlation.r, allData$correlation.o, na.rm = T)  
  
# allData

### Testing methods of removing originals

# extracting just the significant ones (i.e., successful replications according to p < .05 on the replication)

### 

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Included Articles | Mean proportion change | Mean replication ES | Mean original studies ES | Mean ES difference | SD difference | Median proportion change | Median replicaiton ES | Median original ES | Median difference | n included | n criteria calculable for | 95% CI LB Mean ES Change | 95% CI UB Mean ES Change | Number included in model | Estimated decrease model | 95% CI LB Mean ES change model | 95% CI UB Mean ES change model |
| Overall | -0.29 | 0.28 | 0.40 | -0.13 | 0.26 | -0.35 | 0.21 | 0.34 | -0.11 | 314 | 314 | -0.16 | -0.10 | 305 | -0.14 | -0.21 | -0.07 |
| StatisticalSignificance | 0.03 | 0.40 | 0.43 | -0.02 | 0.20 | -0.07 | 0.34 | 0.37 | -0.04 | 220 | 314 | -0.05 | 0.00 | 195 | -0.05 | -0.12 | 0.01 |
| Nonequivalence | -0.07 | 0.36 | 0.43 | -0.07 | 0.24 | -0.16 | 0.31 | 0.36 | -0.06 | 237 | 305 | -0.10 | -0.04 | 235 | -0.08 | -0.16 | -0.01 |
| BF0RepBelow3 | -0.15 | 0.35 | 0.43 | -0.07 | 0.18 | -0.23 | 0.29 | 0.34 | -0.08 | 160 | 243 | -0.10 | -0.05 | 160 | -0.09 | -0.15 | -0.03 |
| BFRep0Above3 | -0.04 | 0.41 | 0.45 | -0.04 | 0.18 | -0.11 | 0.33 | 0.36 | -0.04 | 126 | 243 | -0.07 | -0.01 | 126 | -0.07 | -0.13 | 0.00 |
| BF01Below3 | -0.04 | 0.38 | 0.45 | -0.06 | 0.26 | -0.13 | 0.33 | 0.37 | -0.05 | 219 | 304 | -0.10 | -0.03 | 217 | -0.09 | -0.16 | -0.01 |
| BF10Above3 | 0.08 | 0.42 | 0.44 | -0.01 | 0.21 | -0.04 | 0.37 | 0.37 | -0.01 | 175 | 304 | -0.05 | 0.02 | 173 | -0.05 | -0.11 | 0.02 |
| BF0PBelow3 | -0.04 | 0.37 | 0.44 | -0.07 | 0.25 | -0.14 | 0.32 | 0.37 | -0.05 | 230 | 304 | -0.10 | -0.03 | 228 | -0.09 | -0.16 | -0.02 |
| BFP0Above3 | 0.08 | 0.42 | 0.44 | -0.01 | 0.21 | -0.05 | 0.37 | 0.37 | -0.01 | 184 | 304 | -0.04 | 0.02 | 182 | -0.05 | -0.11 | 0.02 |

TOST - Exclude inconclusive ones -

Bayesian modelling - compare data for and against

## <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]

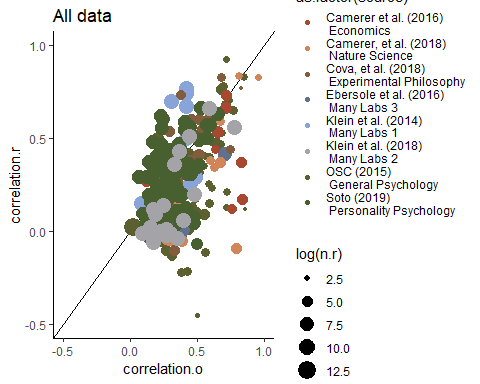
One-sided and () and replication () Bayes Factors for as reported in {Camerer, 2018 #967} and as estimated in the current paper, along with the reported correlation coefficients and sample sizes from the original and replicaiton studies.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 | 5.4000e-01 | 0.31 | 2.65 | 2.06 | 1.04 |
| Aviezer et al. (2012), Science | 0.96 | 15 | 0.83 | 14 | 4.4600e+02 | 57.00 | 226.82 | 274.00 | 137.04 |
| Balafoutas and Sutter (2012), Science | 0.28 | 72 | 0.15 | 243 | 4.2000e+00 | 4.27 | 4.12 | 2.12 | 1.07 |
| Derex et al. (2013), Nature | 0.52 | 51 | 0.36 | 65 | 3.1440e+03 | 3701.00 | 33.42 | 21.78 | 10.91 |
| Duncan et al. (2012), Science | 0.67 | 15 | 0.37 | 128 | 2.6820e+03 | 2513.00 | 2047.92 | 2335.00 | 1167.51 |
| Gervais and Norenzayan (2012), Science | 0.29 | 57 | -0.04 | 755 | 6.0000e-02 | 0.03 | 0.03 | 0.02 | 0.09 |
| Gneezy et al. (2014), Science | 0.22 | 178 | 0.18 | 407 | 2.3190e+02 | 485.27 | 472.06 | 113.00 | 56.51 |
| Hauser et al. (2014), Nature | 0.82 | 40 | 0.83 | 22 | 2.8890e+03 | 10211.00 | 101945.01 | 24633.95 | 12317.00 |
| Janssen et al. (2010), Science | 0.63 | 63 | 0.34 | 42 | 5.8600e+00 | 0.00 | 1.90 | 4.18 | 2.12 |
| Karpicke and Blunt (2011), Science | 0.60 | 40 | 0.38 | 49 | 1.4550e+01 | 11.82 | 13.99 | 13.05 | 6.55 |
| Kidd and Castano (2013), Science | 0.27 | 86 | -0.04 | 999 | 5.0000e-02 | 0.01 | 0.01 | 0.02 | 0.08 |
| Kovacs et al. (2010), Science | 0.45 | 24 | 0.59 | 95 | 5.5700e+07 | 132389304.00 | 86776199.93 | 54424990.81 | 27212495.42 |
| Lee and Schwarz (2010), Science | 0.39 | 40 | -0.05 | 409 | 8.0000e-02 | 0.01 | 0.02 | 0.03 | 0.11 |
| Morewedge et al. (2010), Science | 0.45 | 32 | 0.35 | 89 | 8.6870e+01 | 157.85 | 164.64 | 80.78 | 40.41 |
| Nishi et al. (2015), Nature | 0.20 | 200 | 0.12 | 480 | 7.0500e+00 | 7.77 | 8.39 | 2.87 | 1.44 |
| Pyc and Rawson (2010), Science | 0.38 | 36 | 0.15 | 438 | 6.8300e+00 | 4.04 | 16.79 | 15.99 | 8.00 |
| Ramirez and Beilock (2011), Science | 0.79 | 20 | -0.09 | 105 | 1.4000e-01 | 0.00 | 0.00 | 0.07 | 0.19 |
| Rand et al. (2012), Nature | 0.14 | 343 | 0.03 | 3150 | 1.4000e-01 | 0.10 | 0.13 | 0.13 | 0.07 |
| Shah et al. (2012), Science | 0.27 | 56 | -0.04 | 897 | 7.0000e-02 | 0.04 | 0.04 | 0.02 | 0.08 |
| Sparrow et al. (2011), Science | 0.37 | 69 | 0.07 | 338 | 1.5000e-01 | 0.03 | 0.06 | 0.26 | 0.15 |
| Wilson et al. (2014), Science, | 0.67 | 30 | 0.59 | 39 | 6.0125e+02 | 1871.00 | 1926.34 | 830.41 | 415.22 |

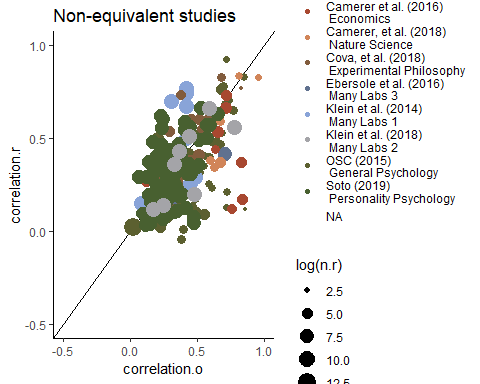
The only large discrepancy included is seen in Balafoutas and Sutter (2012) in which the Bayes Factor reported in {Camerer, 2018 #967} was based on a hypothesis test of orderd binomial probabilities, likely accounting for the large difference.

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

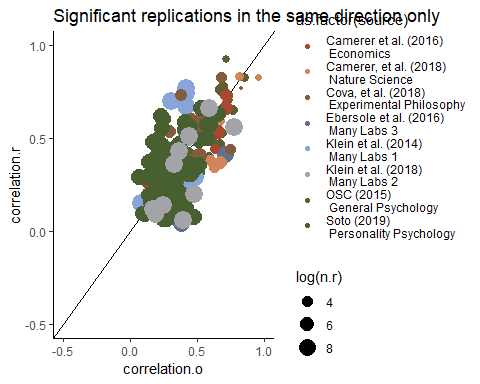
plotAllData



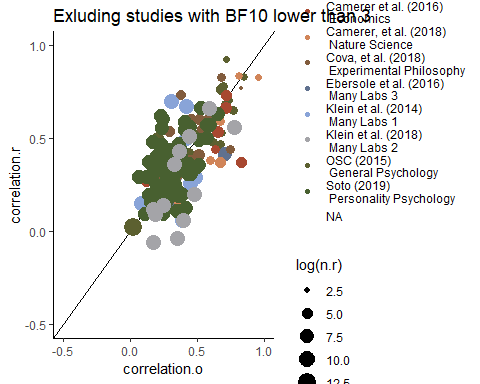
plotNonequiv



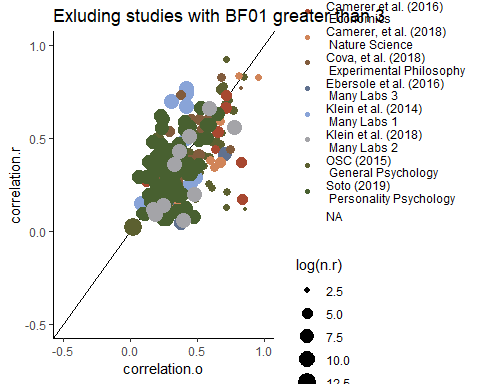
plotSigR



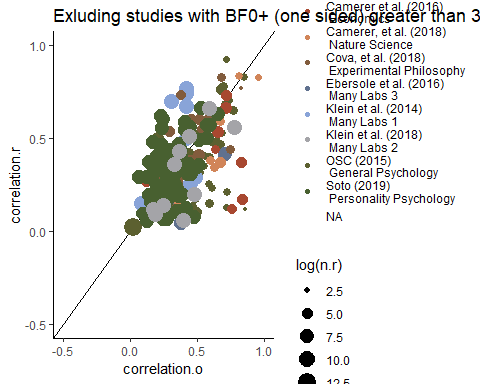
plotBF10Greater3



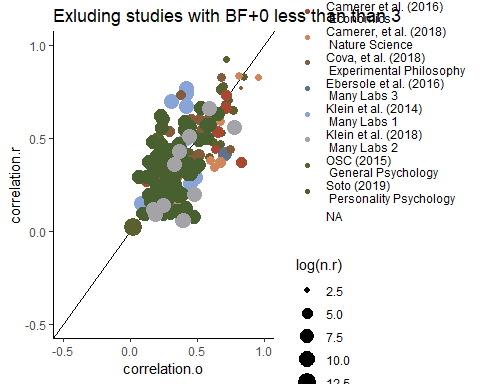
plotBF01Lesser3



plotBF0plusLesser3



plotBFPlus0Greater3



plotBFRep0Lesser3

