Estimating the effect of publication and reporting bias

ADD IN BFs > 10 and Equivalence test < .1

## Introduction

See word document

## Methods

### Data extraction

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

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

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

INSERT TABLE 1 HERE

### Analysis

#### 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 effects (or effects so small as to be effectively null). The average disattenuation could be extremely high, and yet this effect be almost entirely driven by the presence of effectively-null effects. This aspect becomes especially important as the sampling of the literature is non-random, meaning it is plausible that some effects were chosen for replication to a greater or lesser extent as it was expected that they may not replicate. In order to account for this issue, the average effect size attenuation was calculated using multiple methods of excluding original studies. Two methods are presented here, using equivalence testing and using the statistical significance of the replication study.

##### 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 under-powered to detect a non-null but true effect are likely to be excluded from this analysis. Especially as in some of the replication projects the sample size in the second study, 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, 0th, 25th, 50th, 75th and 100th quintiles = [0, 0.1, 0.15, 0.23, 0.74]), suggesting that in some cases the original study may have been massively under-powered to detect even large effects. Ideally, a full reanalysis would be performed for each original study. However, it was not feasible to extract and reperform full analyses for the over 600 total original and replication studies.

#### Simulation studies

All methods function by removing studies which have low effect sizes, so it is a forgone conclusion that the 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 these methods estimate the amount of effect size attenuation under reasonable assumptions, see supplementary materials [simulation] for a full description of the simulation study output. In addition to the two methods presented here, approximate Bayes factors were developed based on the reported correlation coefficient transformed effect sizes and a Bayes factor cut score was used to exclude studies. However, because of the information available on each of the original and replication studies, and the approximate nature of the estimation method that was used (also relying on the correlation coefficients reported in the replication projects), these results were not considered accurate enough to be useful. In fact, in many cases they led to worse estimates of the true proportion of effect size attenuation in many cases (as estimated by (replication effect size-original effect size)/original effect size). For completeness, the results of all tests performed using cut scores are reported in supplementary materials [!]. None of the results of these methods would lead to substantially different conclusions being drawn.

The results of the simulations 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 (i.e., the simulation studies show a Mean Absolute Error of 0.25, 0.15, 0.13 for estimates of the proportion of attenuation seen excluding no studies, those which were not significant and using equivalence testing respectively, over true attenuation and null hypothesis rates of 0 to 1). However, at reasonable levels of attenuation and proportion of null effects being correct, these methods seem to be much more reliable (e.g., excluding simulations with a proportion of null results or attrition of .8 or greater, these methods have a mean absolute errors of 0.23, 0.08, 0.06 for estimates of the mean proportion of attenuation seen when including all studies, excluding non-significant studies, and exlculding statistically-equivilent to the null studies respectivly).

### Leave one out cross validation

There is no accepted default manner of Leave One Out (LOO) cross validation for MLM, so in order to assess whether the main results of this study are sensitive to the inclusion of each of the replication projects and individual findings within each replication project, the models were rerun using leave one out cross validation, excluding both the individual replication attempts and the replication projects one at a time. When leaving out individual studies the range of point estimates (i.e., the difference between the smallest and largest estimate of the difference between original and replication studies) for each of the LOO cross validation models did not exceed more than a Fisher z sore of 0.02. When excluding one replication project at a time, model estimate ranges did not exceed 0.05. See supplementary material [!] for a table of the proportion of model estimate p values below .05, and estimate quintiles for each model from the leave on out cross validation on the study and project levels. None of these changes would lead to substantially different conclusions being drawn from the model output.

## 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 ignoring grouping

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

plotAllData

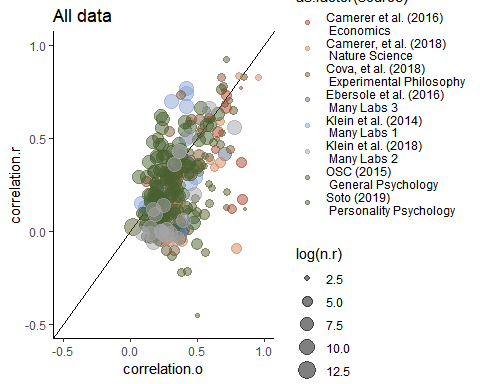


Figure 1. A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes. Points which fall on the the solid, diagonal line represent replication effect sizes equal to the original effect sizes. Point size represents (the log) of the number of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

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n included | n criteria calculable for | Mean original ES | Median original ES | Mean replication ES | Median replicaiton ES | Mean ES difference | 95% CI LB Mean ES Change | 95% CI UB Mean ES Change | Median ES difference | SD difference | Mean proportion change | Median proportion change |
| Overall | 314 | 314 | 0.38 | 0.33 | 0.27 | 0.20 | -0.13 | -0.16 | -0.10 | -0.11 | 0.25 | -0.29 | -0.35 |
| StatisticalSignificance | 219 | 314 | 0.41 | 0.35 | 0.38 | 0.32 | -0.02 | -0.05 | 0.00 | -0.04 | 0.20 | 0.03 | -0.07 |
| Nonequivalence | 237 | 305 | 0.41 | 0.35 | 0.35 | 0.30 | -0.07 | -0.10 | -0.04 | -0.06 | 0.24 | -0.07 | -0.16 |

The same pattern of results are seen in every

### Examining only the statistically significant studies

Looking at the 219 replications in which the replication study was statistically significant, the average effect for original studies was 0.41, and the mean effect size for replication studies was 0.38. There was an average decrease of r = -0.02 (naive 95% CI [-0.05, 0, an average decrease of 3.47%. Removing studies which are not statistically significant, the average effect size decrease is of a smaller magnitude.

Of course, using this rule for excluding studies is likely to lead to a under-exaggeration of the issue, 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

### Equivilence tests

### BFs

#### 

Excluding just those studies which provide at least moderate evidence for the null paints a similar picture,

#### 

#### 

### Multilevel models

## Results

#### Limitiations:

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

All of the methods that were used to the replication studies which were null-or-effectively were bound to decrease the amount of effect size decrease that is seen. At worst, they could be seen as just removing the studies which happened to find low effects as opposed to removing all the true null hypotheses. However, this preliminary analysis does provide suggestive evidence that the degree of effect size attenuation that is seen may be largely att

## <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

## Alternative removal methods

##### Approximate bayes factors

Three different types of Bayes factors were developed for each study using default priors following {Wagenmakers, 2016 #994} using the effect sizes converted into correlation coefficients from each study. 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. 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 and Bayes factors from 3-10 provide “substantial” evidence {Jeﬀreys, 1961 #1001}.

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, in which the prior for the replication correlation coefficient is the posterior based on the original research {Wagenmakers, 2016 #994}. This papers 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 (, , ). 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 table [bayesFactors] 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}. See Table [all estimates output] for a list of all of the model and raw estimates output using all tested cut scores. None lead to substantially different conclusions.

#### 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 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 |

These results were not considered sufficiently accurate, and alongside the results of the following simulations, it seemed reasonable to base the main inferences of this paper on the results of the frequentest analyses which had preferable properties.

### Simulation of removal methods

In order to assess whether the methods that were used to estimate the proportion change in studies excluding null results develop reasonable estimates, a series of simulations were performed. Simulations took as a starting point the observed effects in the original studies, estimating a true effect from these original results based on the Fisher Transformed 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 be 0 randomly. Notably, 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 scenario, 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 at each level of detail.

##### Table [all estimates output]

A table of all of the output from

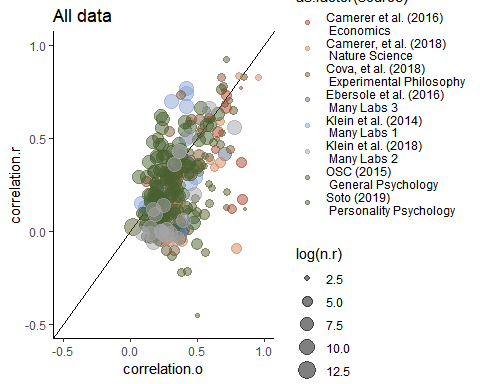
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | n included | n criteria calculable for | Mean original ES | Median original ES | Mean replication ES | Median replicaiton ES | Mean ES difference | 95% CI LB Mean ES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

##### Table [all model output]

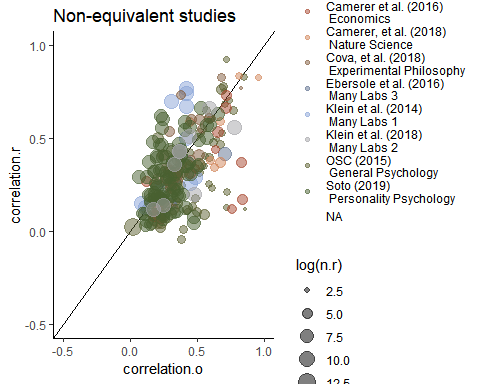
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | modelN | modelEstimate | MLM95lb | MLM95ub |
| Overall | 305 | -0.14 | -0.21 | -0.07 |
| StatisticalSignificance | 194 | -0.05 | -0.12 | 0.01 |
| BF0PBelow3 | 228 | -0.09 | -0.16 | -0.02 |
| BFP0Above3 | 182 | -0.05 | -0.11 | 0.02 |
| BF01Below3 | 217 | -0.09 | -0.16 | -0.01 |
| BF10Above3 | 173 | -0.05 | -0.11 | 0.02 |
| BF0RepBelow3 | 160 | -0.09 | -0.15 | -0.03 |
| BFRep0Above3 | 126 | -0.07 | -0.13 | 0.00 |
| Nonequivalence | 235 | -0.08 | -0.16 | -0.01 |

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

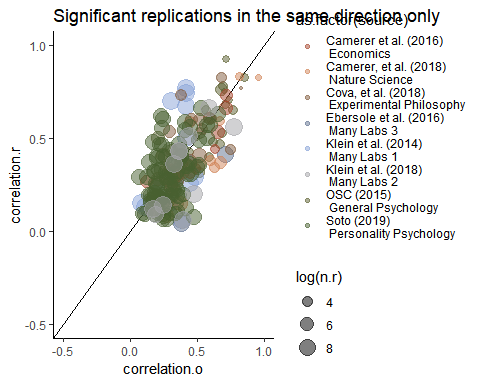
plotAllData



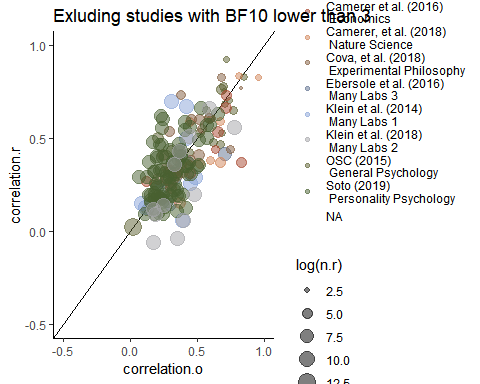
plotNonequiv



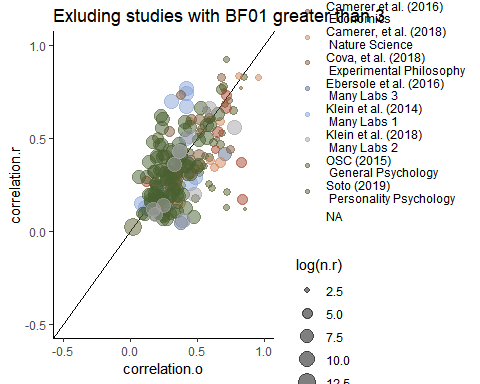
plotSigR



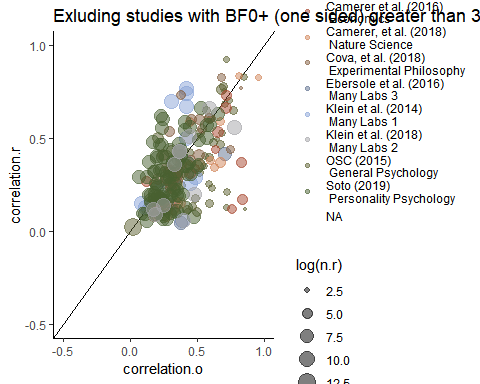
plotBF10Greater3



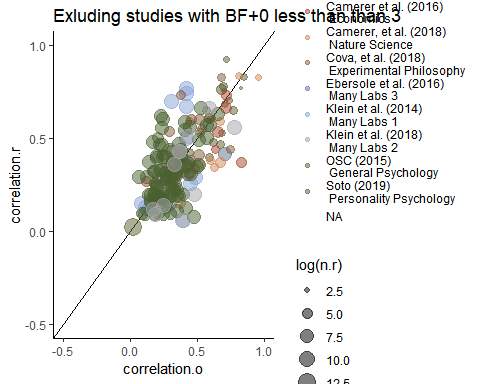
plotBF01Lesser3



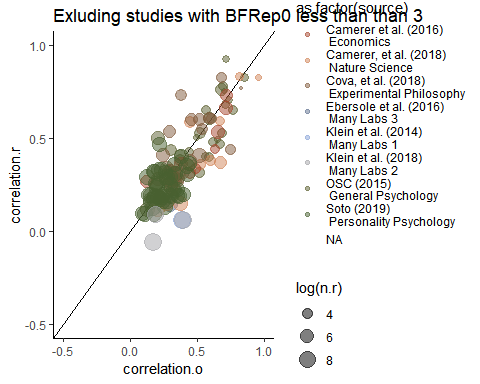
plotBF0plusLesser3



plotBFPlus0Greater3



plotBFRep0Lesser3



### LOO Cross validation output