Examining the Impact of Publication and Reporting Biases on Effect Sizes in Published Behavioural Sciences Research Using Large Scale Replication Projects

F. Singleton Thorn, P. Dudgeon

University of Melbourne, School of Psychological Sciences

Author note:

This research was supported by an Australian Government Research Training Program Scholarship awarded to F. Singleton Thorn.

Correspondence concerning this article should be addressed to F. Singleton Thorn, Faculty of Medicine, Dentistry and Health Sciences, School of Psychological Sciences, University of Melbourne, Parkville, Victoria, 3010, Australia. Email: [fsingletonthorn@gmail.com](mailto:fsingletonthorn@gmail.com) or felixs@unimelb.edu.au

**Abstract**

This paper examines 306 replication studies from eight large-scale replication projects – projects that directly replicated a set of published studies – to present a preliminary estimate of the degree of effect size attenuation between original and replication studies. This value represents the cumulative impact of publication bias, reporting bias, and any other systematic decreases in effect sizes that might be seen in replication studies. We estimate that effect sizes are, on average, smaller in replication studies by r = -0.14 (95% CI [-0.07, -0.2]), equivalent to a Cohen’s d of -0.28 (95% CI [-0.14, -0.42]) or a decrease equivalent to 34% (95% CI [17%, 51%]) of the mean original effect size, and also show that there is substantial heterogeneity across the article, effect and replication project level. Using a Bayesian mixture model to account for the presence of null effects we also estimate that effect sizes are, on average, 20% smaller in replication studies (95% HPDI [11%, 28%]) conditional on the effect under study being non-zero. Researchers should be aware that effect sizes in the published literature are likely to be exaggerated and account for this when planning, reading and interpreting research.

*Keywords*: Publication bias, effect size, QRPs, metascience, metaresearch

Estimating the Effect of Publication and Reporting Biases on Effect Sizes in Published Behavioural Sciences Research using Lage Scale Replication Projects

This paper uses the results of over 300 replication studies conducted as a part of eight large-scale replication projects (henceforth ‘replication projects’) to model the effect size change between original and replication studies in the behavioural sciences research literature. Although the presence or absence of effects may be an interesting question in of itself, an understanding of the magnitude of effects is essential to accurately interpreting an effect and for planning future research effectively. As such the discovery and precise estimation of effects is essential to developing a coherent and reliable scientific literature. Under conditions where results are selectively reported or analysed based on characteristics related to the size of the effect (e.g., statistical significance), the literature no longer provides an unbiased estimate of true outcome effect sizes (Hedges, 1992). The degree to which publication and reporting biases inflates effect sizes in the behavioural sciences literature is currently unknown. Recent large-scale replication projects – projects that have systematically replicated bodies of research – provide a new resource with which to begin to estimate the extent of effect size inflation in the behavioural sciences literature.

All of the included replication projects were primarily conducted in order to assess the degree to which their particular area of research contains results that are not replicable, or to estimate variability in effects among subpopulations. See Table 1 for a list of the included replication projects, the percentages of replication attempts with a statistically significant result in the same direction as the original study, and the number of studies from each project included in the current analysis.

Table 1.

*A list of each included replication project, the number of replication studies performed as a part of each replication project, the percentage of replication studies that were “successful” (defined here as replication studies that found statistically significant in the same direction as the original study), the number of studies for which are included in the current study, and the percentage of each project’s studies that are included in the current analysis.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Replication projects | Number of replication studies performed | Replication results statistically significant in the same direction as original (%) | Included studies | Percent of performed replication studies included in the current project |
| 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 | 100% |
| 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 | 100% |
| 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 | 89% |
| 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 | 89% |
| Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., BahnÃ­k, Å ., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating Variation in Replicability. Social Psychology, 45(3), 142-152. <doi:10.1027/1864-9335/a000178> a | 16 (13 effects) | 88% (85%) | 15 | 94% (92%) |
| 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 | 79% |
| Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. <doi:10.1126/science.aac4716> | 97 | 36% | 89 | 92% |
| Soto, C. J. (in press). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes Of Personality Replication Project. *Psychological Science*. | 121 | 86% | 100 | 83% |
| All projects | 347 | 65% | 306 | 88% |

Note: a Klein et al. (2014) includes 4 operationalisations of a single effect that were input separately for analysis in the current study, the bracketed values in the table refer to the number of results at the effect level. aSoto’s (in press) replication rate was recalculated on the “effect” level (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. Soto (in press) uses as its primary unit of analysis trait-outcome associations, and found that 87% of the 78 trait-outcome associations were supported by a significant result in the same direction.

There are several possible causes of effect size attenuation in replications. One possible cause of effect size attenuation is systematic deviation from the original study’s protocol in replication studies. It’s imaginable that replication teams tend to operationalise the original study’s methods in ways that cause a reduction in effect sizes, possibly in subtle ways that may not be immediately obvious from replication reports. Tacit knowledge that the original study’s authors have access to but that which may not be written in the study’s methods section may mean that effect sizes tend to be smaller in replication studies.

The combination of regression to the mean, selective reporting, and publication bias, all of which are be exacerbated by low average statistical power would also account for a reduction in effect sizes in replication studies. Publication bias is the process by which studies that report finding results supporting their hypotheses, usually by showing statistically significant results, are more likely to be published than those that do not ("The importance of no evidence," 2019; Lane & Dunlap, 1978; Mahoney, 1977; Sterling, 1959). If studies are more likely to be published when they show statistically significant results, effect sizes in the literature will be, on average, exaggerated, and the number of false positives (i.e., true null effects showing statistically significant results) increased (Lane & Dunlap, 1978). The extent to which effect sizes are exaggerated in a literature depends, primarily, on the degree to which non-significant results are excluded, the true statistical power of studies (a product of the effect size and sample size of studies given their experimental design and analysis strategy) and the proportion of true nulls being investigated (Hedges, 1992; Oakes, 1986).

Publication bias towards statistically significant results appears to be particularly acute in the behavioural sciences research literature, with an estimated 92% of papers reporting a significant finding (Fanelli, 2012) and around 75% of all reported p values being below .05 (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016), despite the low estimated average power of psychological research (Cohen, 1962; Szucs & Ioannidis, 2017). There is also direct evidence of publication bias from studies in the biomedical and social sciences literature that have tracked research projects from before data collection and show that studies that find statistically significant effects are more likely to be published than those that do not (Dwan, Gamble, Williamson, & Kirkham, 2013; Franco, Malhotra, & Simonovits, 2014).

Selective reporting among measures and QRPs like p-hacking and hypothesising after the results are known (HARKing) on the basis of statistical significance or the size of effects (Kerr, 1998) can also lead to effect sizes being exaggerated and increase the proportion of false positives in a scientific literature (Bakker, van Dijk, & Wicherts, 2012; Murphy & Aguinis, 2017; Simmons, Nelson, & Simonsohn, 2011). There are several reasons to think that QRPs and publication bias are prevalent in the psychology literature. Recent self-report surveys of psychologists suggest that questionable research practices like HARKing and p-hacking are common across countries and fields of psychological research (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Fiedler & Schwarz, 2015; John, Loewenstein, & Prelec, 2012). The observed correlation between effect sizes and sample sizes (i.e., “small study effects”) also suggests studies that do not report statistically significant effects either remain unpublished or that QRPs may be used to ensure that large-enough effect sizes are obtained to reach statistical significance (Button et al., 2013; Egger, Smith, Schneider, & Minder, 1997).

Given the evidence of practices that should cause effect size exaggeration among studies in the published psychology literature, two important questions follow. Firstly, how much are effect sizes in the literature inflated, and secondarily, how much variability is there in the degree to which effect sizes are inflated across studies? We have identified only two previous studies that attempt to empirically estimate the degree to which effect sizes in the published psychology literature are exaggerated. Stanley, Carter, and Doucouliagos (2018) used several meta-analytic bias-adjustment methods (specifically using three different estimators; WLS, WAAP-WLS, and PET-PEESE) in a reanalysis of 200 meta-analyses published in Psychological Bulletin. They found a median effect size exaggeration of 8 to 15%, with the range depending on the meta-analytic bias reduction method used. As Stanley et al. (2018) point out, their study estimates the effect of publication bias using a likely unrepresentative literature, and it is unclear how much this estimate is likely to generalize outside of the pages of Psychological Bulletin. Schäfer and Schwarz (2019) showed that effect sizes found in preregistered studies are, on average, much smaller (with a median correlation of 0.16) than those reported in non-preregistered published studies (with a median correlation of 0.36) and suggest that this difference is likely to be at least in part driven by the use of QRPs in non-preregistered studies. However, as Schäfer and Schwarz note, it is more likely than not that this decrease is primarily caused by systematic differences between the types of effects being investigated in preregistered and non-preregistered studies (Schäfer & Schwarz, 2019).

In order to examine the change in effect sizes between original and replication studies and estimate the degree of heterogeneity seen in this body of literature, the current study presents an exploratory analysis of recent replication projects using two main analytic approaches. The first analysis uses a multilevel meta-analytic framework to estimate the expected effect size change between original and replication studies, to quantify the degree of heterogeneity seen across the article, effect, and project levels. As this database is likely to include effects that are true null effects (or effects that are so close to true null effects as to be practically dismissible), remaining analytic approaches examine the degree to which effect sizes reported in the literature are exaggerated when the effect under study is non-null. This later value is arguably of more use to people attempting to plan studies based on the published literature, where the accuracy of sample size planning efforts depends on the unknown true effect size. To estimate this quantity, we use simple data exclusions (Analyses 2 and 3) and the Bayesian Mixture Model presented in Camerer et al. (2018, analysis 4).

In reading the paper, it is important to note that these replication projects have not replicated a random selection of effects from the literature. As such the analyses here do not allow us to make simple inferences about what would be seen in a future replication study (i.e., predicting the effect size decrease between a randomly selected psychology research article and its replication) or directly examine the impact of publication and reporting biases on effect sizes. This analysis could therefore be read as producing estimates of the differences we would expect to see in replication studies performed as part of large-scale replication projects in the behavioural sciences literature, under the assumption that the replication projects included here are a random sample of replication projects.

## Methods

### Disclosures

All analyses were exploratory, and additional models which were developed and considered 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. See <https://osf.io/daj8b> for the preregistration of this project specifying the research questions to be addressed and the general analytic approach used for Analyses 1 - 3. All of the data and analysis code used in this study, and an RMarkdown document to allow the current paper to be reproduced, are available from <https://osf.io/7qvna/>.

### Data Extraction

All eight published or in-press large scale replication projects performed within in the behavioral science research literature were included in the current study (see Table 1 for a list of the included studies). 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, Ebersole et al., 2016; Klein et al., 2014). In these cases, these values were manually extracted from the original articles. When sample sizes for original studies were not reported in the data provided by each replication project they were manually extracted from original articles where possible.

For all analyses, the original and replication effect sizes were transformed to Fisher z-transformed correlation coefficients following the methods used in Open Science Collaboration (2015, see Supplementary Materials 5 for details). This conversion used data from the replication project whenever possible (i.e., whenever effect sizes were reported in correlation coefficients in a summary table or in a project’s online data this was directly converted to Fisher z values). If the study-level results were not reported as correlation coefficients, Cohen’s d values, as t-tests, or as F statistics in the original or replication project we excluded the result from this analysis (e.g., cases when no effect size was reported in the original study or in the replication project data set). In cases where sample sizes were not reported per group, sample sizes among groups were assumed to be equal in these conversions. For each of the Many Labs projects the top level result was used (i.e., the results of the analysis that collapsed the data across the multiple labs). 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, along with sample sizes for original and replication studies, was extracted for a total of 306 pairs of studies, excluding a total of 41 study pairs. See Table 1 for the number of valid studies extracted from each project.

### Analysis

All analyses were performed in R version 3.5.1 (R Development Core Team, 2018) and meta-analyses were performed using the Metafor package version 2.1 (Viechtbauer, 2010) using restricted maximum-likelihood estimation. 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 coefficients for easy interpretation (unless otherwise stated).

#### Analysis 1: Multilevel random effects meta-analysis.

The first approach uses a random effects meta-analysis framework to estimate the expected effect size difference between original and replication studies.

This analysis treats each pair of effects, the original and replicated effect sizes, as one “study” in a meta-analytic framework. This model estimates the change from the original to the replication study effect size () with a fixed intercept (), a random effect for replication project (), a random effect for each original article (), and a random effect for each individual replication (). Random effects at the project level were included to account for non-independence between replications from each replication project. Random effects at the original article level were included to account for cases when multiple effects from an original article were replicated or where multiple operationalisations of an original effect were tested. 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. This standard error is an approximation for the studies that reported F tests with a greater than 1 and chi-squared tests. In order to check whether this was strongly impacting results all multilevel meta-analyses were re-performed excluding these studies. No differences in the substantive interpretation of results would follow from this change (i.e., the intercept coefficient and random effects variance estimates changed by less than 0.003).

Using the aggregate summary statistics from the replication projects where a set of labs conducted replications (e.g., the Many Labs Projects) may underestimate the standard error of the difference scores (as their standard error is also a function of the unknown heterogeneity across labs). As a sensitivity analysis, we also ran all multilevel models using a conservative estimate of their sampling variance - calculating their standard errors using the mean sample size included in each replication study as opposed to the total sample size. Again, no differences in the substantive interpretation of results would follow from this change, with the coefficient estimates and estimates of the variance of the random effects changing by less than 0.005.

#### Accounting for null results.

In assessing the degree to which effects are attenuated between original and replication studies it is important to ask how much this effect is driven by the presence of a subset of replication studies where the null hypothesis is true. The average effect size difference between original studies and their replications could be extremely high, and yet this effect could be entirely driven by cases where the null hypothesis is true. For example, if 50% of studies had true null hypotheses, and yet all non-null replication effects were identical to those reported in the original articles, the average attenuation would be 50% despite the fact that the non-null original effect sizes provided unbiased estimates of the replication effect sizes. Analyses 2 to 4 were performed in order to account for this issue.

#### Analysis 2 and 3: Multilevel random effects meta-analysis with exclusions.

Analyses 2 and 3 reperform the above meta-analysis excluding studies using two exclusion criteria. Analysis 2 excluded studies in which the replication study was not statistically significant with an effect in the same direction as the original (using the *p* value reported in the replication projects’ datasets, at an alpha of .05, and using two-tailed tests where applicable). Analysis 3 removed effects in which the replication study effect was “statistically equivalent” to the null according to an equivalence test.

Analysis 2, excluding studies in which the replication study was not significant, means that replication studies with a low level of statistical power to detect the (unknown) true replication effect size are likely to be excluded. This may lead to this analysis underestimating the amount of effect size exaggeration, as replications with non-zero but small effect sizes are likely to be non-significant. This issue is compounded by the fact that some of the replication projects chose the sample sizes that were used in the replication studies using a power analysis of the observed effect in the original study (Open Science Collaboration, 2015). This approach to designing the replication studies means that if effect sizes are, on average, smaller in the replication studies than the original reported result, replication studies will often be underpowered.

In order to avoid excluding under-powered studies erroneously, Analysis 3 excluded studies based on whether the replication study results were statistically equivalent to the null hypothesis or statistically significant in the opposite direction (Lakens, 2017; Lakens, Scheel, & Isager, 2018). A requirement for equivalence testing is that an equivalence bound is selected (i.e., an effect size below which the effect size is said to be for all practical purposes equal to zero). For this, we used the lowest effect size that would have been statistically significant in the original study (assuming an alpha of .05), following a suggestion in Lakens et al. (2018). Equivalence tests were performed using Z tests of the Fisher z-transformed effect sizes, excluding studies where the observed replication effect was significantly smaller than the equivalence bound using a one tailed test at the 95% confidence level. Standard errors of each study were estimated as , except for studies from Camerer et al. (2018) that had more than a single replication attempt. In these cases we used the standard errors derived from the meta-analyses that produced the effect size estimate (see Supplementary Materials 1 for details).

In interpreting results based on this exclusion criterion, it is important to note that the minimum detectable effect was occasionally quite high because original sample sizes were often very small (mean equivalence bound in correlation coefficient terms = 0.18, SD = 0.11, 0th, 25th, 50th, 75th and 100th quintiles = [0, 0.1, 0.15, 0.23, 0.63]). This means that original studies were sometimes under-powered to detect even large effects using the current analysis, and as such this method may exclude studies that have replication effects that the original authors may have considered important (Thompson, 2002). See Supplementary Materials 2 for scatter plots of the dataset using each exclusion rule.

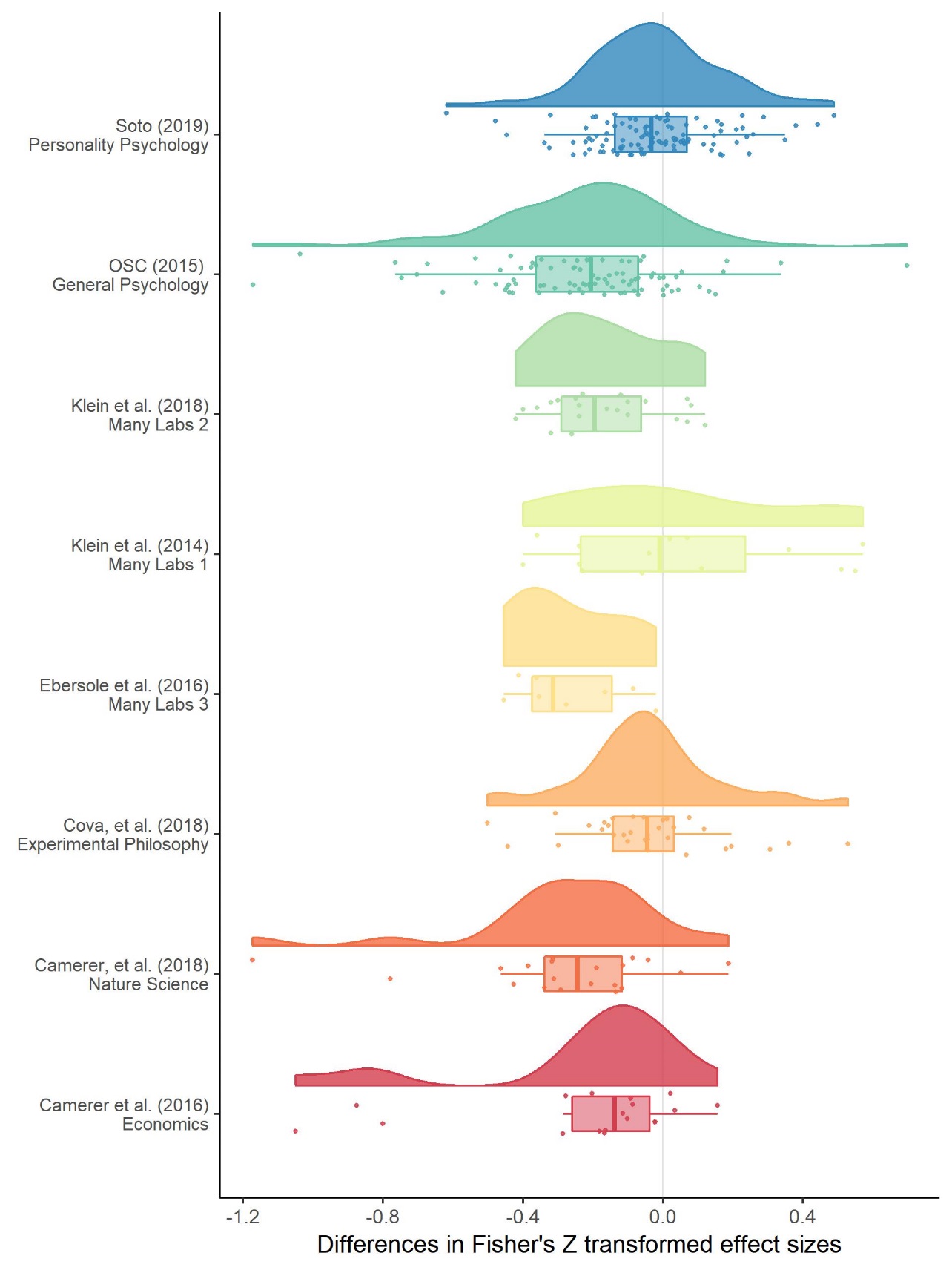
#### Analysis 4: Bayesian mixture model.

Analyses 2 and 3 both rely on excluding studies using exclusion rules that will, respectively, exclude or retain studies due to low statistical power in the replication study. In part in order to avoid this issue the final approach to estimating the amount of effect size attenuation conditional on the effect under study being non-zero was the Bayesian mixture model presented in Camerer et al. (2018). This model assumes that each replication effect size comes from one of two components, either from the null-hypothesis or from the alternative-hypothesis component. If the true replication effect size is drawn from the null-hypothesis component, it is assumed to be drawn from a normal distribution with a mean of 0. If the true replication effect size comes from the alternative-hypothesis component, then it is assumed to be drawn from a normal distribution with a mean equal to the original's estimated true effect size attenuated by an "attenuation factor". The attenuation factor is constrained to a value between zero and one and is assumed to be consistent across studies. The observed replication effect size is then assumed to be drawn from a normal distribution centered on the true replication effect size with a standard deviation equal to the standard error of the replication study (estimated here as with *N2* being the replication sample size).

There are two main parameters of interest in this model: (i) the “attenuation factor” (called a “deflation factor” in Camerer et al., 2018), which is the degree to which effect sizes are attenuated between original and replication studies, and (ii) the "assignment rate", which is the overall rate at which studies are assigned to the null hypothesis. This analysis was performed in JAGS (version 4.3.0; Depaoli, Clifton, & Cobb, 2016) using the rjags interface (version 4.8.0; Plummer, Stukalov, & Denwood, 2018). See Supplementary Materials 4 for model syntax and further analysis details.

## Results

#### Descriptives

Looking at the 306 included original-replication study pairs included in this analysis, the effect size in the replication study was lower than that in the original study for 219 articles, 72% of the included studies. An exact binomial test shows that this is extremely unlikely under the assumption that replication effect sizes are equally likely to be smaller or larger in the replication study, p < .001. The average effect size for original studies was a Fisher z score equivalent to r = 0.39, and the mean effect size for replication studies was r = 0.27, a mean decrease of r = 0.11. Notably, this represents an average decrease in effect sizes from the original to the replication study of 28%. See Table 2 for a comprehensive list of descriptive statistics on the effect size differences seen in this sample and Figure 1 for a raincloud plot of the Fisher z-score change in effect sizes by replication project.

*Figure 1.* A raincloud plot (density, box and scatter 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.

Table 2. *Differences between original and replication studies. All calculations were performed on Fisher z-transformed correlations and presented effect sizes are back-transformed into correlation coefficients for interpretability.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | All studies | statistically significant replications | Nonequivalent studies |
| n included | 306 | 198 | 237 |
| Mean original ES | 0.39 | 0.40 | 0.41 |
| Median original ES | 0.33 | 0.35 | 0.35 |
| Mean replication ES | 0.27 | 0.39 | 0.35 |
| Median replication ES | 0.22 | 0.33 | 0.30 |
| Mean ES difference | -0.13 | -0.02 | -0.07 |
| Median ES difference | -0.11 | -0.03 | -0.06 |
| SD difference | 0.25 | 0.20 | 0.24 |
| Mean proportion change | -0.28 | 0.04 | -0.07 |
| Median proportion change | -0.34 | -0.07 | -0.17 |

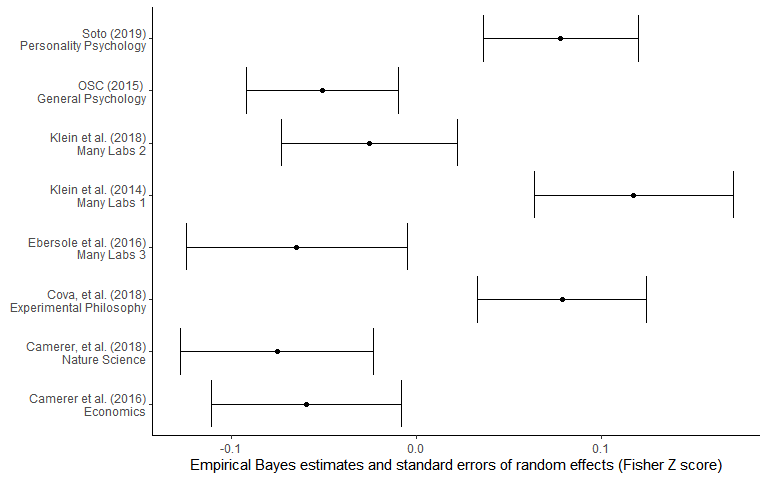
#### Analysis 1: Multilevel random effects meta-analysis results.

The random effects meta-analysis including all data estimated a *r* = -0.14 (95% CI [-0.2, -0.07]) decrease in effect sizes from the original to replication studies. This represents a decrease equivalent to 34% (95% CI [51%, 17%]) of the mean effect size in the original studies (a Fisher z-transformed correlation coefficient equivalent to a correlation of 0.37).

Greater variance was attributable to the article and effect level than to the project (with standard deviations for each random effect of = 0.128, = 0.110, and = 0.088), representing an intraclass correlation (ICC) for the project of 0.215. There was a large amount of unexplained heterogeneity, QE(305) = 3531.9, p < .001, = 92.585 (calculated following Nakagawa & Santos, 2012), suggesting that 93% of variance in effect sizes was due to heterogeneity (i.e., variance in the true effect size differences), while the remaining 7% was attributable to sampling variance. See figure 2 for Empirical Bayes estimates of the random effect for each of the replication projects (calculated following (calculated following Raudenbush & Bryk, 1985)).

Table 3. *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 |
| -0.137 | -0.206 | -0.068 | 0.035 | < .001 |  |
|  |  |  |  |  | Project variance = 0.008, n = 8 |
|  |  |  |  |  | Article variance = 0.016, n = 229 |
|  |  |  |  |  | Effect variance = 0.012, n = 306 |
|  |  |  |  |  | QE(305) = 3531.9, p < .001 |



*Figure 2.* Empirical Bayes estimates of the random effect of each replication project in Fisher Z scores (i.e., estimates of the difference between the replication project’s mean effect size difference and the overall estimated mean effect size difference). Error bars are 1 standard error.

#### Analyses 2 and 3: Results from multilevel random effects meta-analysis with exclusions.

Examining just the 198 cases in which the replication study was statistically significant (65% of all studies), the average effect for the original studies was 0.404, and the mean effect size for replication studies was 0.387. This represents a mean decrease of *r* = 0.017, a mean percentage increase in effect sizes of 4% and a median percentage decrease of 7%. Using equivalence testing 77% of replication studies were not statistically equivalent to the null (n= 237). The average effect size in the original non-equivalent studies was 0.406, compared to a mean effect size for replication studies of *r* = 0.348. This is a mean decrease of r = 0.058, a mean percentage decrease of 7%, and a median percentage decrease of 17%.

Reperforming the meta-analysis only including studies for which the replication was statistically significant and had an effect in the same direction as the original produced an estimated r = -0.051 (95% CI [-0.111, 0.010]) change in effect sizes from original to replication studies. Including only the studies which were not statistically equivalent leads to a predicted r = -0.082 (95% CI [-0.154, -0.010]) decrease in effect sizes. The estimates of the proportion of variance attributable to the article or replication project level did not change considerably in either of these subsets. See Table 4 for the model estimates from each model.

These values represent changes equivalent to a decrease of 12% to 20% of the average original effect size (a correlation coefficient of r = 0.387). However, there was considerable imprecision in these estimates, with 95% confidence intervals for both of these subsamples extending from a considerable decrease equivalent to 38% of the average original effect size, to a small increase equivalent to 2% of the average original effect size.

##### Table 4.

*The number of studies included in each model, and the estimated correlation coefficient decrease from each model.*

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | All studies | Statistically significant replications | Nonequivalent studies |
| Model N | 306 | 198 | 237 |
| Estimated decrease | -0.14 | -0.05 | -0.08 |
| 95% CI LB | -0.20 | -0.11 | -0.15 |
| 95% CI UB | -0.07 | 0.01 | -0.01 |
| Estimated % attenuation | -33.67 | -12.41 | -20.20 |
| LB % attenuation | -50.59 | -27.25 | -38.01 |
| UB % attenuation | -16.74 | 2.43 | -2.40 |

*Note:* 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 of the mean original effect size (r = 0.366).

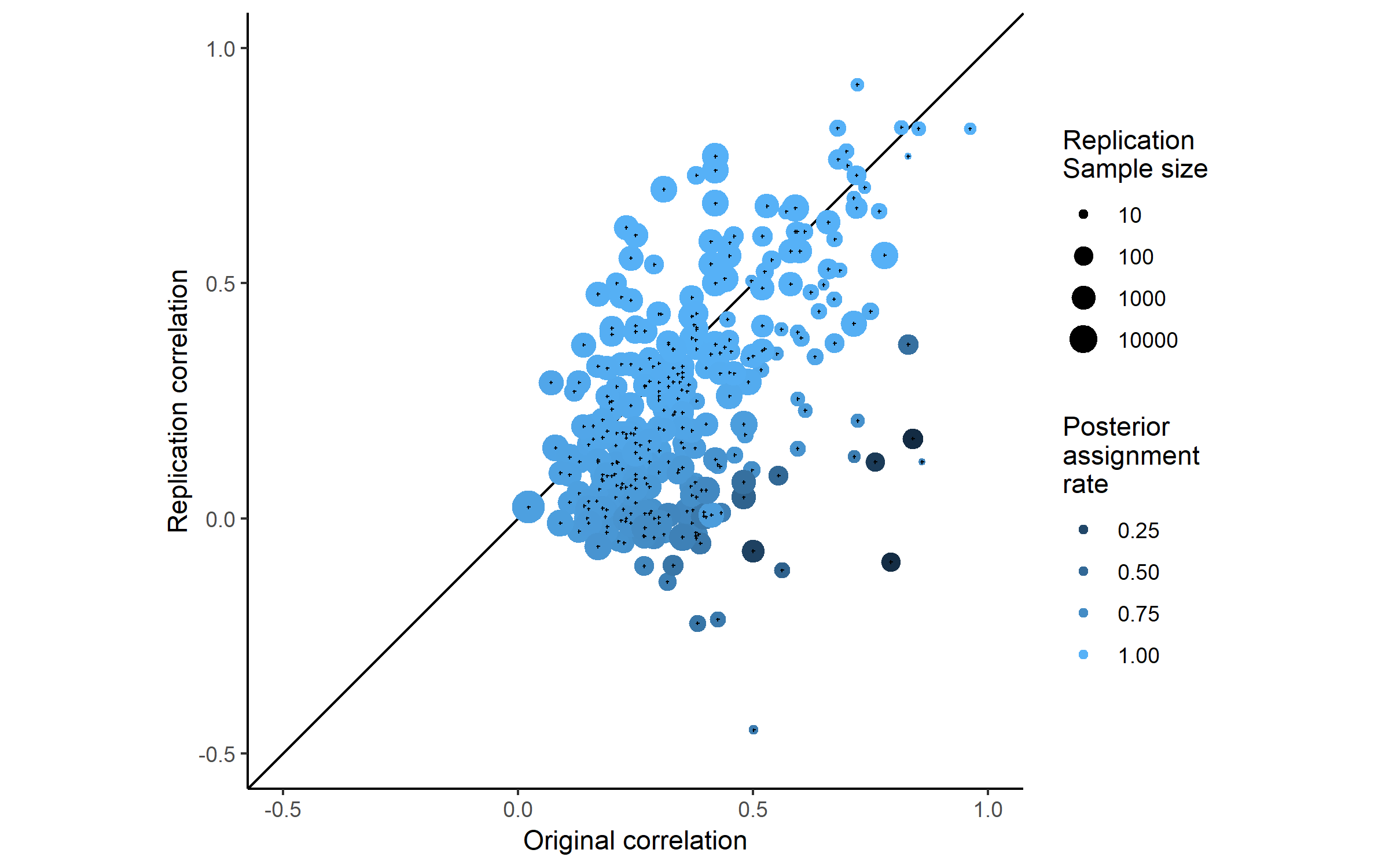
#### Leave one out cross validation of meta-analyses

To assess how sensitive the results of the multilevel models were to the inclusion of each of the replication projects, the included studies, and the individual replicated effects, all of the above multilevel models were rerun using leave one out cross validation, excluding both each effect, effects from each original study (i.e., in cases where multiple effects were tested from the same original source), and each replication project one at a time. None of these analyses led to model estimates (i.e., the expected decrease in effect size between original and replication study or equivalently the intercept estimate) that were further than 0.02 from those given above, suggesting that none of the individual projects, effects or studies included were overly influential. See Supplementary Material 3 for tables summarising the leave-one-out model output.

#### Analysis 4: Bayesian mixture model results.

The Bayesian mixture model was estimated using four Markov chains from each of which 100,000 draws were taken (excluding an 11,000 draw burn-in period). Trace and density plots for the discussed parameters were examined and, along with and values within .001 of 1, appeared to suggest that the model successfully converged (Gelman & Shirley, 2011). The overall posterior assignment rate (i.e., the proportion of studies that were estimated to be from the non-null alternative hypothesis) was 89%, with a 95% highest probability density interval of [79%, 98%]. The overall attenuation factor (i.e., the estimated amount that effect sizes decreases between the original and replication studies) was 19% with a 95% highest probability density interval of [11%, 28%]. Figure 3 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 how often each effect was assigned to the alternative hypothesis.

As was pointed out in Camerer et al. (2018), values close to the diagonal (i.e., cases in which the original and replication effect sizes are similar) were reliably assigned to the alternative hypothesis component whereas effects far below the diagonal were more often assigned to the null hypothesis component. The overall posterior assignment rate might be overly optimistic (i.e., assign studies to the non-null hypothesis at a high rate), likely in part due to the fact that this model allows for “true” effect sizes to be estimated as being extremely low or near zero and still assigned to the alternative hypothesis, with 29% of the estimated “true” replication effect sizes being smaller than a correlation coefficient of .10.



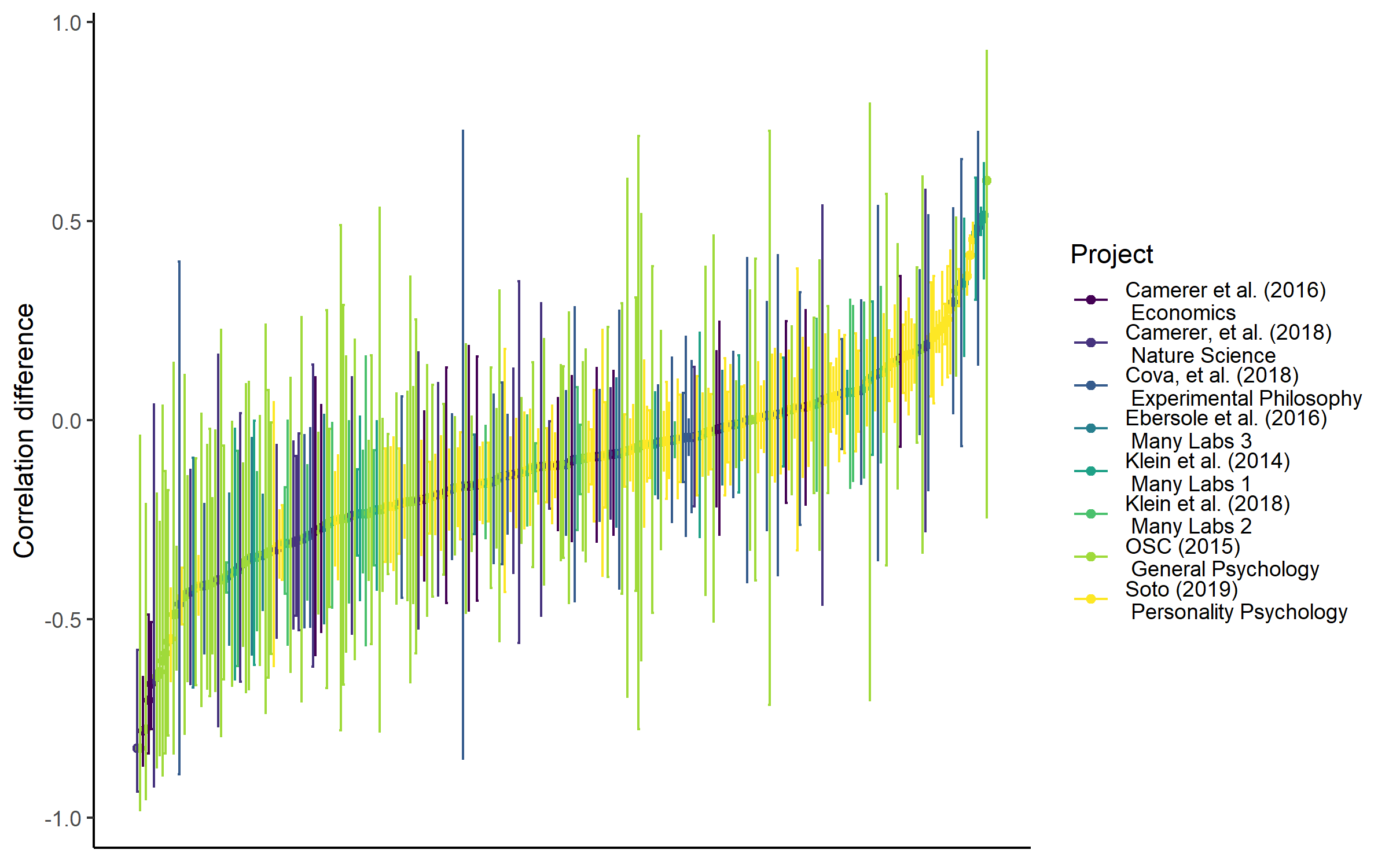
*Figure 3.* A scatterplot of replication study effect sizes (in correlation coefficients) plotted against original study effect sizes, coloured by the posterior assignment rate, the proportion of times each study was assigned to the alternative hypothesis. Points that fall on the solid, diagonal line 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

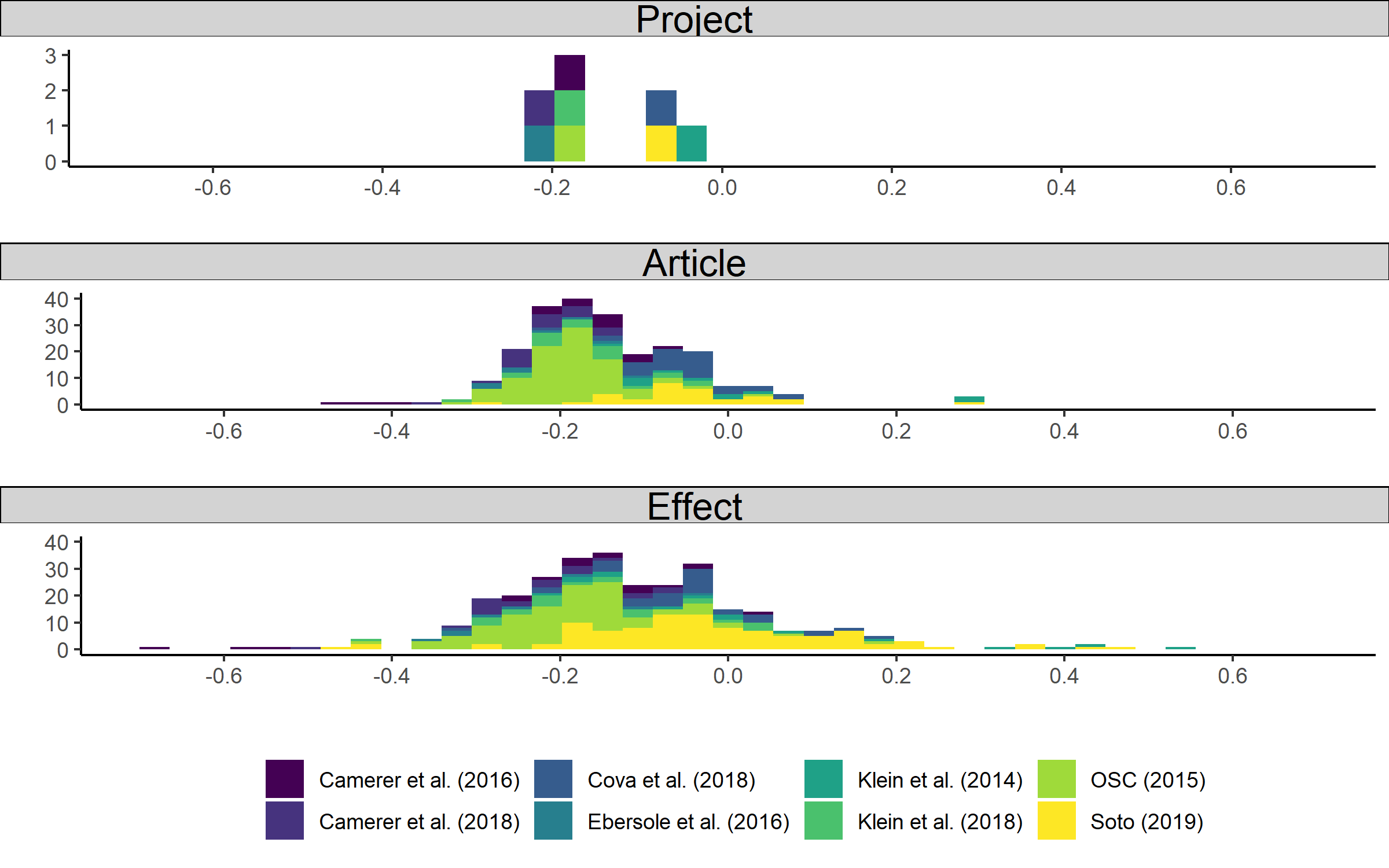
The results show that there was a substantial average decrease in effects sizes between original and replication studies and suggest that this is still the case even after accounting for the presence of null effects. The results of the multilevel meta-analysis show an estimated mean decrease of *r* = -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% (95% CI [51%, 17%]) of the mean effect size in the original studies (a Fisher z equivalent to *r* = 0.39).

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 agreed, although the degree of precision in their estimates differs. The Bayesian mixture model suggests that there is an average decrease of 19%, with a 95% highest probability density interval of [11%, 28%]. The multilevel meta-analysis models that excluded non-significant results and studies in which the replications studies were statistically equivalent to the null both led to similar conclusions, although they give less precise estimates, highlighting the uncertainty in these estimates. For example, the confidence intervals over both of the models in Analyses 2 and 3 extend from a meaningful decrease of *r* = -0.15 to a trivial increase of *r* = 0.01.

In using these results to inform future research (e.g., in sample size planning) and to interpret the published literature, it is essential to take note of the level of heterogeneity in the effect size attenuation across not just replicated effects but also across replication projects. The sampling decisions and replication methods used by each of the included projects appears to have had a considerable effect on the amount of effect size attenuation seen (e.g., looking at the results of Analysis 1, the estimated standard deviation of the mean level of effect size attenuation across projects is 0.13, 95% CI [0.04, 0.18]). The degree of heterogeneity in the amount of effect size attenuation across studies and projects means that using any single estimate of the amount of effect size decrease is likely to be misleading in the case of any individual replication study, see Figure 5 for an illustration showing the degree of heterogeneity expected at each level of the multilevel model.



*Figure 4.* A caterpillar plot of the effect size differences between original and replication study effect sizes ordered by magnitude in correlation coefficients, error bars are 95% confidence intervals around effect size differences.

 *Figure 5.* Histograms of the expected differences (i.e., the sum of the Empirical Bayes estimates for random effects and the model intercept from Analysis 1) between original and replication studies for each replication project (top panel), study (middle panel), and effect (bottom panel) in Fisher Z scores, illustrating the degree of heterogeneity expected at each level of the multilevel model.

### Limitations and future directions

### In interpreting these results it is important to note several limitations. Firstly, the current study cannot distinguish between effect size heterogeneity (i.e., effect sizes that are different due to subtle unobserved moderators (Kenny & Judd, 2019) and effect size exaggeration. However, in so far as effect size heterogeneity tends to lead to smaller effects in replication studies, the distinction may not matter for the purposes of researchers hoping to replicate or plan future similar studies of the same type of effects. It also cannot be ruled out that the effect size differences seen in these large-scale replication projects are larger than would be seen by individuals attempting to replicate particular effects (e.g., if researchers in these large-scale replications have less access to the tacit knowledge that would normally facilitate replicators’ efforts).

The Bayesian mixture model assumes independence between effects, a uniform attenuation factor across all areas of psychological research, and allows for effects sampled from the alternative distribution to be negligibly small or even negative. Future research could help develop a more nuanced account of the data-generation process underlying this dataset by, for example, building a model that allows for the attenuation rate to change across replication studies, or by including more components in order to allowing for studies with negligible or negative but non-null effects in addition to the alternative and null components.

The large amount of heterogeneity across replication projects limits the utility of the fixed effects parameter estimates (i.e., the estimated mean level of attenuation seen across all projects). While the current dataset allows us to make reasonable inferences about future replication projects, the varied sampling strategies employed by the previous projects necessarily limit the inferences we can draw. At the moment, it is difficult to say whether the heterogeneity in effect size attenuation are due to intra-field differences in replicability and effect size attenuation, differences caused by the sampling strategies, or other issues such as differences in the quality of the replication studies.

In order to be able to start to disentangle these different possible causes and develop an account of the predictors and moderators of replication success and effect size attenuation, we need to begin to develop a large representative database of replication studies. Especially if such a database could be augmented with meta-data regarding the type of analysis, design and effects under study, this resource could allow us to make meaningful predictions about individual future replications. However, until a large database of such studies becomes available, analyses like the current one provide the best estimates possible, albeit estimates that should be read and understood with these caveats in mind.

### Conclusion

The findings of this study reinforce the importance of recent efforts to reduce the reliance on underpowered original research designs in psychological research, to circumvent publication bias, and to avoid QRPs like p-hacking and HARKing (Bakker et al., 2012). Efforts to avoid the impact of any of these issues would likely reduce the degree to which effect sizes are attenuated in replications of the primary research literature.

In order to avoid performing future underpowered research, researchers should be aware that their experiments are likely to be underpowered if they plan their sample sizes using the effect size reported in a previous experiment. As a conservative heuristic for researchers performing formal sample size planning on the basis of previous research, researchers could follow the advice given in Camerer et al. (2018) and plan their experiments assuming that the true effect size is 50% of the reported effect size, a value matched by the more extreme 95% confidence interval of the estimated amount of effect size exaggeration across studies in this sample. Alternatively, it may be preferable to use methods of sample size planning that do not rely on precise a priori estimation of the effect size under study, such as planning studies to reliably detect the smallest effect size of interest (Lakens et al., 2018), using sequential analysis strategies (Pocock, 1977), or planning for adequate precision in parameter estimates across a range of possible effect sizes (Kelley, Darku, & Chattopadhyay, 2017; Maxwell, Kelley, & Rausch, 2008). In addition, recent large-scale multinational data collection efforts like the Many labs Projects or the Psychological Science Accelerator (Moshontz et al., 2018) also help to avoid the negative impacts of low statistical power by allowing for extremely high powered studies of even very small effects.

There are several recent efforts to reduce the impact of publication and reporting biases that readers should be aware of, many of which individual researchers can voluntarily and easily take part in. Careful preregistration of analysis plans allows researchers to avoid biases in the analysis of their data that may otherwise lead to inflated effect sizes (Wicherts et al., 2016). Data-sharing platforms such as figshare (figshare.com) and the Open Science Framework (osf.io) make it possible for researchers to easily share the results of research whether or not a study is published in a traditional journal. Similarly, pre-prints (e.g., <https://psyarxiv.com>) allow researchers to report and publicize reports and data that may otherwise remain in the file draw. Both preprints and data repositories make it easier to ensure that non-significant results are accessible to other researchers and meta-analysts. Finally, registered reports, in which papers are reviewed before data-collection on the basis of the research design and analysis strategy as opposed to the results, also show promise in helping to develop a body of literature that is not affected by reporting and publication bias (Nosek & Lakens, 2014). However, until large bodies of research free of publication bias become available, researchers should be aware that effect sizes in published studies may be considerably overstated.

Author contributions

F. Singleton Thorn conceptualized the research question, designed and performed the statistical analysis, collected the data and drafted this manuscript. P. Dudgeon provided critical revisions and advice on the statistical analyses reported in this manuscript.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

### References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE, 12*(3), e0172792. doi:10.1371/journal.pone.0172792

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi:10.1038/nrn3475

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153. doi:10.1037/h0045186

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just Another Gibbs Sampler (JAGS): Flexible Software for MCMC Implementation. *Journal of Educational and Behavioral Statistics, 41*(6), 628-649. doi:10.3102/1076998616664876

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J. (2013). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review. *PLOS ONE, 8*(7), e66844. doi:10.1371/journal.pone.0066844

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:<https://doi.org/10.1016/j.jesp.2015.10.012>

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629. Retrieved from <http://www.bmj.com/content/315/7109/629.abstract>

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. doi:10.1007/s11192-011-0494-7

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. *Social Psychological and Personality Science, 7*(1), 45-52. doi:10.1177/1948550615612150

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502. Retrieved from <http://science.sciencemag.org/content/345/6203/1502.abstract>

Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of markov chain monte carlo* (Vol. 6, pp. 163-174). Boca Raton, Florida: CRC press.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: what is going on? *PeerJ, 4*, e1935. doi:10.7717/peerj.1935

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science, 7*(2), 246-255.

The importance of no evidence. (2019). *Nature Human Behaviour, 3*(3), 197-197. doi:10.1038/s41562-019-0569-7

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2017). Accuracy in Parameter Estimation for a General Class of Effect Sizes: A Sequential Approach. *Psychological Methods*. doi:10.1037/met0000127

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol Methods*. doi:10.1037/met0000209

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217. doi:10.1207/s15327957pspr0203\_4

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., . . . Brumbaugh, C. (2014). Data from investigating variation in replicability: A “many labs” replication project. *Journal of Open Psychology Data, 2*(1).

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances In Methods and Practices in Psychological Science, 1*(2), 259-269. doi:10.1177/2515245918770963

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology, 31*(2), 107-112. doi:10.1111/j.2044-8317.1978.tb00578.x

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175. doi:10.1007/BF01173636

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537-563. doi:doi:10.1146/annurev.psych.59.103006.093735

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances In Methods and Practices in Psychological Science, 1*(4), 501-515. doi:10.1177/2515245918797607

Murphy, K. R., & Aguinis, H. (2017). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*. doi:10.1007/s10869-017-9524-7

Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253-1274. doi:10.1007/s10682-012-9555-5

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology, 45*(3), 137-141. doi:10.1027/1864-9335/a000192

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York, NY: Wiley.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Retrieved from <http://science.sciencemag.org/content/349/6251/aac4716.abstract>

Plummer, M., Stukalov, A., & Denwood, M. (2018). rjags: Bayesian Graphical Models using MCMC. R package version 4.8.0.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191-199. doi:10.1093/biomet/64.2.191

R Development Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes Meta-Analysis. *Journal of Educational Statistics, 10*(2), 75-98. doi:10.3102/10769986010002075

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology, 10*, 813. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2019.00813>

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science, 22*(11), 1359-1366. doi:10.1177/0956797611417632

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin*.

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. *Journal of the American Statistical Association, 54*(285), 30-34. doi:10.2307/2282137

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*(3), e2000797. doi:10.1371/journal.pbio.2000797

Thompson, B. (2002). What Future Quantitative Social Science Research Could Look Like: Confidence Intervals for Effect Sizes. *Educational Researcher, 31*(3), 25-32. doi:10.3102/0013189X031003025

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal Of Statistical Software, 36*(3), 48. doi:10.18637/jss.v036.i03

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832