## Estimating the effect of publication bias in Behavioural research

## Introduction

An important aspect in developing a coherent and reliable scientific literature is the discovery and precise estimation of associations and effects. Although the presence or absence of effects may be an interesting question in of itself, an understanding of the magnitude and importance of effects is a central aspect in many research contexts. A major effort among psychological researchers and methodologists in recent decades in behavioural research has been the movement away from focusing entirely on binary outcome statistical significance testing. These figures have argued against focusing on the presence or absence of effects, but rather on the size and direction of effects and precision in estimation (e.g., Cohen, 1990; Cohen, 1994; Meehl, 1967, 1978).

Under conditions where results are selectively reported based on characteristics related to the size of the effect (e.g., statistical significance), the literature no longer provides an unbiased estimate of the true outcome effect (Hedges, 1992). However, in current practice there is good reason to think that reporting and publication biases lead to exaggeration of effect sizes in the behavioural sciences literature results (Lane & Dunlap, 1978; Mahoney, 1977; Murphy & Aguinis, 2017; Simmons, Nelson, & Simonsohn, 2011). An essential question for the purposes of understanding our scientific literature, as well as for planning future studies, is assessing the degree to which effect sizes are inflated in the published literature. The current paper examines a newly available resource, large scale replication studies which have systematically replicated bodies of research, in order to estimate the degree to which effects reported in the psychological literature are inflated.

The current paper examines the results of over 300 replication attempts conducted as a part of seven large scale replication projects (henceforth just ‘replication projects’). All of these projects were primarily conducted in order to assess the degree to which their particular area of research contains results which are irreproducible, or to estimate variability in effects among subpopulations. All used non-random samples of the literature, and all show that the reproducibility of results is below what would be expected given that all experiments were being analysed and published without regard to the statistical significance of results. See Table 1 for a list of the included replication projects, along with their target populations, and the percentages of replication attempts with a statistically significant results in the same direction as the replicated result.

This new body of literature makes it possible to assess the effect of publication bias on the size of the reported effects, beyond the top line result of showing fewer statistically significant results in the same direction as the original study. In order to do so, the current study presents an exploratory analysis of this large set of data using a multilevel random effects meta-analytic framework to estimate the effect of publication bias on effect sizes seen in the literature. As the replication studies inevitably include a large number of effects which are likely to be true null effects (or effects which are so close to true null effects to be practically dismissible), I also estimate the degree to which true effect sizes in the literature are decreased after excluding those effects which are likely to be true nulls using frequentist and Bayesian exclusion rules. This is especially important as several of the studies reported in the current study report having chosen articles explicitly to represent a range of likelihoods of replicating, as well as based on the study criteria themselves (e.g., the included Many Labs studies).

Table 1.  
*Included large scale replication projects, along with the number of articles replicated, the number included in this analyses of each type, and how sample sizes were determined.*

### Publication and reporting bias’s effect on reported effect sizes

Publication bias describes the observation that studies are more likely published if they find results which support their hypotheses, usually by showing statistically significant results (Lane & Dunlap, 1978; Mahoney, 1977). This is the traditional “file draw problem” (Rosenthal, 1979), the idea that non-significant results get placed in the file draw as opposed to being reported. If studies are more likely to be published given that 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). This occurs because the smaller the sample size included in research, all else being equal, the larger the observed effect has to be to reach statistical significance. When an effect understudy is truly null, or practically indistinguishable from the null, and null effects are rarely published, this can create the appearance of a non-null effect in the literature, and the fallacious appearance of support for particular theories, based on false positive results alone (Oakes, 1986). Selective reporting and Questionable Research Practices (QRPs) can also lead to the same outcomes, when particular outcome measures are reported, emphasised, or not reported because of the results of statistical analyses. QRPs like p-hacking and Hypothesising After the Results are Known (HARKing) on the basis of the some outcome measure such as statistical significance or effect size magnitudes (Kerr, 1998) can also lead to effect sizes being exaggerated and increased proportions of false positives in the scientific literature (Murphy & Aguinis, 2017; Simmons et al., 2011). The degree of effect size exaggeration depends, primarily, on the true statistical power of studies (i.e., the effect size and sample size included in studies given the experimental design and analysis strategy) and the proportion of true nulls being investigated (Oakes, 1986).

If anything, publication bias towards statistically significant results appears to be particularly acute in behavioural research. It is difficult to explain the high proportion of studies in psychology reporting statistically significant results (estimates range from 75% to over 90%; Fanelli, 2010; Fanelli, 2012; Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2016) without suggesting that publication bias, or use of QRPs, are inflating the proportion of studies which report statistically significant findings. This is particularly the case when looking at estimates of the average power of psychological research. Taking a recent estimate of the average power of psychology to detect reasonable estimates of the average effect sizes seen in psychology (44% to detect a cohen's d of .5; Szucs & Ioannidis, 2017), would mean that over `r studiesPerPublishedPaper` would have to be conducted per published paper to account for the proportion of studies that report significant findings[[1]](#footnote-1). Recent surveys in the behavioural research literature also suggest that that questionable research practices activities like HARKing and p-hacking appear are common across fields of psychological research (Fiedler & Schwarz, 2015; John, Loewenstein, & Prelec, 2012). All of these activities lead to increased false positive errors, and equivalently exaggerated effect sizes as represented in the scientific literature. In order to be able to read behavioural research, it is essential to have a good estimate of the degree to which reported effect are exaggerated in the literature. The current paper provides an estimate of the cumulative effect of these behaviours on published effect sizes.

Previous efforts to estimate publication bias

Previous efforts to assess this question have shown using simulation studies that under reasonable assumptions type one error rates could be as high as 40% and effect size exaggeration as high as d = .33 when questionable research practices are in place even, and even without QRPs as high as d = .16, in typical experiments in psychology (Bakker, van Dijk, & Wicherts, 2012). Stanley, Carter, and Doucouliagos used WAAP-WLS, ELS and PET-PEESE estimators in 200 meta-analyses published in Psychological Bulletin suggest that differences in effect sizes between replication and original studies can largely be explained by heterogeneity not selective reporting in either direction. They found just 8 to 15% residual effect size bias depending on the meta-analytic bias reduction method used. However, two of three of these estimation methods are known to be downwardly biased (leading to underestimates of the amount effect size inflation (Stanley & Doucouliagos, 2015; Stanley, Doucouliagos, & Ioannidis, 2017),), and this literature could reasonably be expected to be less biased than others in that they only necessarily sampled studies for which enough papers had been published to perform a meta-analysis.

## Methods

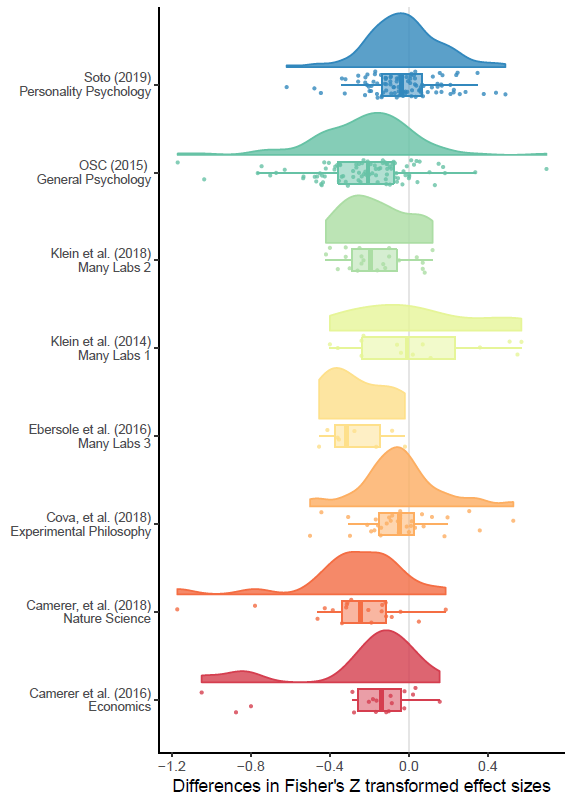
SEE MARKDOWN DOC

(Wagenmakers, Verhagen, & Ly, 2016)

## Results

Markdown

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



### Data extraction

## Conclusion

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312. doi:10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round (p < .05). *American psychologist, The, 49*(12), 997-1003. doi:<http://dx.doi.org/10.1037/0003-066X.49.12.997>

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE, 5*(4), e10068. doi:10.1371/journal.pone.0010068

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

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. Retrieved from <http://www.jstor.org.ezp.lib.unimelb.edu.au/stable/2246311>

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

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

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

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), 103-115. doi:10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834. doi:10.1037/0022-006X.46.4.806

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

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. Retrieved from <http://sfx.unimelb.hosted.exlibrisgroup.com/sfxlcl41?rft.jtitle=Psychological%20Bulletin&rft.volume=86&rft.spage=638&rft_id=info%3Adoi%2F10.1037%2F%2F0033-2909.86.3.638&rft.genre=article&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Ajournal&ctx_ver=Z39.88-2004&url_ver=Z39.88-2004&url_ctx_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Actx&rfr_id=info%3Asid%2Fpps.sagepub.com&ctx_tim=2016-03-08T02%3A06%3A12.902-08%3A00>

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. What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin*. Retrieved from <https://ezp.lib.unimelb.edu.au/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=ovftt&AN=00006823-900000000-99701>

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: weighted least squares meta-analysis. *Statistics in Medicine, 34*(13), 2116-2127. doi:10.1002/sim.6481

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. A. (2017). Finding the power to reduce publication bias. *Statistics in Medicine, 36*(10), 1580-1598. doi:10.1002/sim.7228

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

Wagenmakers, E. J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behav Res Methods, 48*(2), 413-426. doi:10.3758/s13428-015-0593-0

1. to . These estimates assume that all studies performed examine non-null findings and ‘medium’ effects of .5 Cohen’s *d*. More realistic estimates of the proportion of null-experiments would lead to greater numbers of experiments having to be performed. [↑](#footnote-ref-1)