## 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. 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). 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 studies. See Table 1 for a list of the included studies, along with their target populations, and the percentages of replication attempts with a statistically significant results in the same direction as the replicated result. All of these studies 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 of these studies used non-random samples of the literature. 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.

This new body of literature makes it possible to assess the effect of publication bias on the size of the reported effects. 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 so close to true null effects to be practically demisable), I also attempt to estimate the degree to which true effect sizes in the literature are decreased, by excluding those effects which are likely to be true nulls using various methods.

Table 1.  
*Included large scale replication projects, along with the number of articles replicated, the number included in this analysis and …*

### Publication and reporting bias

The most common definition of publication bias is the idea that studies are more likely to be accepted for publication 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). Selective reporting can also lead to the same outcomes, where particular outcome measures are reported, emphasised, or not reported because of the results of statistical analyses. Questionable Research Practices, like p-hacking and Hypothesising After the Results are Known (HARKing) (Kerr, 1998) on the basis of the some outcome measure such as statistical significance or effect size magnitudes, can also lead to effect sizes being exaggerated and increased proportions of false positives in the scientific literature (Murphy & Aguinis, 2017; Simmons, Nelson, & Simonsohn, 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 also appears to be particularly acute in psychology and 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 the use of QRPs, are inflating the proportion of studies which report statistically significant findings. In fact, it seems unlikely that publication bias alone can account for the proportion of studies reporting statistically significant results. 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. The current paper provides an estimate of the cumulative effect of these behaviours on published effect sizes.

### Defining ‘successful’ replication

In order to provide a useful estimate of the effect of studies

Methods of judging replication success used in the current study

**Statistically significant results** in the replication study. The p values used in this instance were taken from the replication studies’ data.

**95% CIs containing the replication effect size**

If there was no publication bias in the literature, under pure sampling variability we would expect to see 83.4% of studies falling in the 95% CI of the original study (Cumming, Williams, & Fidler, 2004). This analysis was performed for all tests were valid SEs could be developed for both the original and replication study (i.e., for t, and correlations, and F tests with a df1 of 1). N = `nSEsvalid`.

**Bayes Factors** - BFs 01 were calculated following (Wagenmakers, Verhagen, & Ly, 2016). The Jeffreys–Zellner–Siow (JZS) default prior was used following (Wagenmakers et al., 2016), although these values should be considered an approximation as they do not take into account the fact that correlations were converted from other test statistics or effects sizes.

## Methods

SEE MARKDOWN DOC

## Results

A paper hoping to estimate the average effect of publication bias in psychology using what effect sizes

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>

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and Researchers' Understanding of Confidence Intervals and Standard Error Bars. *Understanding Statistics, 3*(4), 299-311. doi:10.1207/s15328031us0304\_5

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

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, lowering the number of studies which have to be performed to account for the proportion of significant main findings in psychology. More realistic estimates would lead to greater numbers of non-null studies having to be performed. [↑](#footnote-ref-1)