OPEN ACCESS Freely available online

PLOS ONE

Why Selective Publication of Statistically Significant Results Can Be Effective

Joost de Winter*, Riender Happee

Department of BioMechanical Engineering, Delft University of Technology, Delft, The Netherlands

Abstract

Concerns exist within the medical and psychological sciences that many published research findings are not replicable. Guidelines accordingly recommend that the file drawer effect should be eliminated and that statistical significance should not be a criterion in the decision to submit and publish scientific results. By means of a simulation study, we show that selectively publishing effects that differ significantly from the cumulative meta-analytic effect evokes the Proteus phenomenon of poorly replicable and alternating findings. However, the simulation also shows that the selective publication approach yields a scientific record that is content rich as compared to publishing everything, in the sense that fewer publications are needed for obtaining an accurate meta-analytic estimation of the true effect. We conclude that, under the assumption of self-correcting science, the file drawer effect can be beneficial for the scientific collective.

Citation: de Winter J, Happee R (2013) Why Selective Publication of Statistically Significant Results Can Be Effective. PLoS ONE 8(6): e66463. doi:10.1371/journal.pone.0066463

Editor: K. Brad Wray, State University of New York, Oswego, United States of America

Received March 20, 2013; **Accepted** May 4, 2013; **Published** June 20, 2013

Funding: No current external funding sources for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: j.c.f.dewinter@tudelft.nl

Replicability Crisis and the File Drawer Effect

It is widely held that "replicability of findings is at the heart of any empirical science" [1]. Replication is obtained if applying the same research design in an independent sample of participants yields the same result, meaning that any difference between the observed effect and the true (population) effect is insubstantial [1].

Concerns exist within the medical and psychological communities that many published findings are poorly replicable. Published research findings are often false positives [2] or gross exaggerations of the true effect [3,4], especially in domains where effect sizes and sample sizes are small and the prior probability of a hypothesis being true is low [2,5,6]. According to Pashler and Harris [7], one can legitimately speak of a "replicability crisis".

Poor replicability is, in part, caused by the file drawer effect, meaning that findings that are statistically significant are more likely to be submitted and accepted for publication than null results [8–12]. Selective reporting is typically regarded as a questionable research practice [13] and has been associated with researchers' pressure on productivity and novelty [6], flexibility in data analysis [14], desire for media attention [15], aversion to null results [16], and with the fact that many journals have a low acceptance rate. As pointed out by Giner-Sorolla [17], a publication bottleneck exists because researchers carry out many studies while there are relatively few publication outlets. Young, Ioannidis, and Al-Ubaydli [18] similarly argued that journals create an artificial scarcity of publication opportunity and an illusion of exclusivity.

Many authors recommend that the file drawer effect should be eliminated and that p values and effect sizes should not be a criterion in the decision to submit and publish scientific work [1,11,16,19–22]. Davison and Nevin [23], for example, recommend that editors and reviewers should not be biased towards publishing novel or different results, but should publish also null results. Ioannidis [24] envisions a future ideal in which we publish everything to make "the scientific record complete rather than fragmented and opportunistic". This publication philosophy is also adopted by the journal PLoS One, which states it will publish all papers that are judged to be technically sound [25].

The recommendation to publish both statistically significant and nonsignificant results is valid if the aim is to maximize the replicability of individual research studies. After all, according to the regression-toward-the-mean phenomenon, extreme variables tend to be closer to the true effect on a repeating measurement. However, we argue that this recommendation is less defensible from the perspective of the scientific collective. With a simulation, we show that selective publication eventually yields a more accurate estimate of the true effect than publishing everything.

Assumptions of the Simulation Study

Our simulation study acts on the premise that science is self-correcting, and that what has previously been the alternative hypothesis becomes the null hypothesis which researchers try to refute. This premise is in line with Bronowski [26] who explained that "science is essentially a self-correcting activity … scientists are people who correct the picture of the moment with another one, as a natural evolution towards a 'true' picture of the world". Specifically, we assume that researchers test their hypothesis with respect to the prevailing consensus as assessed by a cumulative meta-analysis of studies published on the same research question.

Ioannidis [2] argued that "negative" results become attractive for publication only if another researcher has published a "positive" result on the same research question. Elsewhere, Ioannidis and Trikalinos [27] coined the term "Proteus phenomenon" to describe their observation of "rapidly alternating extreme

research claims and extremely opposite refutations" [2], particularly during the early accumulation of data. Figure 1 illustrates the Proteus phenomenon as observed in a genetic association study. It can be seen that the first publication substantially overestimates, and that the second publication underestimates the eventual summary effect. The Proteus phenomenon has previously been interpreted as an intricate form of publication bias [27–30]. We suggest that selective publication results in the Proteus phenomenon and contributes to an effective convergence towards the true effect in a cumulative meta-analysis.

Along with the self-correction assumption, our simulation assumes constant study quality and single hypothesis tests, each generating one p value. Of course, in reality, studies can be more complex and multiple hypotheses may be tested in a single assay. We do not contend reporting standards for complex research, such as making research data, protocols, and analytical codes publicly available (cf. [31]). Furthermore, the factor time is not included in our simulation models. That is, results are assessed per publication without taking into account study completion time and the time between study completion and publication. In reality, publication of research findings is not a sequential process as multiple researchers could be working on a topic in parallel.

Simulation of the Publish Everything Approach versus the Selective Publication Approach

Computer simulations can be used to study dynamic processes of complex systems for which analytical solutions are not readily available. Herein, we use simulation to explore researchers' publishing behavior as a function of previously published effects on the same research question. We compare two publication approaches: a Publish Everything Approach and a Selective Publication Approach. The prevailing opinion is that publishing everything is the preferred method and that selective reporting is a questionable research practice [13].

Suppose that researchers worldwide are investigating the strength of an effect by means of identical experiments and that the observed effects appear in published articles. The observed effects (E_{obs}) are generated by independent random sampling of n subjects from a normal distribution with standard deviation of 1 and a mean E_{true}.

In the Publish Everything Approach, observed effects are always published, irrespective of their magnitude or direction. In the Selective Publication Approach, statistically significant findings ($p \leq \alpha$) are published and nonsignificant findings ($p > \alpha$) are not published (i.e., placed in the file drawer). The p value is determined using a two-tailed z test on E_{obs} with respect to the null hypothesis E_{meta}, which is the cumulative meta-analytic effect aggregating studies published on the same hypothesis so far, as in Eq. 1. In other words, a publication occurs only if the observed result (E_{obs}) differs statistically significantly from the prevailing consensus (E_{meta}).

$$E_{meta} = \frac{1}{N}\sum_{i=1}^{i=N} E_{obs,pub,i} \quad (1)$$

$E_{obs,pub,i}$ is the observed effect as published in the i-th publication and N is the number of publications so far. E_{meta} is assumed to be 0 if no studies have been published yet.

We used the following input to the simulation: $\alpha = 0.05$ (the false positive rate or significance level), $E_{true} = 0.3$ (a relatively small true effect), and $n = 50$ (the sample size for each study). The simulation stopped when 40 studies were published in the Selective Publication Approach. The simulation was repeated 5,000 times, to be able to calculate the expected values of $E_{obs,pub}$ and E_{meta}.

The mean observed effect of the published studies as a function of the publication number in Figure 2 shows an oscillating pattern for the Selective Publication Approach, akin to the Proteus phenomenon in Figure 1. The standard deviations around E_{true} illustrate that published effects in the Selective Publication Approach differ more from the true effect than in the Publish Everything Approach. The high standard deviations are caused by the fact that, in the Selective Publication Approach, observed

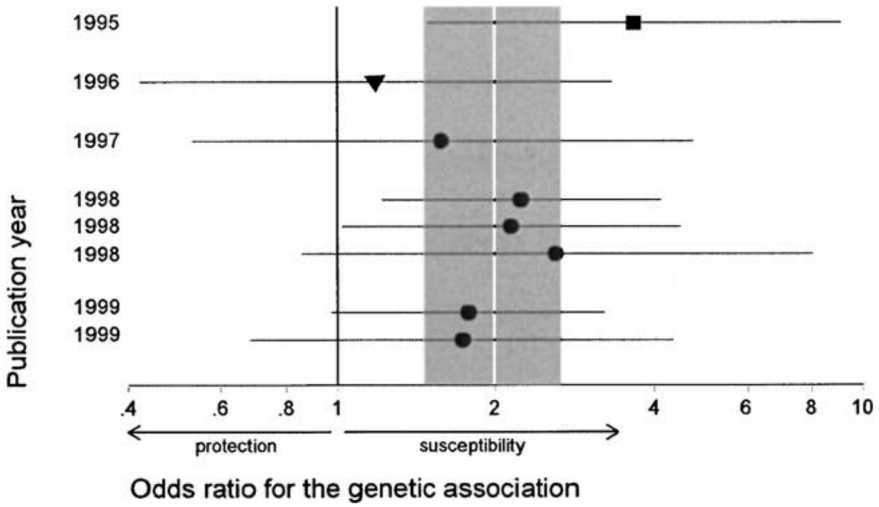

Figure 1. Illustration of Proteus phenemonen from Ioannidis and Trikalinos [27]. (Reprinted from Journal of Clinical Epidemiology, Vol. 58, J. P. Ioannidis and T. A. Trikalinos, Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials, pp. 543–549, 2005, with permission from Elsevier.) The figure shows odds ratios and 95% confidence intervals of "the relationship between the methylenetetrahydrofolate reductase (MTHFR) TT genotype in the mother and the risk of neural tube defects in the child". The study with the strongest effect is shown by a square symbol and the study with the smallest effect is shown by a triangular symbol. The white line represents the summary odds ratio. The shaded area represents the 95% confidence interval of the summary odds ratio.
doi:10.1371/journal.pone.0066463.g001

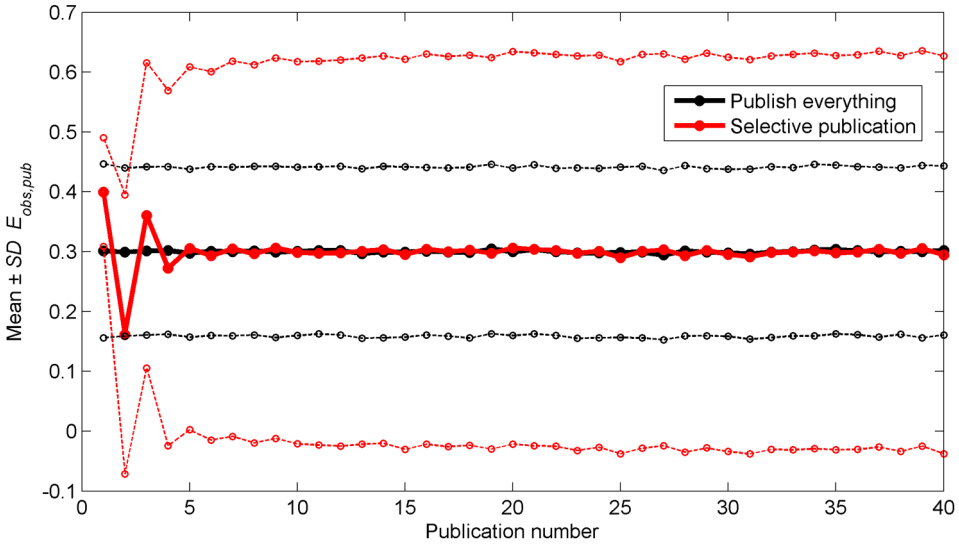

Figure 2. Mean (thicker solid lines) and mean plus/minus one standard deviation (thinner dashed lines) of observed effects of individual studies ($E_{obs,pub}$) as a function of the number of publications. The means and standard deviations are calculated across the 5,000 repetitions of the simulation.
doi:10.1371/journal.pone.0066463.g002

effects are published only if they differ more than 0.277 from E_{meta}. Summarizing, replicability for the Selective Publication Approach is low, as demonstrated by the over- and underestimation of the true effect for the early publications, and by the large variability of published effects around E_{true}.

Figure 3 shows that with the Selective Publication Approach, E_{meta} values based on initial publications are on average biased with respect to E_{true}. E_{meta} converges towards E_{true} after a few publications, indicating that this bias is rapidly nullified. The standard deviations in Figure 3 illustrate that E_{meta} is on average closer to E_{true} in the Selective Publication Approach than in the Publish Everything Approach. At the 40th publication, the *SD* of E_{meta} for the Publish Everything Approach is 0.0222 and the *SD* for the Selective Publication Approach is 0.0170. For the Publish Everything Approach the *SD* value of 0.0170 is reached at the 68th publication.

The results in Figures 2 and 3 are in agreement with Ioannidis [32] who stated that "in some fields of research, we may observe diminishing effects for the strength of research findings and rapid alternations of exaggerated claims and extreme contradictions" (see also [33]). The decreasing support of a scientific claim over time is more commonly known as the decline effect [34].

Figure 4 shows the same results as Figure 3, but now as a function of the number of studies instead of the number of publications. The standard deviations around E_{true} indicate that E_{meta} approximates E_{true} more closely for the Publish Everything Approach. That is, when the results are assessed per study instead of per publication, the Publish Everything Approach performs more favorably than the Selective Publication Approach.

The mean number of studies until publication can be seen in Figure 5. In the Publish Everything Approach, this value equals 1 because each study is published. For the Selective Publication Approach, the probability of publication decreases with publication number, that is, when consensus establishes. The value converges to 20 (i.e., $1/\alpha$), meaning that the literature eventually grows 20 times as fast when publishing everything as compared to the Selective Publication Approach. The number of publications for the Publish Everything Approach is on average 704 ($SD = 112$), whereas the number of publications for the Selective Publication Approach is 40 for each repetition. The corresponding *SD*s of E_{meta} (i.e., after publishing on average 704 and 40 studies) are 0.0053 and 0.0170 for the Publish Everything Approach and Selective Publication Approach, respectively.

The simulation code is provided as Supporting Information S1 and may be used for testing the effect caused by altering the simulation parameters. For example, with a stronger true effect, $E_{true} = 1$ instead of $E_{true} = 0.3$, the statistical power for the first publication of the Selective Publication Approach becomes virtually 1, meaning that the first study is always published. Accordingly, the over- and underestimation pattern does not occur, but the extreme opposite refutations and the comparative advantage of the Selective Publication Approach in terms of *SD* of E_{meta} (cf. Figure 3) remain. In contrast, when using $\alpha = 0.01$ instead of $\alpha = 0.05$, statistical power decreases, and the systematic bias of E_{meta} for the early publications has larger amplitude and takes more publications to fade out.

Simulation of Inadequate Synthesis of the Literature

In reality, researchers may not adequately synthesize the available literature. For example, researchers may not adapt their null hypothesis and simply continue to publish all results that differ statistically significantly from 0. Figure 6 illustrates that this would yield a systematic bias for the Selective Publication Approach; E_{meta} is inflated, being about 0.10 greater than E_{true} (0.3), and does not converge to E_{true} as in Figure 3.

Another example of inadequate synthesis of the literature is ignoring published evidence. Figure 6 shows the effect of ignoring the 3 latest publications in the Selective Publication approach. The first 4 publications accumulate confidence in an exaggerated effect, and from the 5th publication results converge to E_{meta} with a substantial delay and overshoot compared to the results in Figure 3. Our simulation ignored the 3 latest publications, which was considered a realistic situation. If a larger value than 3 is chosen, the period of the oscillation seen in Figure 6 will increase.

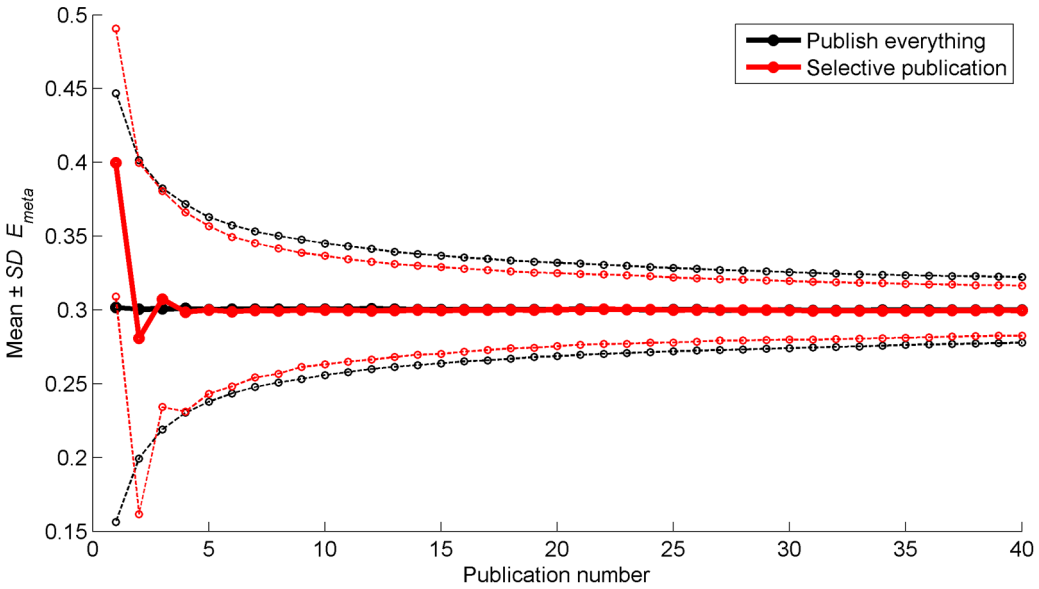

Figure 3. Mean (thicker solid lines) and mean plus/minus one standard deviation (thinner dashed lines) of the cumulative meta-analytic effect (E_{meta}) as a function of the number of publications. The means and standard deviations are calculated across the 5,000 repetitions of the simulation.
doi:10.1371/journal.pone.0066463.g003

Discussion

Our simulation study showed that instead of publishing everything, it is worthwhile to be selective and publish only research findings that are statistically significant. After a number of publications, selective publishing yields a more accurate meta-analytic estimation of the true effect than publishing everything (Figure 3). In other words, publishing nonreplicable results while placing null results in the file drawer can be beneficial for the scientific collective.

Our simulation assumed that science is self-correcting. That is, we assumed that researchers are committed to questioning and refuting previous publications. In some research fields, studies may be more likely to be published as long as the observed effect differs statistically significantly from 0, yielding a systematic bias of the cumulative meta-analytic effect (Figure 6). Another problem is that, in certain research fields such as social and behavioral

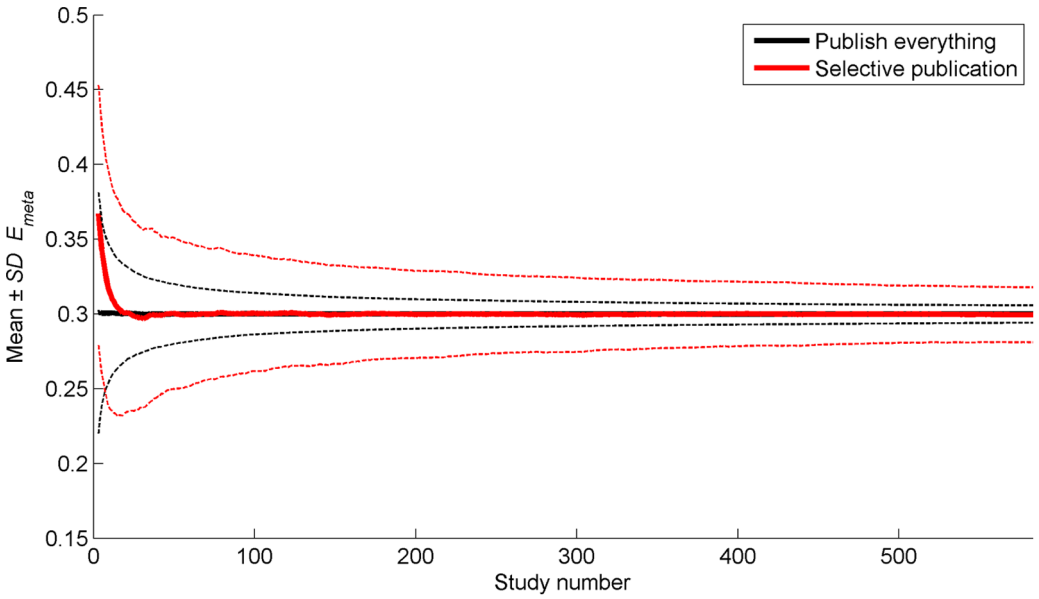

Figure 4. Mean (thicker solid lines) and mean plus/minus one standard deviation (thinner dashed lines) of the cumulative meta-analytic effect as a function of the number of studies. Note that the number of studies can vary per repetition because the simulation was terminated when 40 publications were done under the Selective Publication Approach. Only studies having more than 4,500 out of 5,000 E_{meta} values available are shown (i.e., study numbers 3–585). The means and standard deviations are calculated across the repetitions of the simulation.
doi:10.1371/journal.pone.0066463.g004

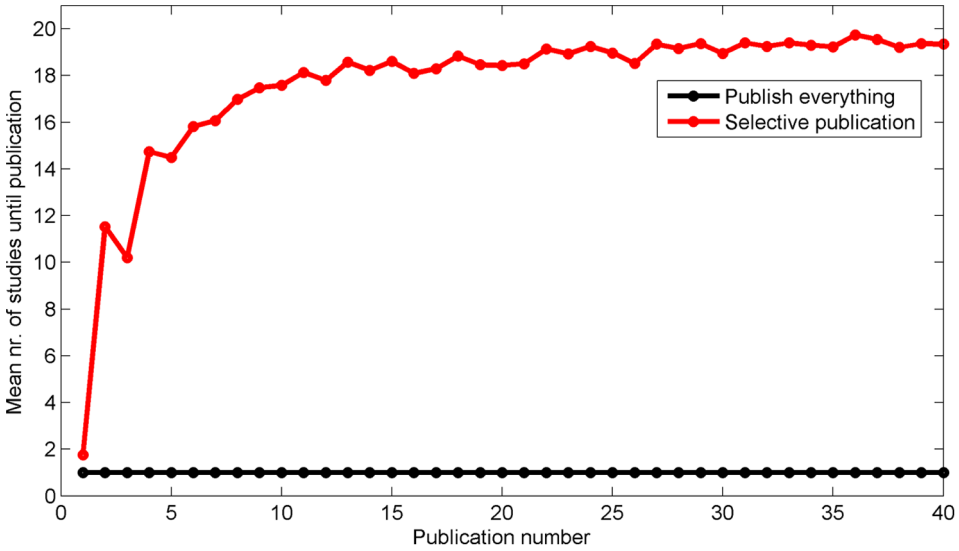

Figure 5. Mean number of studies until publication as a function of the number of publications. The means are calculated across the 5,000 repetitions of the simulation.
doi:10.1371/journal.pone.0066463.g005

sciences [35], replication studies may be unlikely and cumulative meta-analyses may never be done, resulting in unchallenged fallacies (cf. [36]). For example, if the true effect equals 0, the first selectively published effect will always deviate strongly from the true effect, and replication studies are required to refute this published claim. Because the self-correction assumption is probably untenable in many fields of science, we do not encourage selective publication. In line with this, we argue that the problem is not that researchers are averse to null results. The problem is the prejudiced researcher and the researcher who ignores or misrepresents previously published evidence on the same topic. Accordingly, efforts should go toward enhancing the self-correction mechanism and conducting a comprehensive literature synthesis prior to doing experiments.

According to our simulation, the Publish Everything Approach implies that content density of the literature database, defined as the information gained after synthesizing a given number of publications, will be suboptimal. Specifically, 68 publications were needed for the Publish Everything Approach for reaching the level of meta-analytic accuracy (i.e., *SD* of E_{meta}) obtained after 40 publications in the Selective Publication Approach. Selective publishing yields a more accurate estimation of the true effect than publishing everything, as a function of the number of publications. However, publishing everything will yield a more accurate

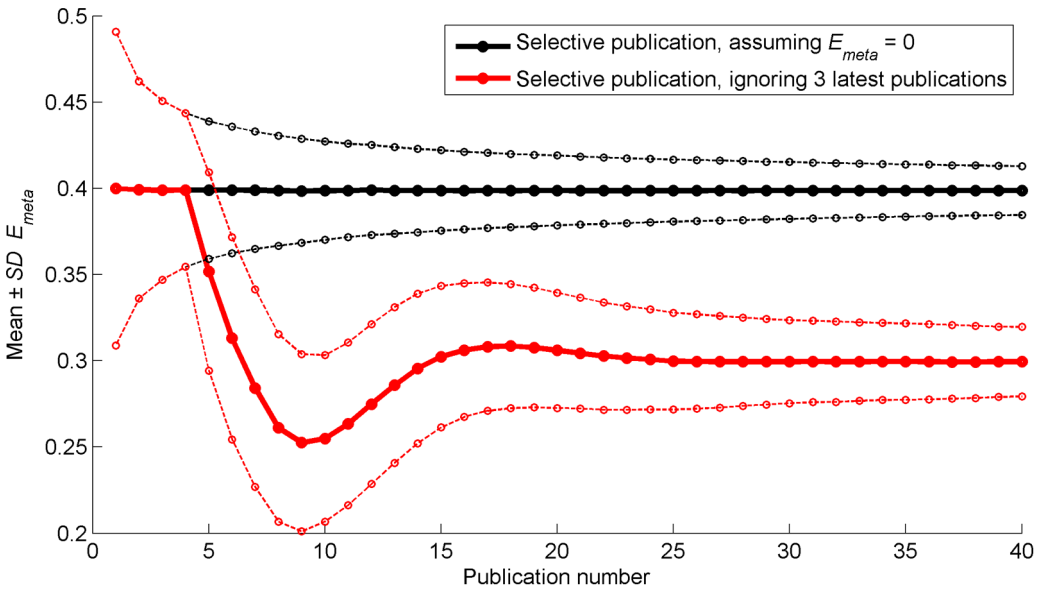

Figure 6. Mean (thicker solid lines) and mean plus/minus one standard deviation (thinner dashed lines) of cumulative meta-analytic effect as a function of the number of publications. The black lines represent the situation where E_{obs} is tested with respect to 0. The red lines represent the situation when ignoring the 3 latest publications for determining E_{meta}. The means and standard deviations are calculated across the 5,000 repetitions of the simulation.
doi:10.1371/journal.pone.0066463.g006

estimation than selective publishing, if taking into account all publications (cf. Figure 4). We argue that a given number of publications is the preferred criterion. Increasing the number of publications may place an unwanted burden on reviewers and editors, and we expect that no more than a fixed number of publications on a specific research question will be desired by a research community. This statement is in line with Nelson, Simmons, and Simonsohn [37] who argued that we should publish fewer papers in order to prevent what they called the "cluttered office effect".

Science is becoming more competitive and researchers are pressured to publish frequently and in highly ranked journals, a phenomenon which has been associated with a rising prevalence of statistically significant effects in research journals [38]. We suggest that publication of significant effects, and the corresponding Proteus phenomenon, may in some cases be desirable or even optimal. Young, Ioannidis, and Al-Ubaydli [18] stated that we may have to "accept the current system as having evolved to be the optimal solution to complex and competing problems". An analogy can be made with control theory, a discipline in the engineering sciences that deals with the behavior of dynamical systems and which is concerned with finding corrective actions that effectively reduce the sensed discrepancy between the system state and a reference value. Scientific discovery may be seen as an endeavor that minimizes the error between the prevailing opinion (E_{meta}) and a reference value, the true effect (E_{true}). Just like a person adjusting a shower spigot to reach a desired temperature (cf. [39]), researchers may publish their results in order to adjust a discrepancy between the prevailing consensus and the true effect. The strength of the corrective actions (cf. the amount of hot or cold water entering the shower) influences the rapidity with which errors are nullified, and is similar to the inverse of the α value used in the Selective Publication Approach. Selecting a low α results in a rapid response, but contributes to overshoot of the target value. A high α (e.g., $\alpha = 1$; publishing everything) results in a sluggish response. This is qualitatively similar to adjusting the shower spigot with equal rapidity irrespective of the difference between the current temperature and the target temperature. Hence, it is legitimate to respond more strongly to effects that deviate more from the null hypothesis. As also pointed out by Drummond [40] and Fiedler et al. [41], being indifferent with respect to novelty or statistical significance is counterproductive.

Supporting Information

Supporting Information S1 Simulation code.
(M)

Author Contributions

Analyzed the data: JW RH. Contributed reagents/materials/analysis tools: JW RH. Wrote the paper: JW RH.

References

1. Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, et al. (2013) Recommendations for increasing replicability in psychology. Eur J Pers 27: 108–119.
2. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124.
3. Ioannidis JP (2008) Why most discovered true associations are inflated. Epidemiology 19: 640–648.
4. Vul E, Harris C, Winkielman P, Pashler H (2009) Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect Psychol Sci 4: 274–290.
5. Wacholder S, Chanock S, Garcia-Closas M, Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J Natl Cancer Inst 96: 434–442.
6. Bertamini M, Munafò MR (2012) Bite-size science and its undesired side effects. Perspect Psychol Sci 7: 67–71.
7. Pashler H, Harris CR (2012) Is the replicability crisis overblown? Three arguments examined. Perspect Psychol Sci 7: 531–536.
8. Callaham ML, Wears RL, Weber EJ, Barton C, Young G (1998) Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. JAMA 280: 254–257.
9. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR (1991) Publication bias in clinical research. The Lancet 337: 867–872.
10. Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychol Bull 86: 638–641.
11. Pautasso M (2010) Worsening file-drawer problem in the abstracts of natural, medical and social science databases. Scientometrics 85: 193–202.
12. Sena ES, Van der Worp HB, Bath PM, Howells DW, Macleod MR (2010) Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol 8: e1000344.
13. John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci 23: 524–532.
14. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22: 1359–1366.
15. Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, et al. (2008) False-positive results in cancer epidemiology: a plea for epistemological modesty. J Natl Cancer Inst 100: 988–995.
16. Ferguson CJ, Heene M (2012) A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. Perspect Psychol Sci 7: 555–561.
17. Giner-Sorolla R (2012) Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. Perspect Psychol Sci 7: 562–571.
18. Young NS, Ioannidis JP, Al-Ubaydli O (2008) Why current publication practices may distort science. PLoS Med 5: e201.
19. Dirnagl U (2010) Fighting publication bias: Introducing the Negative Results section. J Cereb Blood Flow Metab 30: 1263–1264.
20. Sterling TD, Rosenbaum W, Weinkam J (1995) Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. Am Stat 49: 108–112.
21. Suñé P, Suñé JM, Montoro JB (2013) Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. PLoS One 8: e54583.
22. Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan A-W, et al. (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 3: e3081.
23. Davison M, Nevin JA (2005) On science and the discriminative law of effect. J Exp Anal Behav 83: 85–92.
24. Ioannidis J (2012) Reporting and reproducible research: Salvaging the self-correction principle of science. Freiburg, Germany: Annual Lecture given at the EQUATOR Network Scientific Symposium. Available: http://www.equator-network.org/index.aspx?o = 5599. Accessed 1 March 2013.
25. Binfield P (2009) PLoS One: Background, future development, and article-level metrics. ELPUB2009 Conference on Electronic Publishing. Milan, Italy. 69–86.
26. Bronowski J (1979) The origins of knowledge and imagination: Yale University Press.
27. Ioannidis JP, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol 58: 543–549.
28. Pfeiffer T, Bertram L, Ioannidis JP (2011) Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. PLoS One 6: e18362.
29. Cook C (2010) Mode of administration bias. J Man Manip Ther 18: 61.
30. Ioannidis J (2013) Clarifications on the application and interpretation of the test for excess significance and its extensions. J Math Psychol: in press.
31. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP (2011) Public availability of published research data in high-impact journals. PLoS One 6: e24357.
32. Ioannidis JP (2006) Evolution and translation of research findings: From bench to where. PLoS Clin Trials 1: e36.
33. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306–309.
34. Schooler J (2011) Unpublished results hide the decline effect. Nature 470: 437.
35. Makel MC, Plucker JA, Hegarty B (2012) Replications in Psychology Research How Often Do They Really Occur? Perspect Psychol Sci 7: 537–542.
36. Ioannidis JP (2012) Why science is not necessarily self-correcting. Perspect Psychol Sci 7: 645–654.
37. Nelson LD, Simmons JP, Simonsohn U (2012) Let's publish fewer papers. Psychol Inq 23: 291–293.
38. Fanelli D (2012) Negative results are disappearing from most disciplines and countries. Scientometrics 90: 891–904.

39. Jagacinski RJ (1977) A qualitative look at feedback control theory as a style of describing behavior. Hum Factors 19: 331–347.
40. Drummond C (2009) Replicability is not reproducibility: Nor is it good science. International Conference on Machine Learning. Montreal, Canada. Available: http://www.csi.uottawa.ca/~cdrummon/pubs/ICMLws09.pdf. Accessed 1 March 2013.
41. Fiedler K, Kutzner F, Krueger JI (2012) The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. Perspect Psychol Sci 7: 661–669.