aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
7(6) 597–599
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691612463399
http://pps.sagepub.com
SAGE

It Does Not Follow: Evaluating the One-Off Publication Bias Critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press)

Uri Simonsohn
The Wharton School, University of Pennsylvania

Abstract

Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press) attacks individual papers through critiques that apply faulty logic to analyses ironically biased by cherry picking. However well intentioned, the critiques are probably counterproductive to their stipulated goal and certainly unfair to the targeted authors.

Keywords

publication bias

In a growing number of critiques, Francis (2012a, 2012b, 2012c, 2012d, 2012e, in press) reports statistically significant evidence of publication bias tests in individual psychology papers and invites readers to completely ignore the results from those papers. These analyses are ironically biased because of cherry picking. The published critiques are but a small subset of those attempted: the subset with $p < .1$. More important, even if the analyses were correct, the conclusion to ignore evidence simply does not follow from the presence of publication bias; this fallacious inference confuses practical with statistical significance.

Cherry Picking

Replicate or show your drawer

Publication bias, the tendency to publish only statistically significant evidence, reduces the validity of reported *p* values. In the extreme, journals could be filled with the 5% of studies that are false positive, and authors' file drawers could hold the remaining 95% (Rosenthal, 1979).

We can think of publication bias as a multiple comparisons problem. Suppose we run two studies but only one worked. The *p* value we ought to compute and report is the probability of one of the two studies working— not the default *p* value, which is the likelihood that one study works if only one study is run.

When studies are statistically independent (e.g., based on different sets of subjects), the likelihood that at least one of two works, under the null, is $1-.95^2$ or 9.75%. For one of three it is $1-.95^3$ or 14.3%. If we try enough studies we are all but guaranteed to get at least one to work. For example, with 44 attempts, 90% of the time ($1-.95^{44}$) we will get at least one to work.

Although we should, we never do disclose—let alone correct for—the size of our file drawers. Instead, we address this problem with replications. Even if we got a study to work only after 44 attempts, there is still just a 5% chance of it working again under the null: replication *p* values are kosher.[1] When cherry picking, then, we should replicate or show our drawer.

Francis cherry picks but neither replicates nor shows his drawer

The question being asked in the critiques is "does Paper X suffer from publication bias?" Such a question does not lend itself to a replication because there is only one Paper X. We must, then, look into the file drawer and take into account how many papers other than X were tested.

This number, Francis acknowledges, is considerably greater than 0. Figure 1 plots the impact on false-positive rates as we increase the number of total attempts when, as is the case in the critiques, $p < .1$ is considered significant. While contemplating Figure 1, imagine how easy it is to arrive at false-positive evidence of publication bias when one is unconstrained by number of attempts (imagination aid: most journals have more than 10 papers per issue).[2]

Note that it is irrelevant whether we think of the study that worked as conceptually related to the failures before it or not.

Corresponding Author:
Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

Downloaded from pps.sagepub.com at Statsbiblioteket on November 13, 2012

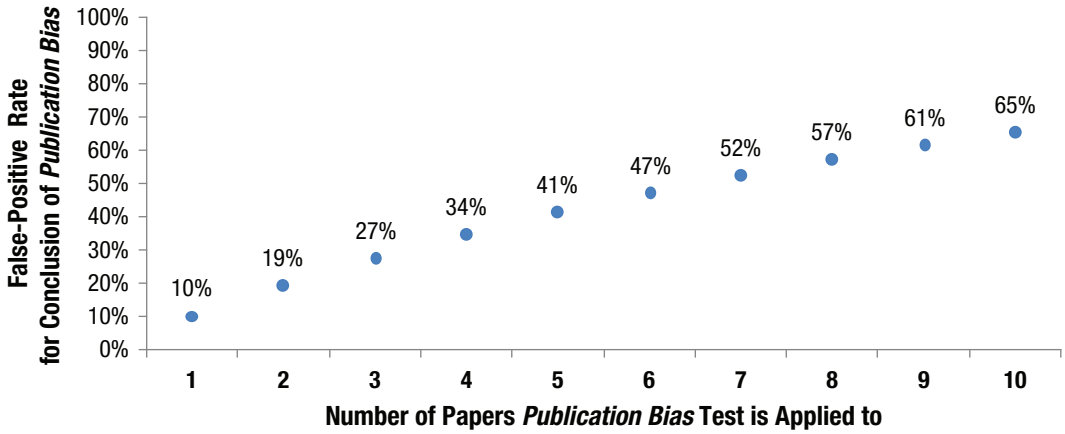

Fig. 1. Likelihood of at least one analysis of publication bias obtaining $p < .1$ as more and more papers without publication bias are examined. The *y* axis reports the likelihood of at least one analysis resulting in $p < .1$ (under the null); this probability is $(1-.9^x)$ where *x* is the number of papers examined. For example, somebody applying the publication bias test to 10 papers would have a 65% chance of concluding at least one of them has publication bias, when none of them do in fact suffer from it.

The math involved in compounding *p* values is the same when studies are about the same topic, and when they are not, and when the studies involve experiments or publication-bias tests. There is no way around it. Because the critiques were cherry picked without conducting replications, their *p* values are larger than reported.

The "my analyses are not supposed to generalize" defense

In personal communications, and in responses to authors whose work he has critiqued, Francis argues that it is fine for him to cherry pick. For example:

> [critiqued authors] suggested that my investigations of publication bias engage in the very practice that I criticize. I would be susceptible to this criticism if I were making inferences about publication bias for the field in general.[3]

We are concerned with a herring of a different color. We worry not whether the conclusion about Paper X applies also to Papers Y and Z. We worry that cherry picking increases the false-positive rate for Paper X (because it does).

Ignore the Advice to Ignore the Data

Let's now consider the conclusion that if a paper has statistically significant evidence of publication bias its findings should be ignored.

> Whatever its source, the presence of a publication bias means that the findings […] do not provide useful information about the claimed effect. (Francis, 2012a)

> Now that the data are known to be contaminated with publication bias, […]. Researchers […] are advised to ignore the findings […] and run new experiments. (Francis, 2012d, p. 177)

The conclusions are at odds with meta-analyses textbooks that propose correcting for publication bias, or assessing its potential impact, rather than eliminating data (see, e.g., Cooper, Hedges, & Valentine, 2009; Pigott, 2012; Rothstein, Sutton, & Borenstein, 2005). Francis has argued that because corrections are imperfect, we should not attempt them. The alternative of dropping all data, however, is merely another (even more imperfect) correction. Furthermore, because all journals exhibit some degree of publication bias, the logic behind this correction leads to the absurd conclusion that all published scientific knowledge should be ignored.

At its core, the "delete-all" correction confuses statistical with practical significance. Consider a literature with 100 studies, all with $p < .05$, but where the implied statistical power is "just" 97%. Three expected failed studies are missing. The test from the critiques would conclude there is statistically significant publication bias; its magnitude, however, is trivial. Ignoring the 100 studies is unwarranted by evidence and more generally counterproductive for the advancement of knowledge.

The test used by Francis merely examines if publication bias is present, not how consequential it is. We should hence draw conclusions only as to whether publication bias is present, not how consequential it is.

No Contradiction

Responding to an early draft of this article, Francis noted an apparent contradiction: I would appear to argue that publication

Downloaded from pps.sagepub.com at Statsbiblioteket on November 13, 2012

bias invalidates his critiques but not the criticized papers.[4] Let's decompose the critiques into a premise (critiqued Paper X suffers from publication bias) and a conclusion (Paper X contains no valid data). I have argued that the evidence for the premise is tainted by cherry picking and that the conclusion does not logically follow from the premise. Because only the critiques, not the criticized papers, suffer from such faulty logic, only the critiques, not the criticized papers, ought to be ignored. There is no contradiction.

By the Way, of Course There is Publication Bias

Virtually all published studies are significant (see, e.g., Fanelli, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995), and most studies are underpowered (see, e.g., Cohen, 1962). It follows that a considerable number of unpublished failed studies must exist. With this knowledge already in hand, testing for publication bias on paper after paper makes little sense: One is guaranteed to eventually reject a null we already know is false, but whose rejection tells us nothing about what we ultimately need to know—whether that specific finding is likely to be.

Jacob Cohen (1994) titled an article discussing shortcomings in our understanding of hypothesis testing "The Earth is Round ($p < .05$)." A similarly obvious statement followed by an italicized p is: "Some Failed Studies Are not Published ($p < .1$)".

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. Replications' p values are often not really kosher because of undisclosed flexibility in data collection and analyses (Simmons, Nelson, & Simonsohn, 2011); this does not reduce the importance of replications, but rather, highlights the importance of properly reporting them.
2. The issues raised in this note are generic enough that they do not require getting into the specifics of the test Francis employs. This footnote will do. Francis applies the excessive significance test first put forward by Ioannidis and Trikalinos (2007), in which the share of studies that are significant is compared with the statistical power of those analyses. If the share that is significant is significantly higher, one may conclude that publication bias is present. This test suffers from important limitations that may deem it invalid: on the one hand, it ignores sampling error in the estimation of power, making it anticonservative, but on the other hand, it takes power estimation at face value, making it conservative. The net effect of both issue has not been formally explored yet, but for small sets of studies, the anticonservative impact will likely dominate. For simplicity of exposition, however, I assume the test is valid if the likelihood of a $p < .1$ under the null is 10%. Regardless of what it really is, if we can run enough tests, the false-positive rate will eventually raise to 100%, so the comments made here are not dependent on the validity of the test.
3. http://i-perception.perceptionweb.com/misc/i03/i0519ic–greg_response.doc
4. His exact words were "If he [Simonsohn] really believes that researchers should compensate for the presence of bias (his second criticism), then the existence of bias in my investigations should only lead to a call for a correction, not a claim that the investigations are invalid (his first criticism)" (personal communication, July 29, 2012).

References

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

Cohen, J. (1994). The earth is round ($p <$. 05). *American Psychologist*, *49*, 997.

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Francis, G. (2012a). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences, USA*, *109*, e1587. Advance online publication. doi: 10.1073/pnas.1203591109

Francis, G. (2012b). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*, 585–594.

Francis, G. (2012c). Publication bias and the failure of replication in experimental psychology. *Psychomic Bulletin & Review*. Advance online publication. doi:10.3758/s13423-012-0322-y

Francis, G. (2012d). The same old new look: Publication bias in a study of wishful seeing. *i-Perception*, *3*(3), 176–178.

Francis, G. (2012e). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156.

Francis, G. (in press). Publication bias in "Red, Rank, and Romance in Women Viewing Men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*.

Ioannidis, J., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–253.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer Verlag.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. Malden, MA: Wiley.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

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*, 30–34.

Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*, 108–112.

Downloaded from pps.sagepub.com at Statsbiblioteket on November 13, 2012