aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

False-Positive Citations

Perspectives on Psychological Science
2018, Vol. 13(2) 255–259

DOI: 10.1177/1745691617698146
www.psychologicalscience.org/PPS
SAGE

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract
We describe why we wrote "False-Positive Psychology," analyze how it has been cited, and explain why the integrity of experimental psychology hinges on the full disclosure of methods, the sharing of materials and data, and, especially, the preregistration of analyses.

Keywords
research methods, *p*-hacking, replicability, preregistration

> Although our sample size may appear small to some readers, it is important to note that we obtained the necessary power and representativeness to generalize from our results while purposefully avoiding an unnecessarily large sample that could have biased our results toward a false-positive type I error (Simmons, Nelson, & Simonsohn, 2011).
>
> —Article published in the *Proceedings of the National Academy of Sciences* in 2012

When we wrote "False-Positive Psychology" (Simmons et al., 2011), we believed that it was unlikely to be published, less likely to be read, virtually uncitable, and generally best of service as a three-authored effort at catharsis.

Five years later, we have been asked to write about our article for a special issue on the most-cited APS articles. We are supposed to summarize the article, say how we came up with the idea to write it, comment on its influence, and identify what we wish we had done differently. Along the way, we will also look into how we have been cited. The results of that exercise are a good reminder that not every citation is something to be proud of.

Why We Wrote the Article

In 2010 or thereabouts, we stopped believing that many published findings were true. We discussed recently published articles in our weekly journal clubs (we were all at different universities then), and those discussions frequently devolved into statements of disbelief. We did not think the findings were fraudulent, but it was just impossible to believe that, with only 14 participants per cell, researchers had found that people will pay more for a chocolate bar when it is presented at a 45° angle, but only if they are below the median on the self-monitoring scale.[1] When results in the scientific literature disagree with our intuition, we should be able to trust the literature enough to question our beliefs rather than to question the findings. We were questioning the findings. Something was broken.

After much discussion, our best guess was that so many published findings were false because researchers were conducting many analyses on the same data set and just reporting those that were statistically significant, a behavior that we later labeled "*p*-hacking" (Simonsohn, Nelson, & Simmons, 2014). We knew many researchers—including ourselves—who readily admitted to dropping dependent variables, conditions, or participants to achieve significance. Everyone knew it was wrong, but they thought it was wrong the way it is wrong to jaywalk. We decided to write "False-Positive Psychology" when simulations revealed that it was wrong the way it is wrong to rob a bank.

Corresponding Author:
Joseph P. Simmons, University of Pennsylvania, 500 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: jsimmo@wharton.upenn.edu

Influence

An article cannot be influential if it is not read, and no one likes to read boring or hard-to-understand articles. So we tried to make sure that our article was accessible and at least a little bit entertaining. To help accomplish this, we ran two experiments demonstrating how *p*-hacking could allow us to find significant evidence for an obviously false hypothesis. It was not hard to generate a false hypothesis to test, but in a field that seemed ready to believe in lots of things, it was hard to generate one that was obviously false. We eventually decided to test whether listening to a song could change somebody's age, as we thought it unlikely that psychologists would believe that listening to "When I'm 64" or "Hot Potato" truly altered people's ages. Inevitably, the experiments "worked."

We also knew that our article could not lead to real change if we just complained about the problem. So we spent a long time thinking about solutions, seeking out one that would require the least of researchers and journals while achieving the most for knowledge and truth. We eventually identified a solution that seemed trivially easy to implement and impossible to oppose: asking authors to simply describe what they did in their studies. Specifically, we proposed that authors be required to disclose how they arrived at their sample sizes and to report all of their measures, manipulations, and exclusions. We were ready to be ignored, but we were not ready to be opposed. How could any half-serious scientist actively oppose a rule requiring authors to accurately describe their research?[2]

More famously (or perhaps infamously), we also recommended that authors be required to have at least 20 observations per cell (or to justify why they did not). This recommendation was universally hated because 20-per-cell is arbitrary (as all cutoffs are), but it should have been hated because 20-per-cell is a comically low threshold, insufficient to detect in a representative sample that men are heavier than women (Simmons, Nelson, & Simonsohn, 2013). This requirement is the least important of the set, because if you are not *p*-hacking, then you will be forced to get much larger samples anyway. But it garnered the most attention, and, as detailed below, it made it much easier for researchers to cite us . . . in order to boast about how they collected 21 participants per cell.

Our article has had some influence. Many psychologists have read it, and it is required reading in at least a few methods courses. And a few journals—most notably, *Psychological Science* and *Social Psychological and Personality Science*—have implemented disclosure requirements of the sort that we proposed (Eich, 2014; Vazire, 2015). At the same time, it is worth pointing out that none of the top American Psychological Association journals have implemented disclosure requirements and that some powerful psychologists (and journal interests) remain hostile to costless, common sense proposals to improve the integrity of our field.

The attention that our article has received undoubtedly owes more to forces outside of our control than to forces within it. Yes, our article had to be readable and actionable, but it also had to resonate with those who read it. And it did. The article happened to come at a time when many psychologists were ready and willing to accept our message, as evidenced by the fact that others were independently beginning to write about these issues (e.g., John, Loewenstein, & Prelec, 2012). The discovery of a prominent case of serial fraud (Diederik Stapel) and the *Journal of Personality and Social Psychology*'s publication of a transparently outlandish finding (Bem, 2011) helped. It also did not hurt that the article quickly became controversial, thanks in large part to public denouncements by Professor Norbert Schwarz, first during the "False-Positive Findings" symposium at the 2012 Society for Personality and Social Psychology (SPSP) conference and then again in an e-mail he sent to several thousand members of the SPSP community.[3] To know what all the fuss was about, you had to read "False-Positive Psychology."

Our field has changed a lot since then. Most notably, there is now a dramatically increased focus on replicability and transparency. In 2010, approximately 0% of researchers were disclosing all of their methodological details, posting their data and materials, and preregistering their studies. Today, disclosure, data posting, and preregistration are slowly becoming the norm, particularly among the younger generation of researchers. We would like to think that our article had something to do with all of this, but honestly, it is impossible to say, because hundreds of psychologists have worked incredibly hard to improve our science. Without them, our article would have had no influence whatsoever. And without our article, these changes may have happened anyway. It was time.

It *is* time.

How We Have Been Cited

At the time of this writing, "False-Positive Psychology" has been cited 887 times according to the Web of Science. We took a look at how we are cited and discovered some fun facts.

The article has been cited in 380 different journals. You are reading the journal that has cited it the most (thanks in part to its endorsement by Bobbie Spellman, the editor at the time), but the other tail of the citation-count distribution is more interesting. If you eliminate

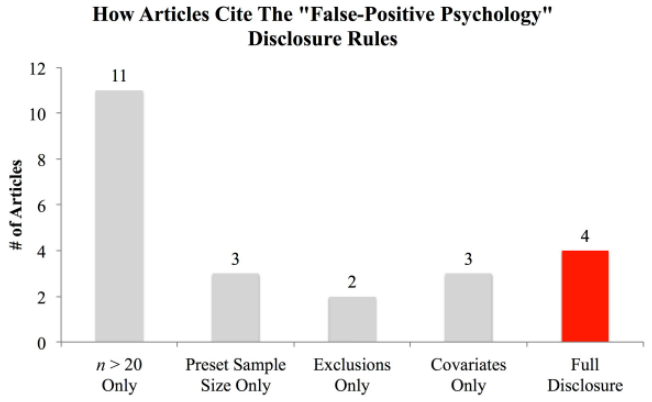

Fig. 1. Empirical articles in the *Journal of Experimental Psychology: General*, the *Journal of Personality and Social Psychology*, and *Psychological Science* that have cited each (or all) of the disclosure requirements of "False-Positive Psychology."

self-citations, it has been cited one time each in the journals *Applied Thermal Engineering*, the *Journal of Wildlife Management, Metaphor and Symbol, Microcirculation, Rapid Communications in Mass Spectrometry*, and the *Journal of Consumer Research*, the flagship journal of our own field of consumer behavior.[4] So "False-Positive Psychology" has arguably had no greater influence on a field whose conferences we attend and whose PhD students we hire than it has on the field of wildlife management. Mission accomplished.

Because we expected "False-Positive Psychology" to be hard to cite, we were curious about just how it is being cited, particularly in articles that report new studies rather than in articles that are about research methods. To get a sense of this, we looked up all of the citations within the *Journal of Experimental Psychology: General* (26), the *Journal of Personality and Social Psychology* (23), and *Psychological Science* (18). For analysis, we excluded 19 nonempirical articles, an additional 4 articles that were coauthored by one of us, and 1 article that referenced our article without actually citing it (a literal false-positive citation).

Among the remaining 44 articles, "False-Positive Psychology" was cited 12 times to say something like "*p*-hacking exists" or "somebody else *p*-hacked" or "the field is in a crisis," 4 times to say something like "replications are a good idea,"[5] and twice to say something like "my results are robust to doing an arcsine transformation." The remaining 23 citations reference us for our disclosure recommendations, and this is where we would like to focus our attention (we posted the relevant quotes here: https://osf.io/uw6m4/).

Ideally, researchers who cite us to say that they are following our disclosure recommendations will cite us to say that they are following *all* of our disclosure recommendations. Instead, researchers seem to be choosing among them. As illustrated in Figure 1, the modal disclosure-related citation references only the least important of our recommendations: having more than 20 participants per cell. For example, "Sample size was predetermined to be between 20 and 30 participants per cell, based on the recommendations of at least 20 per cell (Simmons et al., 2011)," and "Consistent with the recommendation of Simmons et al. (2011), . . . we stopped collecting new data when all cells contained at least 20 people." Moreover, many of them are barely following that recommendation, collecting just a shade above 20 participants per cell. This is not exactly what we had in mind.

Only 4 articles in the set of 23 cited us for saying that they were disclosing all measures, conditions, exclusions, and their sample-size rule; 1 of those was coauthored by one of Leif's former doctoral students, and another explicitly decided to get only 15 participants per cell. Of course, it is possible that many more than 4 of these articles fully disclosed all of their methodological details. But if they did, that is not why we were cited.

Our conclusion, then, is perhaps an obvious one: Although we are fortunate that "False-Positive Psychology" has been widely read and frequently cited, it is not the case that every citation is a cause for celebration. Although a citation necessarily indicates that a researcher was influenced to reference the article, it does not necessarily indicate that the researcher was influenced to carry out the intentions of the article.

What We Would Have Done Differently

Sample-size recommendations

If we went back in time to 2010, we would not recommend that authors be required to have more than 20 observations per cell, because that led people to focus on the wrong aspect of disclosure. Instead, we would emphasize that you cannot consistently get underpowered studies to work without *p*-hacking (or an implausible amount of luck). Thus, underpowered studies are diagnostic of *p*-hacking and do not constitute sufficient evidence for the existence of an effect (Nelson, Simmons, & Simonsohn, 2018).[6]

In addition, we would modify the $n > 20$ rule in two ways. First, we would choose a larger reference point. We now know that even obviously true effects, such as "People who like eggs report eating egg salad more often than those who dislike eggs," are not large enough to be consistently detectable with fewer than 50 participants per cell, and an obvious effect, such as "Smokers think that smoking is less likely to result in death than do nonsmokers," is not consistently detectable with fewer than 150 per cell (Simmons et al., 2013).

Table 1. The Easy Preregistration Questions on AsPredicted.org

1.	Have any data been collected for this study already?
2.	What's the main question being asked or hypothesis being tested in this study?
3.	Describe the key dependent variable(s) specifying how they will be measured.
4.	How many and which conditions will participants be assigned to?
5.	Specify exactly which analyses you will conduct to examine the main question/hypothesis.
6.	Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
7.	How many observations will be collected or what will determine sample size?
8.	Anything else you would like to preregister?

Second, we would not advocate for a strict sample-size cutoff; rather, we would emphasize that samples smaller than these are usually diagnostic of insufficient power. Thus, we would suggest that you should assume that a study with fewer than 50 per cell is underpowered unless there is evidence to the contrary. For example, the study may have used a highly powered within-subjects design, or it may be a pilot for an obvious manipulation, or it may follow a large-sampled study that showed the effect size to be very large. On the flip side, the study may be investigating a between-subjects attenuated interaction, in which case the researchers need twice as many observations per cell to maintain the same level of statistical power as that obtained in an investigation of the two-cell simple effect. Thus, we should expect between-subjects studies of attenuated interactions to have at least 100 per cell (Simonsohn, 2014).

Preregistration

Since 2010, we have become strong believers in and advocates for preregistration, the act of specifying your manipulations, measures, and analyses of interest before any data are collected. Preregistration is a form of disclosure that has two key advantages over the version we had previously advocated for. First, it gives researchers the freedom to conduct analyses that could, if disclosed afterward, seem suspicious, such as excluding participants who failed an attention check or running an unusual statistical test. Second, it is simply a more verifiable form of disclosure. Indeed, preregistration is the only way for authors to irrefutably demonstrate that their key analyses were not *p*-hacked. When the skeptic says, "There is no way that you planned to control for father's age," the researcher can now say, "Actually, here it is specified in my preregistration," and bask in the subsequent plaudits. Preregistration makes you immune to suspicions of *p*-hacking.

Preregistration is now routine in our own labs, and, if you are in the business of collecting and analyzing new data, we see no counterargument to doing it.[7] Preregistration does not restrict the ability to conduct exploratory analyses; it merely allows the researcher and the reader to properly distinguish between analyses that were planned and exploratory. In addition, it does not prevent researchers from publishing results that do not confirm their hypothesis; the critical aspect of preregistration is not the prediction that the researcher makes but, rather, the methodological and analytical plan that the researcher specifies. It is perfectly acceptable to simply pose a research question and describe exactly how you intend to answer it, and it is perfectly acceptable to publish a finding that did not conform to your original prediction.

The only reason not to preregister future studies is if preregistrations are too difficult to generate or to read. Therefore, in December 2015, we launched the web site AsPredicted.org, which makes all aspects of preregistration spectacularly easy. In its 1st year, more than 1,000 different people completed AsPredicted preregistrations. To preregister on the site, authors answer eight easy questions (see Table 1), and a time-stamped, one-page PDF document is generated with their answers. The document includes a link to verify its authenticity online (for a sample, see https://aspredicted.org/nfj4s.pdf). To prevent ideas from being scooped, the preregistration remains private until the authors make it public. They can also make an anonymized version public for the review process.

Conclusion

If we want the output of psychological research to enhance our understanding of the world around us, we need it to be easier for scientists to document true facts than to document false "facts." We need it to be easier to document that egg-likers are more likely to prefer egg salad than to document that listening to "When I'm 64" makes people younger. Without full disclosure, this basic property of the scientific enterprise is lost. True-positive and false-positive findings are equally easy to

generate and to publish. For experimental psychology to be an actual science, we must require researchers to fully disclose the methods of the studies they publish, to post their materials and data, and to preregister the analyses of studies that have not yet been conducted. In 2011, this seemed like a prohibitively outrageous thing to say. In 2016, it is merely the obvious next step. We should embrace disclosure and preregistration as if the credibility of our profession depended on it.

Because it does.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Notes

1. This is (probably) not an actual study.
2. We were not prepared for our solution to be rejected by the *Journal of Consumer Research* editorial board on the grounds that it threatened to "dull . . . some of the joy scholars may find in their craft" (Luce, McGill, & Peracchio, 2012).
3. We have archived Professor Schwarz's message (and our response) here: https://osf.io/8u4vc/.
4. The *Journal of Marketing Research* also publishes consumer-behavior articles and is also considered top of the field; it has cited "False-Positive Psychology" seven times.
5. One of these articles cites us to say, "Such [conceptual] replications are vital to establishing the reliability of research findings (Simmons et al., 2011)," although in our article, we wrote that conceptual replications "are unfortunately misleading as a solution to the problem" (p. 1365).
6. In hindsight, we also regret that we were not explicit about the fact that our sample size recommendation does not necessarily apply to all areas of psychology, such as those that collect many observations per person.
7. Those who are instead in the business of analyzing existing data sets cannot preregister before data collection; in such cases, we advocate for researchers to disclose all measures, exclusions, and so on. In addition, researchers analyzing existing data should demonstrate that their results are robust to many alternate and equally valid specifications. We have begun developing a technique called *specification curve* for this purpose (Simonsohn, Simmons, & Nelson, 2015).

References

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi:10.1037/a0021524

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Luce, M. F., McGill, A., & Peracchio, L. (2012). Promoting an environment of scientific integrity: Individual and community responsibilities. *Journal of Consumer Research*, *39*(2), iii–viii.

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511–534.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013, January). *Life after* p-*hacking*. Paper presented at the 14th Annual Meeting of the Society for Personality and Social Psychology, New Orleans, LA. doi:10.2139/ssrn.2205186

Simonsohn, U. (2014). *No-way interactions*. Retrieved from http://datacolada.org/17

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. doi:10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Specification curve: Descriptive and inferential statistics on all reasonable specifications* (Working Paper). doi:10.2139/ssrn.2694998

Vazire, S. (2015). Editorial. *Social Psychological and Personality Science*, *7*, 3–7. doi:10.1177/1948550615603955