Check for updates

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
1–10
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916211072525
www.psychologicalscience.org/PPS
SAGE

Is Psychological Science Self-Correcting? Citations Before and After Successful and Failed Replications

Paul T. von Hippel
LBJ School of Public Affairs, University of Texas, Austin

Abstract

In principle, successful replications should enhance the credibility of scientific findings, and failed replications should reduce credibility. Yet it is unknown how replication typically affects the influence of research. We analyzed the citation history of 98 articles. Each was published by a selective psychology journal in 2008 and subjected to a replication attempt published in 2015. Relative to successful replications, failed replications reduced citations of replicated studies by only 5% to 9% on average, an amount that did not differ significantly from zero. Less than 3% of articles citing the original studies cited the replication attempt. It does not appear that replication failure much reduced the influence of nonreplicated findings in psychology. To increase the influence of replications, we recommend (a) requiring authors to cite replication studies alongside the individual findings and (b) enhancing reference databases and search engines to give higher priority to replication studies.

Keywords

citation, history, scientific methodology, motivation, goals, reward, replicability, replication, reproducibility

In recent years, systematic efforts have revealed that a large number of published findings cannot be replicated. Eighty-nine percent of “landmark” findings in preclinical cancer research (Begley & Ellis, 2012), 32% of highly cited clinical trials in medicine (Ioannidis, 2005), and 60% of experiments published in top psychology journals (Open Science Collaboration, 2015) have failed to produce similar results when repeated with new samples. Some failed replication studies find no effect at all; others find an effect that, though larger than zero, is much smaller than the effect reported in the original study (Ioannidis, 2008).

The prevalence of nonreplicable results poses two challenges to scientific progress. The first challenge is the widespread use of questionable research practices that increase the risk of nonreplicable findings (John et al., 2012). In response, reformers have promoted more rigorous and transparent practices, known as *open-science* reforms. Before a study begins, open-science reforms include preregistration of hypotheses (Nosek et al., 2018); formal power analysis to determine the sample size required to detect an expected effect size (Turner et al., 2018); and, in some cases, submission of a registered report motivating the hypotheses and describing plans for data collection, research design, and analysis (Nosek & Lakens, 2014), which some journals will agree to either accept or reject before the results are known (Center for Open Science, 2021). After a study is complete, open-science practices include sharing code, data, and other research materials in a publicly accessible repository (Alter & Gonzalez, 2018).

The second challenge is how the influence of nonreplicated results can be tempered. New investigators need more than sound research practices. They also rely on the published literature to judge what is known and unknown, what warrants further investigation and what can be taken for granted. If a large fraction of published findings are not trustworthy, then new scientists may be something like explorers setting off to survey the Kong Mountains or the coast of New South Greenland—discoveries claimed by early explorers and copied for decades from one map to another until later

Corresponding Author:
Paul T. von Hippel, LBJ School of Public Affairs, University of Texas, Austin
Email: paulvonhippel@utexas.edu

explorers, after wasting time on futile expeditions, convinced cartographers that there was nothing there (Brooke-Hitching, 2018).

How can we more quickly correct the map of knowledge on which any research community must rely? A first step—and a key recommendation of open science reformers—is to encourage and publish more efforts to replicate influential studies (e.g., Ioannidis, 2012). Yet publishing replication studies promotes progress only if the results affect the credibility that future investigators ascribe to the original result. A single replication failure does not necessarily make the original finding false, but if replication studies carry some weight in the research community, then an unsuccessful replication attempt should reduce faith in the finding's veracity or generalizability, whereas a successful replication attempt should bolster the original finding. At the very least, successful and failed replication attempts should be cited alongside the original finding, rewarding replicators for their efforts and giving readers a sense of the weight of evidence.

Replication can play an important role in correcting or qualifying the scientific record, and there are episodes in science history when it played that role well. In physics, multiple replication failures (reported in Taubes, 1993) immediately followed two electrochemists' claim to have achieved "cold fusion" (Fleischmann & Pons, 1989), and annual citations of the original cold-fusion article fell by 72% in the 3 years after publication (Fig. 1).[1] In social psychology, citations of an article on the effects of "power posing" (Carney et al., 2010) fell by 42% after several replication studies and meta-analyses failed to support the claim that the poses affected hormone levels or risk taking (Cesario et al., 2017, summarizes an entire journal issue on power posing; see also Ranehill et al., 2015; Simmons & Simonsohn, 2017). More broadly, repeated replication failures are causing some psychologists to question the soundness of the whole field of social priming (Chivers, 2019).

Yet there are also instances in which replication failure appears to go almost unnoticed. In sociology, for example, a Dutch graduate student discovered data errors (Stojmenovska et al., 2017) that almost completely invalidated the results of an article claiming that firms with more diverse boards were more profitable (Herring, 2009). Yet over the next 3 years, the original study received nearly 500 citations, whereas the report of the replication failure received just 11.

Even when a replication failure receives more attention, the original finding often continues to exert outsized influence. A University of Massachusetts graduate student discovered errors (Herndon et al., 2014) invalidating the claim, published by two Harvard professors in a top economic journal, that national debt exceeding 90% of gross domestic product (GDP) reduced economic growth to practically zero (Reinhart & Rogoff, 2010). Because the original claim had been used to support austerity policies during the Great Recession of 2008, the replication failure received national and international media coverage (e.g., Colbert, 2013; Krugman, 2013). Within 2 years of the replication failure, citations of the original article ceased to grow, yet the original article continued to be cited approximately 400 times per year, and one of its authors became Chief Economist of the World Bank. The replication failure was cited only one third as often.

Likewise, in education research, a Princeton assistant professor published findings (Rothstein, 2007b), first discovered while he was a graduate student at Berkeley, showing that the estimated effect of school competition on test scores, published by a Harvard associate professor in a top economics journal (Hoxby, 2000), was hard to reproduce and sensitive to changes in operational definition. The exchange that followed (Hoxby, 2007; Rothstein, 2007a), even before it was published in an academic journal, received coverage in *The Wall Street Journal* (Hilsenrath, 2005). Afterward, citations of the original article did decline somewhat, but more than a decade later, the original article continues to be cited 60 to 80 times per year, whereas the replication failure is cited only one quarter as often.

In this article, we investigate the impact of replication studies in psychology. Taking citations as a measure of the credibility and importance that the research community ascribes to a published article, we ask the following questions: What is the effect of replication failure on citations? Do psychology studies that were not replicable see their citations fall relative to similar studies that were successfully replicated?

Citation counts, by themselves, are an imperfect measure of credibility. After a replication failure, authors may continue to cite an article for various reasons. Authors may still believe the article's findings, they may believe the article contributes evidence on one side of a point that has become debatable, or they may believe the article represents a cautionary tale on the perils of questionable research practices and nonreplicable results. Although we cannot interpret the nuanced meaning of citations without reading every citation in context, a simple first-level question is whether the replication attempt is cited alongside the original article or the article continues to be cited as though no replication attempt took place.

This leads to our other questions: After a replication attempt, how often do articles that cite the original research also cite the replication attempt? Is the answer different for research was not replicated successfully than for research that was replicated successfully?

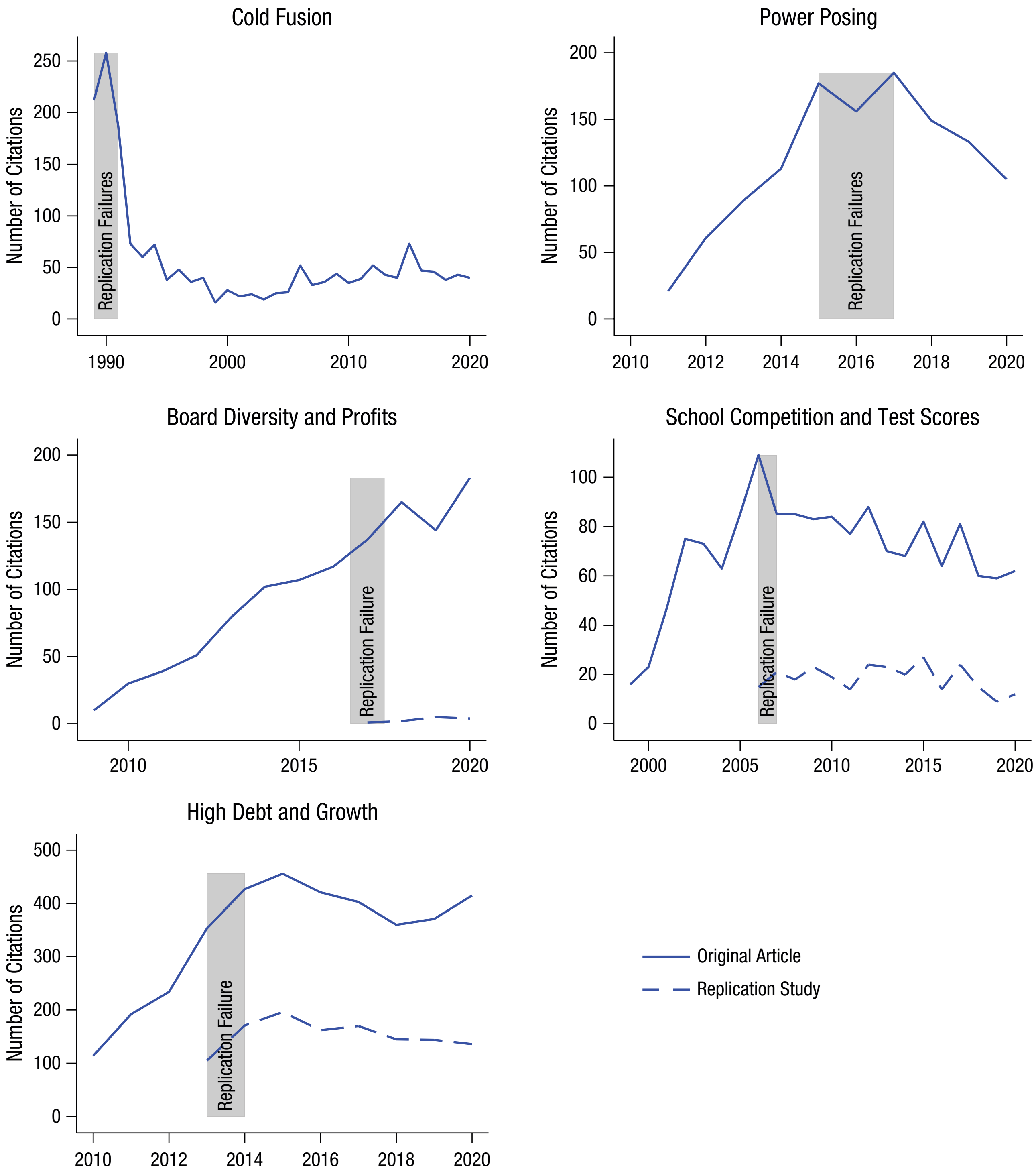

Fig. 1. Number of citations of five influential articles before and after replication failure.

All questions were preregistered at OSF in January 2020. Since then, another article has been published on the relationship between replicability and citations, but that article focused on whether nonreplicated findings were cited more than replicated ones (Serra-Garcia & Gneezy, 2021). Our article focuses on different questions, estimating the effect of replication failure on citations and developing recommendations for increasing the influence of replications in correcting the scientific record.

Method

Data

Replication studies. Starting in November 2011, the Open Science Collaboration (OSC) invited independent investigators to try to replicate 100 studies originally published in 2008 by three selective psychology journals. Replicators followed a standard protocol that involved consultation with the original authors. Replication sample sizes were chosen to have 80% power to detect the original effect at a significance level of p = .05 or better (Open Science Collaboration, 2012).

In 2015, the OSC published the results of the 100 replication studies in a widely cited article (Open Science Collaboration, 2015). At the same time, the OSC published supplementary materials at OSF, including written reports, data, and code for each of the 100 replication studies, as well as a data set containing the results of each original study and the results of the replication study (Open Science Collaboration, 2016). We used Version 7 of the data summarizing the replication results (Open Science Collaboration, 2016) and corrected two minor errors as we conducted our analysis.

The OSC employed four different operational definitions of replication success. Because we focused on how the research community perceived replication failure, our primary definition was the most subjective one, the replication's team answer to the question: "Did your results replicate the original effect?" (yes/no). Results for three alternative definitions appear in the Supplemental Material available online.

In all, investigators attempted to replicate a total of 100 studies from 98 articles—one study from each of 96 articles, and two studies from each of two articles. After averaging results to the article level, we had a data set with 98 rows summarizing the replicability of studies from 98 articles. Thirty-nine studies were successfully replicated, and 59 were not successfully replicated (including the two articles in which two studies were selected for replication; in both of those, the replication attempts were unsuccessful).

Citation data. We counted citations for every year from 2008 (when the original articles were published) through 2020. For each article and year, the data included the number of articles that cited the original 2008 article (citations) and the number of articles that cited the original 2008 article and also cited the 2015 OSC replication attempt (cocitations).

Citations and cocitations were collected by querying Google Scholar in April 2021. Citations were counted automatically using the *scholar* package (Version 0.2.0; Yu et al., 2021) for the R software environment (Version 4.0.5; R Core Team, 2021) and the Octoparse web scraper (https://www.octoparse.com/). Cocitations were counted manually. Counts were collected independently by different workers and then checked for consistency by the author.

In longitudinal data such as these, the econometric literature suggests two different models (Angrist & Pischke, 2009; Ding & Li, 2019) to estimate the causal effect of replication success or failure on citation counts.

Fixed-effects model. The first is a two-way fixed-effects regression:

$$\ln(E(Y_{it})) = \alpha_i + \beta_t + \gamma \text{ AfterFailure}_{it}. \quad (1)$$

Here, $\ln(E(Y_{it}))$ is the log of the expected citations received by article $i = 1, \ldots, 98$ in year $t = 2008, \ldots, 2020$. AfterFailure$_{it}$ is a dummy variable coded 1 after 2015 for studies that were not successfully replicated and 0 otherwise.[2] The coefficient of AfterFailure$_{it}$, γ, represents the effect of replication failure on the log of expected citations, and the transformation[3] $100 \times (\exp(y) - 1)$ reexpresses the effect as the percentage change in expected citations. α_i, a fixed effect or dummy[4] for each article, controls for all article characteristics that had no change in effect on citations over time. β_t, a fixed effect for each year, controls for time trends in citations that were similar across articles. We report robust standard errors clustered at the article level (Cameron & Miller, 2015).

Because the number of citations is a count variable, we modeled Y_{it} with a negative binomial distribution with dispersion parameter k estimated from the data (Allison & Waterman, 2002). The fixed-effects model can also be specified as an ordinary linear regression that does not log the left side of Equation 1; we report results for the linear specification in the Supplemental Material.

Fixed-effects models return consistent estimates of the effect of replication failure if the *parallel-trends assumption* holds (Angrist & Pischke, 2009). The parallel-trends assumption means that citations received by successfully and unsuccessfully replicated studies (a) changed at similar rates until 2015 and (b) would have continued to change at similar rates after 2015 had there been no replication attempt. The first part of the parallel-trends assumption is testable; the second part is untestable but equivalent to assuming that replication success or failure is uncorrelated with unobserved variables that might affect citation counts after 2015 but not before. Note that the parallel-trends assumption does not require that successfully and unsuccessfully replicated studies be cited at similar *levels*; it means only that citations change at similar *rates* and would continue to do so if no replication were attempted.

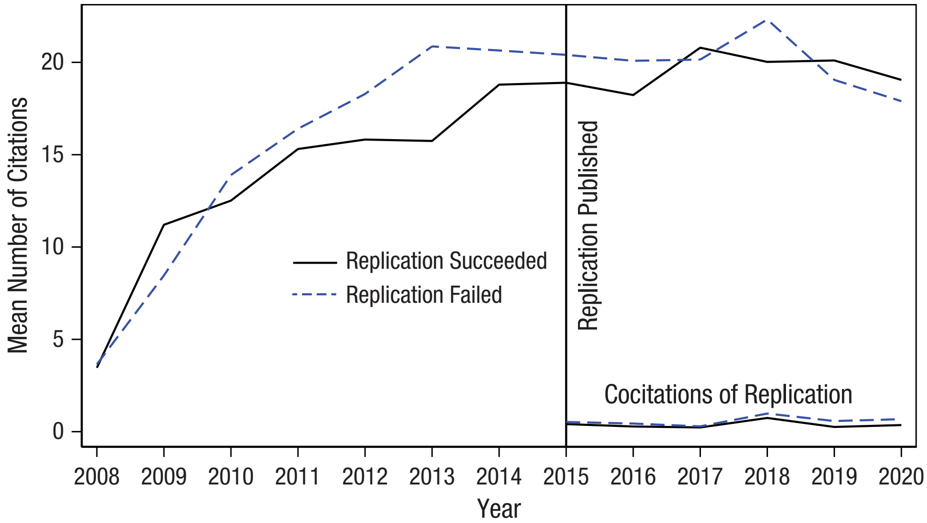

Fig. 2. Mean annual citations of 98 psychology articles (published in 2008) before and after a replication attempt (published in 2015) that either succeeded or failed to reproduce the original result. Mean cocitations of the original study and the replication attempt are also shown.

Lagged model. If the parallel-trends assumption is not met, a better model is a lagged dependent variable regression:

$$\ln\left(Y_{i,t>2015}\right) = \alpha + \beta \ln\left(Y_{i,t<2015}\right) + \gamma \text{ Failure}_i + \delta \text{ Failure}_i \times \ln\left(Y_{it<2015}\right) + e_i. \quad (2)$$

Here, $Y_{i,t>2015}$ and $Y_{i,t<2015}$ are the total citations received by article *i* in the years before and after the results of the OSC replication studies were published in 2015. Both citation counts are logged to linearize the relationship. Failure_i is a dummy variable indicating whether article *i* was unsuccessfully replicated in 2015. The coefficient γ represents the average effect of replication failure on the log of total citations after replication; because citations are logged, 100 × (exp(*y*) − 1) may also be interpreted as the percentage by which replication failure reduced total citations after replication. The model includes an interaction between Failure_i and the log of lagged citations; this allows the slope of lagged citations to be different for articles that were successfully replicated than for articles that were unsuccessfully replicated. The log of lagged citations was mean-centered to reduce the correlation between the interaction and the main effect of replication failure (Iacobucci et al., 2016).[5] We report standard errors that are robust to heteroscedasticity in the random residual e_{it} (Long & Ervin, 2000).

The fixed-effects and lagged models have a "bracketing" relationship: They estimate upper and lower bounds on the causal effect of replication failure (Angrist & Pischke, 2009; Ding & Li, 2019). We preregistered the fixed-effects negative binomial model at OSF (https://osf.io/um6fg/) and specified the other models later.

Results

Fixed-effects model

Figure 2 shows trends in the annual number of citations and cocitations per article. Both before and after the OSC replication study was published in 2015, replicable findings were cited about as often as nonreplicable findings. Specifically, before the OSC replication study, there were on average 13.3 annual citations for each article with successfully replicated results and 14.6 annual citations for each article with unsuccessfully replicated results. The small difference was not statistically significant ($t = 0.55$, $p = .58$).[6] After the OSC replication study, there were on average 19.6 annual citations for each article with successfully replicated results and 19.9 annual citations for each article with unsuccessfully replicated results. Again, the small difference was not statistically significant ($t = 0.05$, $p = .95$).

According to our fixed-effects negative binomial model, replication failure reduced citations of the replicated studies by approximately 9%, an effect that did not differ significantly from zero ($p = .2$, 95% confidence interval [CI] = [21% reduction, 5% increase]).

After 2015, less than 3% of citations of the original studies cocited the OSC replication study. On average, there were only 0.4 annual cocitations for each article with successfully replicated results and 0.6 annual

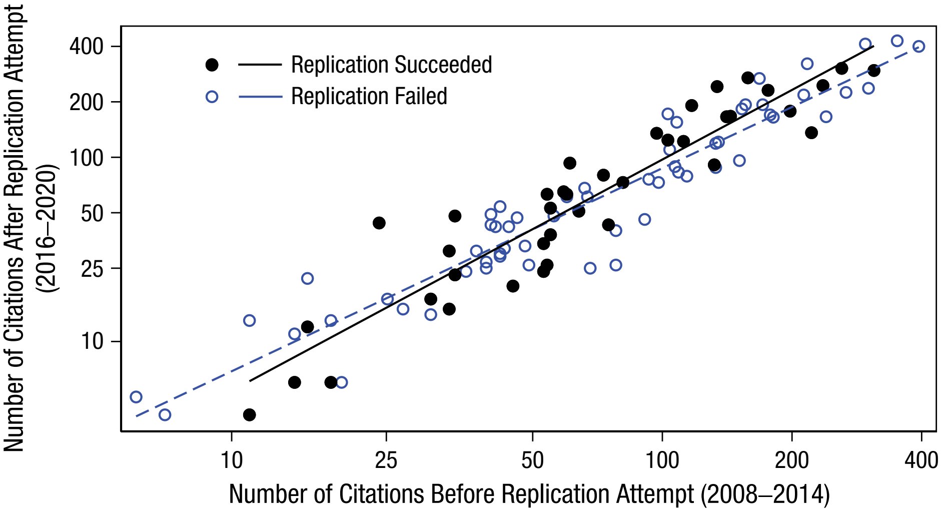

Fig. 3. For 98 psychology articles published in 2008, this scatterplot shows the total number of times that each article was cited in the years before and after a replication attempt published in 2015. Different symbols and regression lines are plotted for successful and failed replications. Axes are scaled logarithmically, and regression lines were fitted to logged citations.

cocitations for each article with unsuccessfully replicated results. The difference was not quite statistically significant, $t(91) = 1.8$, $p = .08$.

These counts actually overstate the number of times that authors cited the OSC replication studies to qualify assertions about one of the original studies. In fact, among the 248 total cocitations, 111 came from two methodological articles about how to define replication or predict whether a result will be replicable (44 cocitations in Dreber et al., 2015; 67 cocitations in van Aert & van Assen, 2018). These articles reanalyzed the OSC replication data without commenting on the substance of any of the replicated studies.

Remember that our fixed-effects regression assumes that citations of successfully and unsuccessfully replicated studies were increasing at a similar rate (i.e., had "parallel trends") before the OSC replication studies were published in 2015. Any violation of this assumption appears to be mild. Between 2008 and 2014, annual citations increased by an average of 15.3 for articles that were successfully replicated and by 17.0 for articles that were not; the difference was small and not statistically significant, $t = 0.45$, $p = .64$. Still, power to reject parallel trends was limited, and the trends in Figure 1 are not visually parallel. It is worth reestimating the effect using a lagged regression model that is more robust to nonparallel trends—as we do next.

Lagged model

Figure 3 is a scatterplot, using a log-log scale, of the number of citations each article received before replication against the number of citations it received afterward. The scatterplot includes regression lines fitted to logged citations, separately for articles that were successfully or unsuccessfully replicated. The lines represent the lagged regression model in Equation 2.

According to the lagged model in Equation 2, replication failure reduced expected citations by 5% on average—a reduction that was not statistically significant ($p = .53$, 95% CI = [20% reduction, 11% increase]). The interaction between replication failure and the log of citations was small and nonsignificant ($p = .09$), as reflected by the trivial difference between the two regression lines in Figure 3.

Discussion

Open-science reformers have encouraged scholars to conduct more replication studies, hoping that the results will help science to correct itself. If replications carry substantial weight in the research community, then successful replication studies should bolster confidence in the replicated findings, and failed replication studies should reduce confidence. The authors of replication studies should be recognized for their contribution.

Our results, though, suggest that many replication studies carry far less weight than advocates for scientific self-correction might hope. Although the OSC's replication project has done a great deal to increase the recognition of the replication crisis in general, it has done far less to shape confidence in specific psychological findings. On average, we found that replication success or failure had little or no impact on citations of the replicated studies. In the vast majority of cases, the original article continued to be cited approximately as much after the replication attempt as it was before, and more than 95% of articles citing the original study failed to acknowledge that a replication had even been attempted.

Our findings are limited to the 100 studies replicated by the OSC, and we should acknowledge that the purpose of those replication studies was unusual. The goal of the OSC replication project was to estimate the reproducibility of psychological research in general—not to call out specific findings that were or were not successfully replicated. Although the results of all 100 replication studies were made publicly available, there was no effort to draw attention to individual results. If the OSC had called out individual findings that were or were not successfully replicated, it might have had more impact on the research community's confidence in those findings. However, a project that promised to call out individual replication failures might have been received less collegially and attracted less cooperation from the authors of the replicated studies.

Outside of the OSC, though, it is not hard to find similar examples. In the introduction, we highlighted three highly cited articles that were unsuccessfully replicated, yet afterward, the number of annual citations actually rose for one article (Herring, 2009), plateaued for another article (Reinhart & Rogoff, 2010), and declined at a rate of just 2% per year for the third article (Hoxby, 2007). Two of the replication failures received mass-media coverage yet were still cited only 20% to 40% as often as the original finding (Herndon et al., 2014; Rothstein, 2007b). The third replication failure was published more quietly, and cocited, like the OSC replication studies, only 2% as often as the original finding (Stojmenovska et al., 2017).

Remember: These were replication failures that succeeded in getting published. Unpublished "file drawer" replication studies presumably have even less impact.

Why do replication failures have so little impact on citations of the original results? In addition to any confirmation bias that afflicts citing researchers themselves, an automated kind of confirmation bias is built into the search engines that researchers use to query the scholarly literature. The documentation for Google Scholar (Google, 2021), for example, states that it "aims to rank documents the way researchers do, weighing the full text of each document, where it was published, who it was written by, as well as how often and how recently it has been cited in other scholarly literature" (para. 3)—not whether the result was correct or has been replicated. Past citations receive the bulk of the weight in search engine rankings; past citations have a 0.9 correlation with the rank ordering of search results returned by Google Scholar, Microsoft Academic, and Web of Science (Rovira et al., 2019). Scopus, by contrast, practically ignores citation histories and tries to return the most relevant results, whether they are widely cited or not (Rovira et al., 2019).

What this implies is that a "classic" article by well-known authors, published in a top journal and cited hundreds of times, will typically rank well above a more recent and less-cited article, perhaps by a graduate student, that reports an unsuccessful replication of the classic result. Even as time passes, there is practically no way for a replication failure to get more citations than the original article. The original article has a head start, and it will be cited whenever the replication study is cited, although the reverse is not true. As a result, three of the four most popular search engines will tend to rank the original article higher—perhaps much higher—than any replication. The relative blindness of popular search engines to replication studies may be a kind of epistemological bug—a bug that affects not just the quality of search results but perhaps the progress of science itself.

Despite the recent surge of interest in replication studies, the rewards for would-be replicators remain uncertain. Although recent reforms may have assuaged fears that replication studies are unpublishable, even a published replication failure may have little effect on the influence of the original findings. The replication failure will almost never be cited more than the original study and will typically be cited much less. And even so, it may antagonize the authors of the replicated study.

Recommendations

What can the research community do to increase the influence of published replication studies and accelerate the potential of science to self-correct? A straightforward reform would be to ensure that published articles cite replication studies alongside the original studies. Authors could be required to cite replication studies, and editorial staff and reviewers could check for overlooked replication studies when finalizing an accepted article for publication. This would improve science in two ways. First, it would force authors and readers to better acknowledge the weight of evidence on a given topic. Second, it would increase incentives to conduct replications by making it harder to ignore replication studies and increasing the number of times that they get cited.

It might be challenging to increase citations of replication studies if leading bibliographic search engines continue to give replication studies low rankings. In the short run, authors and editors can increase their chances of finding replication studies by including keywords such as "replication" or "reanalysis" in queries, although this would catch only the most self-conscious and literal replication studies, overlooking studies that resemble each other but do not use obvious keywords. In the longer run, search engines could be modified to identify replication studies by text mining, and authors of replication studies could use tags to make their contributions more discoverable. Wider use of search engines such as Scopus, which aims to surface relevant results whether they are widely cited or not, might also lead researchers to a more balanced understanding of the literature, rather than relying on a single classic study that may not be replicable.

Third parties could maintain a searchable database of replication studies, much as Retraction Watch (http://retractiondatabase.org/RetractionSearch.aspx?) and OpenRetractions.com maintain databases of retracted studies. Like replication studies, retractions are commonly overlooked, and retracted articles are too often cited as though the result were true (Bolland et al., 2022), but new automated tools to check bibliographies for retracted items are beginning to address the problem (Chawla, 2021; Oransky, 2019; Zotero, 2019). The development of similar tools to check for overlooked replication studies would be a step forward for the integrity and progress of science.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
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.

ORCID iD

Paul T. von Hippel https://orcid.org/0000-0003-4498-4374

Acknowledgments

I thank four workers for collecting data from Google Scholar. Matt Worthington and Hatem M. collected citation data using software; Satish K. and Janaka Nuwan Priyadarshana K. collected cocitation data manually. Worthington was a staff data scientist in the LBJ School of Public Affairs, University of Texas, Austin, and the other workers were hired through Freelancer.com.

Notes

[1] According to Google Scholar, citations of the cold fusion article have rebounded slightly since 2005. This may be an artifact of Google Scholar having better coverage in more recent years. The most recent citation rate of 40 citations in 2020 is well below the 250 citations that the study garnered in 1990 and dramatically below the thousands of citations that the article would receive today if the physics community believed that Fleischmann and Pons had truly discovered cold fusion.
[2] More explicitly, AfterFailure$_{it}$ can be interpreted as an interaction between a dummy variable that is 1 for years after 2015 and a dummy variable that is 1 for articles that were not successfully replicated. The first dummy variable drops out of Equation 1 because it is collinear with the year fixed effects, and the second dummy variable drops out because it is collinear with the article fixed effects. Only the interaction is left as an identifiable parameter (Puhani, 2012).
[3] The transformation is approximately equal to 100γ if γ is small.
[4] There are several ways to estimate fixed-effects models. We use dummy variables for each fixed effect, a simple approach that produces approximately consistent estimates in both linear and negative binomial regression models (Allison & Waterman, 2002).
[5] The correlation between Failure$_i$ and the interaction was .94 before mean centering and less than .01 afterward.
[6] All *t* tests use the Welch-Satterthwaite approximate degrees of freedom for unequal variances.

References

Allison, P. D., & Waterman, R. P. (2002). Fixed–effects negative binomial regression models. *Sociological Methodology*, *32*(1), 247–265. https://doi.org/10.1111/1467-9531.00117

Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, *73*(2), 146–156. https://doi.org/10.1037/amp0000258

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion* (1st ed.). Princeton University Press.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Bolland, M. J., Grey, A., & Avenell, A. (2022). Citation of retracted publications: A challenging problem. *Accountability in Research*, *29*(1), 18–25. https://doi.org/10.1080/08989621.2021.1886933

Brooke-Hitching, E. (2018). *The phantom atlas: The greatest myths, lies and blunders on maps*. Chronicle Books.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*(10), 1363–1368. https://doi.org/10.1177/0956797610383437

Center for Open Science. (2021). *Registered reports*. https://www.cos.io/initiatives/registered-reports

Cesario, J., Jonas, K. J., & Carney, D. R. (2017). CRSP special issue on power poses: What was the point and what did we learn? *Comprehensive Results in Social Psychology*, *2*(1), 1–5. https://doi.org/10.1080/23743603.2017.1309876

Chawla, D. S. (2021, February 2). New bot flags scientific studies that cite retracted papers. *Nature Index*. https://

www.natureindex.com/news-blog/new-bot-flags-scientific-research-studies-that-cite-retracted-papers

Chivers, T. (2019). What's next for psychology's embattled field of social priming. *Nature*, *576*(7786), 200–202. https://doi.org/10.1038/d41586-019-03755-2

Colbert, S. (Writer). (2013, April 23). Austerity's spreadsheet error—Thomas Herndon (Season 9, Episode 90) [TV series episode]. In *The Colbert Report*. Comedy Central. https://www.cc.com/video/kbgnf0/the-colbert-report-austerity-s-spreadsheet-error-thomas-herndon

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, *27*(4), 605–615. https://doi.org/10.1017/pan.2019.25

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, USA*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Fleischmann, M., & Pons, S. (1989). Electrochemically induced nuclear fusion of deuterium. *Journal of Electroanalytical Chemistry*, *261*(2A), 301–308.

Google. (2021). *About Google Scholar*. https://scholar.google.com/intl/en/scholar/about.html

Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, *38*(2), 257–279. https://doi.org/10.1093/cje/bet075

Herring, C. (2009). Does diversity pay? Race, gender, and the business case for diversity. *American Sociological Review*, *74*(2), 208–224. https://doi.org/10.1177/000312240907400203

Hilsenrath, J. E. (2005, October 24). Novel way to assess school competition stirs academic row. *Wall Street Journal*. https://www.wsj.com/articles/SB113011672134577225

Hoxby, C. M. (2000). Does competition among public schools benefit students and taxpayers? *The American Economic Review*, *90*(5), 1209–1238.

Hoxby, C. M. (2007). Does competition among public schools benefit students and taxpayers? Reply. *The American Economic Review*, *97*(5), 2038–2055. https://doi.org/10.1257/aer.97.5.2038

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods*, *48*(4), 1308–1317. https://doi.org/10.3758/s13428-015-0624-x

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. https://doi.org/10.1001/jama.294.2.218

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

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. https://doi.org/10.1177/0956797611430953

Krugman, P. (2013, April 19). The excel depression. *The New York Times*. https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of Psychological Science. *Perspectives on Psychological Science*, *7*(6), 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Open Science Collaboration. (2016). *Reproducibility Project: Psychology*. OSF. https://osf.io/ezcuj/

Oransky, I. (2019, June 12). Want to check for retractions in your personal library—And get alerts—For free? Now you can. *Retraction Watch*. https://retractionwatch.com/2019/06/12/want-to-check-for-retractions-in-your-personal-library-and-get-alerts-for-free-now-you-can/

Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters*, *115*(1), 85–87. https://doi.org/10.1016/j.econlet.2011.11.025

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, *26*(5), 653–656. https://doi.org/10.1177/0956797614553946

R Core Team. (2021). *R: A language and environment for statistical computing* (Version 4.0.5) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Reinhart, C., & Rogoff, K. (2010). Growth in a time of debt. *American Economic Review: Papers & Proceedings*, *100*, 573–578.

Rothstein, J. (2007a). *Rejoinder to Hoxby*. https://eml.berkeley.edu/~jrothst/publications/rothstein-hoxbycomment-rejoinder.pdf

Rothstein, J. (2007b). Does competition among public schools benefit students and taxpayers? Comment. *American Economic Review*, *97*(5), 2026–2037. https://doi.org/10.1257/aer.97.5.2026

Rovira, C., Codina, L., Guerrero-Solé, F., & Lopezosa, C. (2019). Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic, WoS and Scopus. *Future Internet*, *11*(9), 202. https://doi.org/10.3390/fi11090202

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), Article eabd1705. https://doi.org/10.1126/sciadv.abd1705

Simmons, J. P., & Simonsohn, U. (2017). Power posing: *P*-curving the evidence. *Psychological Science*, *28*(5), 687–693. https://doi.org/10.1177/0956797616658563

Stojmenovska, D., Bol, T., & Leopold, T. (2017). Does diversity pay? A replication of Herring (2009). *American Sociological Review*, *82*(4), 857–867. https://doi.org/10.1177/0003122417714422

Taubes, G. (1993). *Bad science: The short life and weird times of cold fusion* (1st ed.). Random House.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(1), Article 62. https://doi.org/10.1038/s42003-018-0073-z

van Aert, R. C. M., & van Assen, M. A. L. M. (2018). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, *50*(4), 1515–1539. https://doi.org/10.3758/s13428-017-0967-6

Yu, G., Keirstead, J., Jefferis, G., Getzinger, G., Cimentada, J., Czapanskiy, M., & Makowski, D. (2021). *scholar: Analyse citation data from Google Scholar* (Version 0.2.0) [Computer software]. Retrieved from http://cran.r-project.org/package=scholar

Zotero. (2019). *Retracted item notifications with Retraction Watch integration*. https://www.zotero.org/blog/retracted-item-notifications/