Propagation of Error: Approving Citations to Problematic Research*

Ken Cor† Gaurav Sood‡

October 26, 2018

*We are grateful to Danielle Portnoy and Xiaoran Huang for assisting us with research, and to Andrew Gelman, Kabir Khanna, and Daniel Stone for offering valuable comments. The data and scripts for replicating the analysis can be found at: https://github.com/recite/propagation_of_error

†Ken is Assistant Dean, Assessment in the Faculty of Pharmacy and Pharmaceutical Sciences at the University of Alberta, Edmonton. Ken can be reached at mcor@ualberta.edu

‡Gaurav can be reached at gsood07@gmail.com

Abstract

Many claims in a scientific article rest on research done by others. But when the claims are based on flawed research, scientific articles potentially spread misinformation. To shed light on how often scientists base their claims on problematic research, we exploit data on cases where problems with research are broadly publicized. Using data from over 3,000 retracted articles and over 74,000 citations to these articles, we find that at least 31.2% of the citations to retracted articles happen a year after they have been retracted. And that 91.4% of the post-retraction citations are *approving*—note no concern with the cited article. We augment the analysis with data from an article published in *Nature Neuroscience* highlighting a serious statistical error in articles published in prominent journals. Data suggest that problematic research was approvingly cited *more frequently* after the problem was publicized. Our results have implications for the design of scholarship discovery systems and scientific practice more generally.

In April 2005, Rubio et al. (2005) published a disturbing finding in *Cancer Research*. They reported discovering that stem cells can spontaneously transform into cancerous cells during *in vitro* experiments. The finding was a blow to research in the use of stem cells to treat cancer. By 2010, according to *Web of Science*, the article had been cited over 300 times.

In August 2010, the article was retracted (De la Fuente et al. 2010). The authors had been unable to replicate the results. Worse, there was mounting evidence that transformations like the one reported were due to a basic error in such research: cross-contamination during cell culturing. But the episode can be seen in a favorable light. Errors were caught, and information about the error was disseminated in the same venue where the original article was published. The story, however, does not end there.

Since retraction, the article has been cited another 300 plus times, with most citations not noting any concern with the original work. For instance, a year after retraction, Firinci et al. (2011) published an article in *International Immunopharmacology* citing Rubio et al. as basis for warning scientists that stem cells can spontaneously transform. Two years after retraction, Kosaka et al. (2012) published an article in *Cancer Gene Therapy* in which they cited spontaneous transformation of human cells into cancerous cells, crediting Rubio et al. for the discovery, as a hurdle to implementation of treatment they find effective. Three years after retraction, Chang et al. (2013) published an article in *Aesthetic Plastic Surgery* citing Rubio et al. as evidence against studies that suggest that spontaneous transformation of human stem cells is not a risk. As we find, these are not isolated examples. It is a pattern.

In this paper, we study citations to research with serious errors. We tally how often research with serious errors is (approvingly) cited before and after the errors have been brought to light. To study the question, we exploit two datasets. The first dataset is a large original dataset of retracted articles—over 3,000 retracted articles and nearly 74,000 citations to retracted articles. The second dataset includes data from an article that highlights a potentially serious statistical error in articles published in prominent journals. Both sets of data suggest that articles with po-

tentially serious errors continue to be cited approvingly long after the error has been publicized, either via a retraction notice or publication of an article noting the error. For instance, we find that at least 31.2% of the citations to retracted articles happen a year after the publication of the retraction notice. And that 91.4% of the post-retraction citations are *approving*—note no concerns with the cited article. Our results have implications for the design of scholarship discovery systems, and for scientific practice more generally.

Potential Impact of Approvingly Citing Research With Serious Errors

Citations are the bedrock of the scientific process. Scientists use citations to give credit for being first ("x, y, and z have studied a"), to debate methods and inferences ("the method used in study x fails to account for s"), as evidence ("x shows a", "we use data from x for our meta-analysis"), and to contextualize results ("our results are consistent with results from y"). And unless the researcher notes problems with cited research, citations cue that the data, results, inferences, or in some cases, the entire article, can be trusted.

When researchers *approvingly* cite, i.e., note no concerns along with the citation, articles with serious errors, problems ensue. First, when erroneous research is approvingly cited as evidence (e.g., Chang et al. 2013; Torsvik et al. 2010), it cues that the evidence for the claim is good. Such citations thus unduly increase the reader's confidence in the result or argument. When the claim being buffeted by the citation is wrong, such citations also misinform. In the extremum, a reader may become persuaded that the incorrect point is right. And such a reader, generally another academic, may go on to write other articles influenced by the incorrect point, citing the erroneous article for support, or may share the point as fact with colleagues and students, propagating the error.

Second, when erroneous research is approvingly cited to contextualize results (e.g., Kosaka

et al. 2012), readers get a misleading impression of the plausibility of the numbers reported in the study.

Third, such citations give full credit to research (and researchers) when at best partial credit is deserved. And since citation tallies cue credibility, such citations make erroneous research appear yet more credible.[1]

Lastly, sometimes research is approvingly cited to acknowledge the source of the data. For instance, Lin, Zhang and Yang (2013) used data from "two retracted studies ... without acknowledgment of their retractions, both of which were for fraudulent data..." (p. 1 Paul, Haughom and Hansen 2015) in a meta-analysis. In such cases, the consequence is obvious and extreme—the key findings in the published work are incorrect.

In sum, *approving* citations to problematic research propagate the error.

Data, Research Design, and Expectations

We expect approving citations to problematic research to stop once the problems have been made public. But some research suggests otherwise. For instance, research by John Darsee continued to be approvingly cited after a highly publicized retraction of his work (Kochan and Budd 1992). Similarly, a study using a database of 235 retracted biomedical articles found that nearly 94% of the citations after retraction treated research as valid (Budd, Sievert and Schultz 1998). Yet another study, exploiting a dataset of 82 retracted articles, came to similar conclusions (Pfeifer and Snodgrass 1990).

All these studies, however, suffer from three weaknesses. First, the studies use small samples, often spanning a single discipline. Use of small, selective samples means that we still do not know how widespread the problem is. Second, the studies solely focus on citations after

[1]Even citations that note concerns with the retracted article will inflate the number of times a retracted article is cited. But when concerns are noted, better metrics can be built on top of the data. When no concern is noted, the task of building pro-rated metrics is considerably harder.

retraction. This means that we do not know how common approving citations are before the problems are publicized, and if the rate changes after publicity. And third, the studies only focus on retracted articles. Retraction is generally a result of serious scientific malpractice. By focusing on retractions alone, the studies fail to illuminate the much more common problem of approving citations to studies with major errors with potentially serious implications for the key results.

We address all three issues. We study approving citations to articles that make potentially serious errors that don't lead to a retraction by leveraging data from an article published in *Nature* that highlights a potentially serious statistical error in articles published in prominent journals. To more extensively study approving citations to retracted articles, we assemble a large original dataset of retracted articles—over 3,000 retracted articles and nearly 74,000 citations to the retracted articles.

Our first dataset comes from an article that identifies articles that mistake the difference between a statistically significant and statistically insignificant result as evidence that the difference is statistically significant (Nieuwenhuis, Forstmann and Wagenmakers 2011). (For an explanation of why this is problematic, see Gelman and Stern (2006).) Nieuwenhuis, Forstmann and Wagenmakers (2011) analyzed 170 articles published in *Nature*, *Science*, *Neuron*, and *Journal of Neuroscience* between 2009 and 2010. They found that roughly half of the 170 articles made this mistake. We used Web of Science (WoS) (we talk more about WoS below) (Reuters 2012) to download citations to all the 170 articles.

Our second dataset is a large novel dataset of retracted articles. We used WoS to assemble the data. WoS indexes articles from over 9,500 natural science journals and 3,500 social science journals (Yong-Hak 2013). WoS indexes articles from over 12,000 international journals and 148,000 conferences (Yong-Hak 2013). WoS contains key citation indices including the *Science Citation Index Expanded* (over 9,500 journals; 1900–present), *Social Sciences Citation Index* (over 3,500 journals; 1900–present), *Arts & Humanities Citation Index* (over 1,700 journals; 1975–present), *Conference Proceedings Citation Index* (over 170,000 conferences; 1990-present), *Book*

Citation Index (over 30,000 titles; 2005–present), among others. For a full list of titles included in the *Science Citation Index Expanded*, *Social Sciences Citation Index*, *Arts & Humanities Citation Index*, and *Conference Proceedings Citation Index*, and a synopsis of the *Book Citation Index*, see here.

To assemble the data, we searched WoS for retraction notices, used information in the retraction notice records to download retracted articles, and then downloaded citations to the retracted articles using WoS feature that provides citations associated with an article. (For details about the method and robustness checks around data, see SI SI 1.) Our final retraction dataset has 3,029 retracted articles and 73,564 citations to the retracted articles.

We augmented the Nieuwenhuis, Forstmann and Wagenmakers and retracted article datasets in two ways. Firstly, to understand why the articles were retracted, we coded the reasons given for retraction in a random set of 100 retraction notices. Second, to measure whether citations to retracted articles were approving or not, we took a random sample of 100 articles citing the retracted research and coded whether it acknowledged the error. (See SI 4 for details of the coding.)

Using the two datasets, we describe various features of citations to erroneous articles and assess how the frequency of citations changes when the errors are publicized. We expect publication of retraction notice or an article noting a potentially serious error in an article to increase awareness about specific problematic articles. We also expect publications noting a potentially serious kind of error to increase awareness about the error. For instance, we expect publication of Nieuwenhuis, Forstmann and Wagenmakers (2011) to increase awareness of the particular error in statistical reasoning. Either pathway should lead to a decline in approving citations to the article. Though for the reasons noted above, we expect the average decline to be modest.

Lastly, we expect the decline in citations due to greater publicity about a general error to be considerably more tepid than the decline in the rate of citations due to a retraction. For one, retractions are unequivocal indications of severe problems with the article. For two, retractions generate a greater response from the publishers, who often switch titles of the retracted articles

in their online databases to reflect that they have been retracted. For three, retractions are tied to specific articles. To find out articles affected by the point made by Nieuwenhuis, Forstmann and Wagenmakers (2011), scientists need to closely read the article they are citing.

To estimate the impact of the publication of error on citation rates, we track citation rate a few years before and after the information about the error is made public. Given long publication cycles, and assuming the article would have been accepted for publication before the discovery of the error, we test the impact of citations one, two, and three years out. For Nieuwenhuis, Forstmann and Wagenmakers (2011), we also use a Difference-in-Differences estimator, exploiting the fact that roughly half of the articles published in the same journals did not have the same potentially serious error.

Results

We start by describing the results from the Nieuwenhuis, Forstmann and Wagenmakers (2011) data and follow it with results from the much larger retracted article data.

Prima facie evidence suggests little impact of the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011) on citations to articles mistaking the difference between significant effect and insignificant effect as evidence for a significant difference. In the two years before the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011), and the year Nieuwenhuis, Forstmann and Wagenmakers (2011) was published (2011), the articles making the mistake were cited 2,267 times. Between 2012 and 2015, the articles were cited an additional 6,604 times.

Figure 1 offers a closer look. It plots the total number of citations received per year by each of the papers making the mistake, the average number of citations received per year by articles making the mistake and smoothed `loess` growth curves. The plot also makes one more thing clear—there is a skew in citation rates (skewness based on the method of moments = 2). To account for the skew, we switched means with medians. Doing so yields a pretty similar pattern except

for the expected intercept shift (see Figure SI 5.1). Not all articles making the error, however, have results similarly affected by the error. Nieuwenhuis, Forstmann and Wagenmakers (2011) flag articles where the error has potentially serious consequences for the results. Thus, next, we track what happens to citations to such articles. We track how the median number of citations vary across years, and whether or not they are affected by the publication of the Nieuwenhuis, Forstmann and Wagenmakers (2011). As Figure SI 5.2 shows, the median number of citations steadily and modestly increase over time with the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011) not.

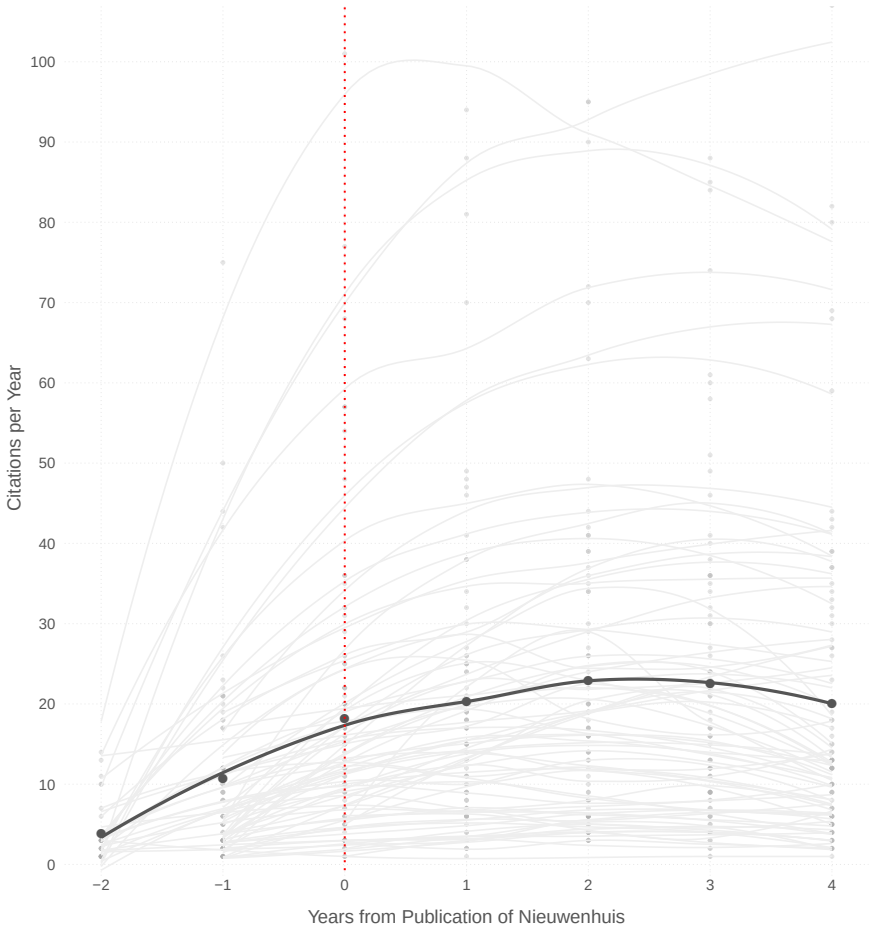

Figure 1: Number of Citations to Articles Containing the Error Per Year

But citations to erroneous research don't need to decline after the error in the research is publicized. We only expect approving citations to decline. To estimate how many citations

after the publication of Nieuwenhuis et al. are approving, we coded whether the citation was approving or not in 100 randomly chosen articles citing articles with the mistake (see SI SI 4 for further details about how the citations were coded.) Of the 100, only one article noted concerns with the cited article, citing (Nieuwenhuis, Forstmann and Wagenmakers 2011) for support. In all, there is strong evidence that approving citations are extremely common after the error is publicized.

To more formally explore the change in citation rate as a consequence of the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011), we started with an 'event study design' focusing on the citation rates to articles with the error. We regressed citations per year on a dummy for the year Nieuwenhuis, Forstmann and Wagenmakers (2011) was published, a linear time trend, and fixed-effect for the article. We also cluster by articles to account for multiple observations per article. In effect, we are getting an average of within article changes after regressing out a linear time trend. Results show, if anything, a modest uptick in citations after Nieuwenhuis, Forstmann and Wagenmakers (2011) is published—a year after publication of Nieuwenhuis, Forstmann and Wagenmakers (2011), articles containing the error get about four more citations per year compared to what they were getting before it (see Table SI 5.2).

Our main analysis for the Nieuwenhuis, Forstmann and Wagenmakers data is a Difference-in-Differences (DID) analysis. DID gives us a better way to control for over time trends, though as you will see, the results echo the results from the simpler analysis. We estimated whether the difference in citation rates of articles making the error and those not making the error changed after the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011). In particular, we regressed citations per year on whether or not the article makes the error, the year(s) after the publication of Nieuwenhuis, Forstmann and Wagenmakers (2011), and interaction between the two. We also include fixed-effects for each article to do within article estimation. Including fixed-effects allows us to circumvent concerns around skew in citations. And we again clustered the standard errors by article.

Table 1 tabulates the results. Models (1), (3), and (5) define error as all article making the error. And Models (2), (4), and (6) refer to error as articles making "potentially serious errors." As the table shows, 1 or 2 years after Nieuwenhuis, Forstmann and Wagenmakers (2011), articles making the error were being cited more frequently vis-à-vis articles not making the error (Diff. $\sim$ 3). Three years out, we cannot still reject the 0, suggesting that there is no evidence of a decline. For articles making "potentially serious errors", the story is much the same, except that the 1 and 2-year out estimates are closer to 3.5 additional citations per year than 3. And three years later, we still cannot say that the articles making "potentially serious errors" were being cited any less frequently. In all, there is strong evidence that approving citations are extremely common after the error is publicized.

An article published in a prominent journal flagging potentially serious concerns about statistical analysis in a set of articles is one thing, a retraction notice is quite another. A majority of the articles that are retracted are retracted, as we show below, because of serious error or fraud. We expect a much stronger response to the publication of retraction notices. We expect retracted articles to *never* be approvingly cited a year or more—taking account of long publication windows—after the retraction notice has been published. As we show below, the data suggests otherwise.

Over the last thirty or so years, the number of retractions has increased sharply (see Figure 2). The first retraction notice that we have in our database is from 1989. That year and the decade after it, the number of retraction notices being published per year never crossed 20. Since then, there has been a sharp and accelerating rise in the number of retraction notices per year. Between 2001, when 16 retraction notices were published, and 2015, last year for which we have complete data, there was a near 30 fold increase; a total of 451 retraction notices were published in 2015. The pattern that we find is consistent with results from Steen, Casadevall and Fang (2013), who also find a rapid increase in retractions over time.

Table 1: Difference-in-Difference Analysis of the Impact of Publication of Nieuwenhuis on the Number of Times per Year Articles Containing the Error Are Cited Vis-a-Vis Articles that Didn't Contain the Error

	Dependent variable:					
	Citations per year					
	1 year out		2 years out		3 years out	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Date	7.2***	7.7***	5.1***	5.5***	3.7***	3.9***
	(0.9)	(0.7)	(0.9)	(0.7)	(1.0)	(0.8)
Error or Not	1.5	0.004	2.1	0.6	2.8	1.6
	(2.5)	(2.8)	(2.4)	(2.7)	(2.4)	(2.6)
Makes Error*Treatment Date	3.1**	3.7***	2.7**	3.5***	1.7	2.1
	(1.2)	(1.3)	(1.2)	(1.3)	(1.3)	(1.5)
Constant	9.5***	10.2***	11.7***	12.6***	13.1***	14.0***
	(1.8)	(1.5)	(1.7)	(1.5)	(1.7)	(1.4)
Observations	957	957	957	957	957	957
Akaike Inf. Crit.	7,328.2	7,327.8	7,408.8	7,407.5	7,474.9	7,475.2
Bayesian Inf. Crit.	7,357.4	7,357.0	7,437.9	7,436.7	7,504.0	7,504.4

Note: *p<0.1; **p<0.05; ***p<0.01
Models (1), (3), and (5) define error as any article making the error.
And Models (2), (4), and (6) refer to error as articles making "potentially serious errors."

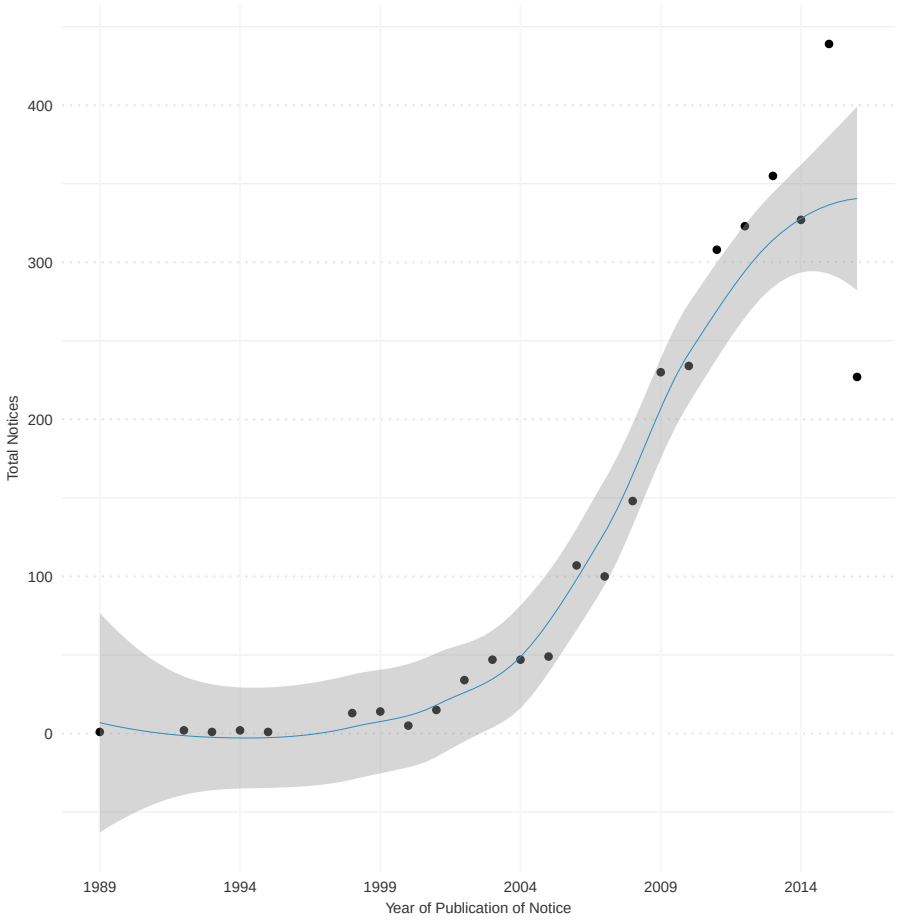

Figure 2: Retraction Notices Per Year

The rapid rise in the number of retractions is likely a combination of increasing production and improvements in detection. But the bottom line is that there is an ever faster growing number of articles which the scientific community thinks should not have have been published in the first place.

These retractions are spread unevenly across disciplines. About 65.2% of the retracted articles are from Life Sciences and Biomedicine, and 13% are from Physical Sciences. (For a breakdown by field, see Table SI 3.1.) We cannot be sure about the source of the variation. It could be that the variation is entirely explained by how many articles are published in each field.

To understand why the articles are retracted, we coded a random sample of 100 retraction notices. 39% of the notices mentioned plagiarism as one of the major reasons for retracting

the article. (Plagiarism includes self-plagiarism, duplication of data, words, and publishing the same or similar article in multiple journals.) Major errors or fraud contributed to another 51% of the retractions, with fraud alone contributing to 24% of the retractions. Ethics violations (2) and conflict over authorship or approval from other authors (5) contributed to the rest. The percentage of retractions attributable to major errors or fraud in our data is similar to the number obtained by other research on reasons for retraction in other corpora. For instance, a study of 1,112 Biomedicine articles retracted between 1997 and 2009 found that 55% were retracted for some type of misconduct (Budd, Coble and Anderson 2011) (see also Steen (2010)). All in all, articles are mostly retracted because the research cannot be trusted.

These flawed articles often accrue a fair number of citations before being retracted. In our data, the articles had been cited 39,792 times before being retracted. This is not unsurprising, given that it took, on average, 2.85 years for the article to be retracted. The median time before the article was retracted was two years (see Figure 3) with 28.1% of the articles taking 4 or more years to be retracted. These numbers are similar to those obtained elsewhere. A study on time to retraction in the PubMed corpus found that the average time to retraction was approximately three years (Steen, Casadevall and Fang 2013).

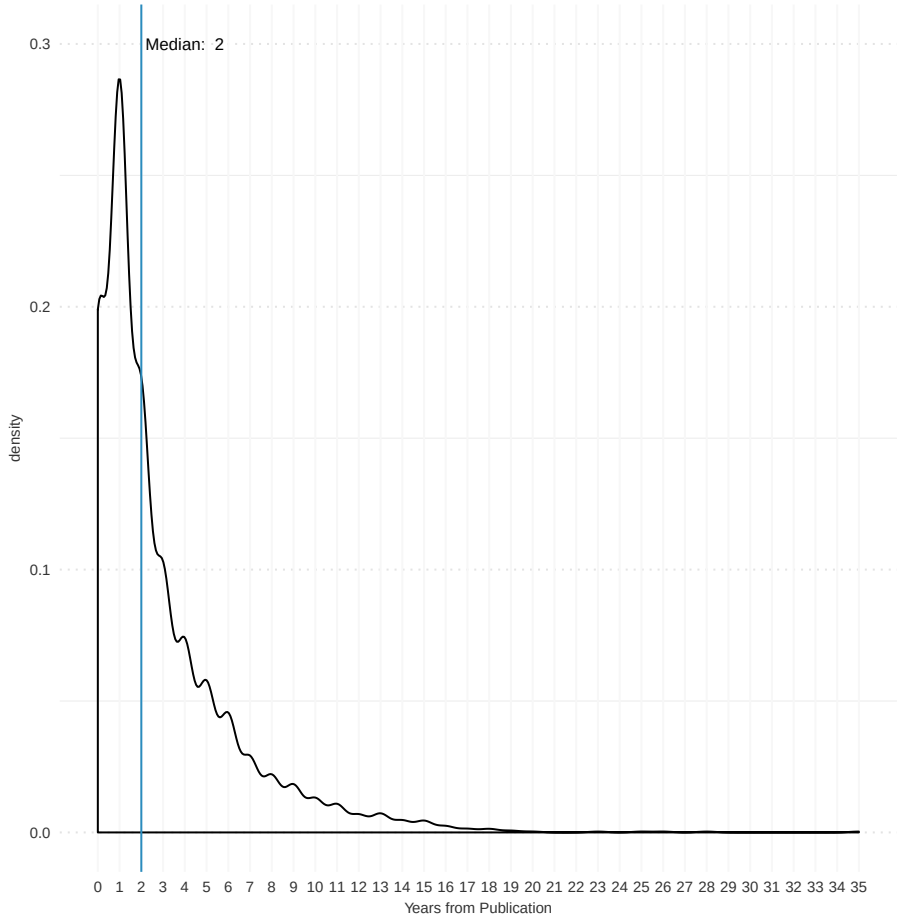

Figure 3: Time to Retraction

On the hunch that greater readership of more prominent journals would mean that problematic articles are flagged more quickly, we estimated the relationship between journal impact factor and average time to retraction. As Figure 4 shows, the relationship is flat—flawed articles in low ranked journals are retracted as quickly as flawed articles in higher ranked journals.

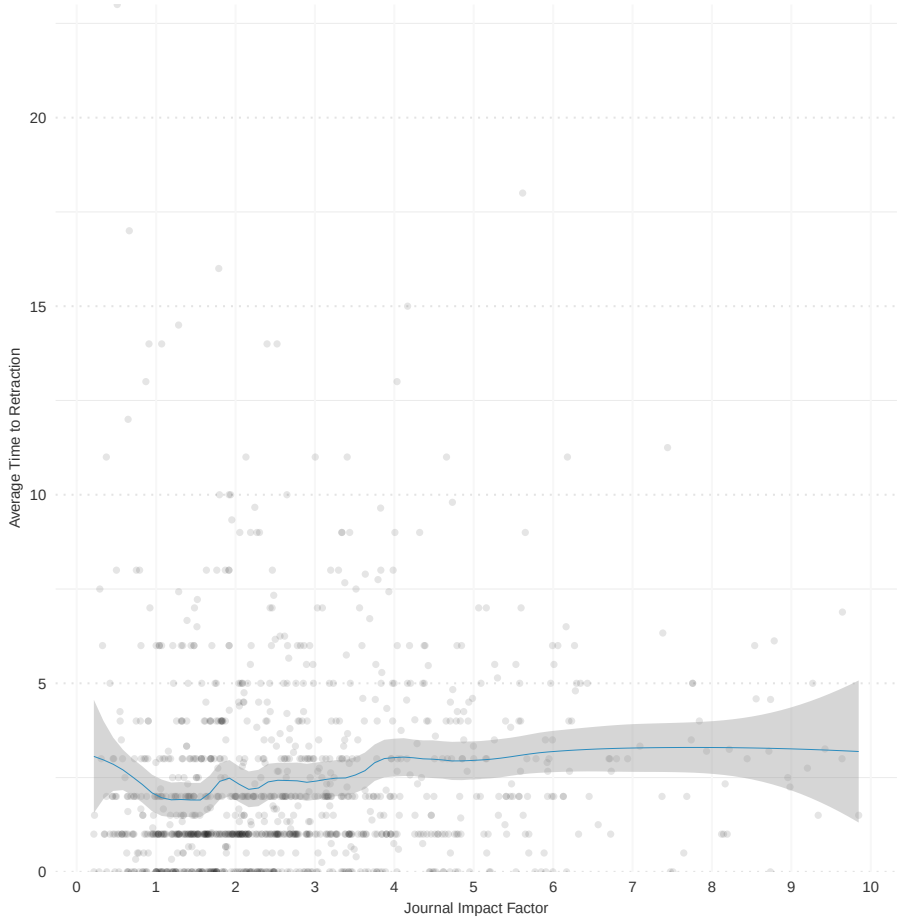

Figure 4: Relationship Between Journal Impact Factor and Time to Retraction

Given that a majority of retracted articles are retracted because of serious error or fraud, we expect retracted articles to *never* be approvingly cited a year or more—taking account of long publication windows—after the retraction notice has been published. However, retracted articles were cited another 22,932 times between the year after they were retracted and August 2016. Thus, on average, the retracted articles received an additional 7.57 citations. Given the skew in retraction notices, with the bulk being published in recent years, these totals include very little post retraction data for many of the articles. In other words, the results are a *lower bound* of the percentage of citations that happen after an article has been retracted.

To explore the frequency of citations before and after retraction, we plotted line graphs of total citations per article per year against year from the publication of retraction notice (see

Figure 5). And we overlaid the lines with the median number of citations per article per year. We limit ourselves to 10 years before and after the publication of retraction notice as the data are very sparse beyond that. Total citations decline when the retraction notice is published—the median goes from 3 to 2 between the year retraction notice is published after next year. But the decline is followed by a plateauing.

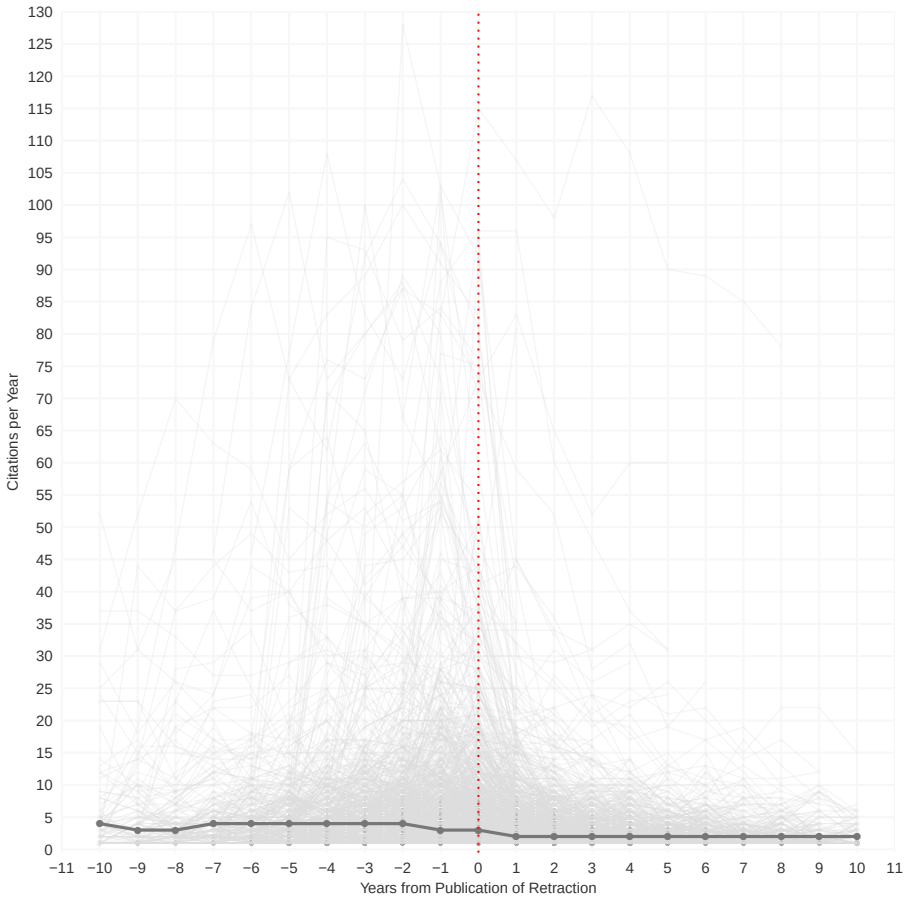

Figure 5: Number of Citations to Retracted Articles per Year

Figure 5 elides over the fact that we do not observe data on all articles after all the plotted years after the publication notice. For instance, if the retraction notice was published in 2014, we only observe one more full year of citations—2015. Thus, to look more formally at the impact of publication of retraction notices a year, 2 years, and 3 years after, we create subsets of data where we have all the articles for which we have at least 1 year, 2 years, and 3 years worth of data after

the publication of the retraction notice. To analyze these subsets, we use a pretty simple model. We regress the number of citations per year per article on years from retraction notice, a dummy for the cliff (1-, 2-, and 3- years after the publication of the retraction notice) and an interaction between the two. We cluster the standard errors by article to account for the fact that we have multiple observations per article. The results of the model can be seen in Table 2.

Table 2: Impact of Publication of Retraction Notice on the Number of Times Retracted Articles Are Cited per Year

	Dependent variable:		
	Citations Per Year		
	1 Year Later	2 Years Later	3 Years Later
	(1)	(2)	(3)
(1, 2, 3) Years After Notice	-2.4^{***}	-2.1^{***}	-1.9^{***}
	(0.2)	(0.2)	(0.3)
Years to Notice	-0.01	-0.2^{***}	-0.3^{***}
	(0.03)	(0.03)	(0.03)
(1, 2, 3) Years After Notice*Years to Notice	-0.4^{***}	-0.2^{***}	0.002
	(0.04)	(0.05)	(0.1)
Constant	6.0^{***}	5.2^{***}	4.7^{***}
	(0.2)	(0.1)	(0.2)
Observations	12,511	11,486	10,428
Akaike Inf. Crit.	83,835.6	76,511.3	69,534.9
Bayesian Inf. Crit.	83,880.2	76,555.4	69,578.4
Note:			*p<0.1; **p<0.05; ***p<0.01

As Table 2 shows, on average an article is cited about 5–5.5 times per year. But 1, 2, and three years later, an average article accrues about two fewer citations per year, a drop that is statistically and substantively significant. There is also a small negative slope in the number of citations per year for models that estimate the effect 1 and two years out. So the number of citations is slowly decreasing. But note that our priors are post-retraction, articles would not be cited. So we must compare the citation rate to 0. And there we can comfortably reject the 0—citation rate post publication of the retraction notice is not zero.

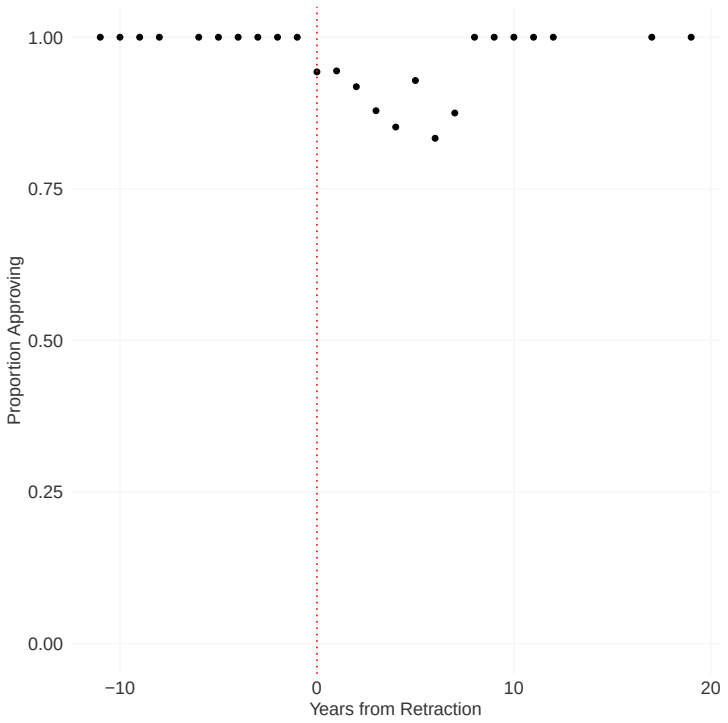

Figure 6: Proportion of Citations That Are Approving Per Year

But we are being imprecise. Our claim is about approving citations—after the publication of the retraction notice, the number of approving citations should go to 0. To estimate how many of the citations are *approving*, we coded a random sample of 100 articles that cited a retracted article pre-retraction and 275 articles that cited retracted research a year or more after the publication of the retraction notice. We could not locate 32 articles, leaving us with 343 articles. There were no false positives. Of the 87 articles citing retracted articles before or in the same year the retraction notice was published, 85 (97.7%) were approving. And of the 256 articles citing retracted articles the year after the retraction notice was published, 234 (91.4%) were approving. (Figure 6 plots the proportion of approving citations by year.)

In all, the data suggest that retracted articles continue to be approvingly cited long after the publication of retraction notice.

Discussion

Only some of the points that scientists make in a paper rely on original work. Many, if not most of the points made in a typical paper, rely on work done by others. That is the nature of the scientific enterprise. But sometimes the work done by others is problematic. For instance, sometimes the claims made by the work being cited are not only wrong but made with fraudulent data. If at all scientists refer to such work, we expect scientists to acknowledge the issues carefully, especially when the errors have been publicized. We expect approving citations to problematic research to stop once the problems have been made public. But data suggest that publicizing serious errors via public retractions or publication of research highlighting the problem at prominent venues leads to, at best, a modest decline in citations. Retracted articles continue to be cited approvingly years after they have been publicly retracted.

But why do scientists fall short? We suspect that they fall short because they place too much faith in others and because they are resource constrained. Scientists sometimes do not scrutinize the article they cite because they trust the published work or trust the integrity of references in old databases or are under intense pressure to publish.

Perhaps the single most important reason why researchers approvingly cite research with serious errors, even after the errors have been publicized, is that scientists trust other scientists, especially peer-reviewed work produced by other scientists. One likely reason for that trust is the belief that scientific misconduct is limited to a few bad people. And that optimism is likely driven by the fact that a few cases of fraud get a bulk of the attention, with reporting often focusing on personalities than processes. Cases of Diederik Stapel, who fabricated data behind at least 30 papers (Levelt, Drenth and Noort 2012), John Darsee, who faked data behind nearly 100 publications (Stewart and Feder 1987; Anderson et al. 2013; Wallis 1983), and Jan Hendrik Schön, who during a period in 2001 published a research paper every 8 days based on fabricated data (Service 2003; Anderson et al. 2013), are legend. So are cases of Andrew Wakefield, who

published an article linking MMR vaccine to autism using fabricated data (Wakefield et al. 1998; Deer 2011; Godlee, Smith and Marcovitch 2011), and recently Michael Lacour, who published a paper in *Science* based on fabricated data (Broockman, Kalla and Aronow 2015; McNutt 2015).[2] Each of these cases was framed as an example of misconduct by a bad actor, the subtext often being that a bad actor is an exception, not the rule.

Misconduct, however, is not limited to a few bad actors. A large anonymous survey of early- and mid-career scientists found that about 2% of scientists admitted to engaging in fabricating, falsifying, or plagiarizing in the last *three* years Martinson, Anderson and De Vries (2005) (see also Titus, Wells and Rhoades (2008)). Another study found that nearly 34% of the respondents in past surveys had admitted to engaging in questionable research practices Fanelli (2009).

The other likely reason behind trust in peer-reviewed research is the fact that the rate of retractions is extremely low. For instance, of the nearly 9.4 million articles published between 1950 and 2004 and available on PubMed, only 596 have been retracted (Cokol et al. 2007). In all likelihood, however, the actual rate of serious errors in manuscripts is manifolds that rate. For instance, Cokol et al. (2007) estimate the rate at which articles ought to be retracted to be anywhere between 16.7 times to 167.8 times the actual rate. And these estimates do not account for research that involves harder-to-prove malpractice such as stuffing non-significant results in the file-drawer (Franco, Malhotra and Simonovits 2014), conducting specification searches, and other more fundamental concerns like low power, which reduces the likelihood that a nominally statistically significant finding reflects a true effect (Button et al. 2013; Ioannidis 2005). All in all, while the belief that most research that is produced is reliable is very likely unfounded, it probably explains why scientists approvingly cite erroneous research.

According to us, the second biggest reason why scientists cite erroneous research is lack of time. Given the pressure to publish, many researchers likely do not spend enough time vetting

[2]Other prominent cases include that of William Summerlin, who painted mice rather than transplant skin (Basu 2006; Anderson et al. 2013), Woo Suk Hwang, who claimed to have cloned embryos, Eric Poehlman, who fabricated data behind at least ten papers and numerous grant applications.

the research they cite. Pressed for time, scientists often likely default to credulousness when evaluating the research they cite. There is also likely some 'motivated vetting,' with articles cited in 'support' likely receiving less scrutiny than those making the 'opposing' argument.

Thirdly, incentives to cite carefully are mostly absent. More often than not, the only thing researchers are ever knocked on when it comes to citations is failing to cite someone or missing the journal's formatting requirements. Citing incorrectly or citing bad research flatteringly generally attracts little opprobrium.

Fourthly, when researchers are searching for relevant research, there are no tools that reliably alert researchers about errors in research. At the time of writing, Google Scholar, for instance, does not flag if an article has been retracted, much less flag articles that have found serious problems with the article.

Lastly, often, researchers rely on old reference databases sitting on their computer for citations. And researchers likely don't check if these databases contain articles that have since been retracted because of the reasons we discuss above—chances are low. For instance, Davis (2012) finds that personal Mendeley libraries contained 1,340 retracted articles. All in all, there are a lot of reasons to suspect that scientists would cite erroneous research, even when errors have been publicized via the publication of a retraction notice or an article noting the problem.

These 'mistakes'—citing flawed research when flaws have been made public—are avoidable. Assuming that researchers do not knowingly approvingly cite retracted articles, the data imply that the discovery of errors, even when public retraction notices are issued, is still a problem.

To ameliorate the problem, we need to improve access to information about problems in research. One way to improve access to information about problems is to build tools that provide the information as part of existing research discovery and production processes. For instance, altering interfaces of heavily used portals such as Google Scholar, JSTOR, journal publishers' sites, etc. so that they thread reproduction attempts, retractions, and other research that directly bears

on the evidence presented in an article along with the article are liable to be effective. Rather than effect change in multiple interfaces, which requires coordination with multiple strategic actors, however, a better strategy may be to create a browser plug-in that highlights problematic articles listed on a web page. Providing such a tool to editors or copy editors at academic publishers may also help ameliorate the problem. Flagging problems during the scientific discovery process, however, is better than flagging them during the production process. Flagging during discovery likely preempts the temptation to engage in post hoc rationalization. Alternately, one could build tools that automatically create pull requests to personal bibliography libraries posted on open publication platforms like GitHub. Lastly, while our study only tallies research that cites known flawed research, it is quite likely that the effect of flawed research extends to studies that cite studies that approvingly cite flawed research (and thereon). And any modifications to the interface should extend to papers that cite flawed research so that people citing them, in turn, are also warned.

Egregious errors like approving citations to flawed research after the flaws have been made public serve to highlight more significant problems with how science is practiced. Scholars do not appear to carefully vet research they cite. To improve the reliability of scientific production, besides innovating on better tools, we may also need also to penalize research that makes such errors.

References

Anderson, Melissa S, Marta A Shaw, Nicholas H Steneck, Erin Konkle and Takehito Kamata. 2013. Research integrity and misconduct in the academic profession. In *Higher education: handbook of theory and research*. Springer pp. 217–261.

Basu, Paroma. 2006. "Where are they now?" *Nature medicine* 12(5):492–493.

Broockman, David, Joshua Kalla and Peter Aronow. 2015. "Irregularities in LaCour (2014)." *Work. pap., Stanford Univ. http://stanford. edu/ dbroock/broockman_kalla_aronow_lg_irregularities. pdf* .

Budd, John M, MaryEllen Sievert and Tom R Schultz. 1998. "Phenomena of retraction: reasons for retraction and citations to the publications." *JAMA* 280(3):296–297.

Budd, John M, Zach C Coble and Katherine M Anderson. 2011. Retracted publications in biomedicine: Cause for concern. In *Association of College and Research Libraries Conference*. pp. 390–5.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson and Marcus R Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14(5):365–376.

Chang, Hak, Byung-Rok Do, Jeong-Hwan Che, Byeong-Cheol Kang, Ji-Hyang Kim, Euna Kwon, Ji-Young Kim and Kyung-Hee Min. 2013. "Safety of adipose-derived stem cells and collagenase in fat tissue preparation." *Aesthetic plastic surgery* 37(4):802–808.

Cokol, Murat, Ivan Iossifov, Raul Rodriguez-Esteban and Andrey Rzhetsky. 2007. "How many scientific papers should be retracted?" *EMBO reports* 8(5):422–423.

Davis, Philip M. 2012. "The persistence of error: a study of retracted articles on the Internet and in personal libraries." *Journal of the Medical Library Association* 100(3):184.

De la Fuente, Ricardo, Antonio Bernad, Javier Garcia-Castro, Maria C Martin and Juan C Cigudosa. 2010. "Retraction: Spontaneous human adult stem cell transformation." *Cancer Res* 70(16):6682.

Deer, Brian. 2011. "How the case against the MMR vaccine was fixed." *BMJ* 342:c5347.

Fanelli, Daniele. 2009. "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data." *PloS one* 4(5):e5738.

Firinci, Fatih, Meral Karaman, Yusuf Baran, Alper Bagriyanik, Zeynep Arikan Ayyildiz, Muge Kiray, Ilknur Kozanoglu, Osman Yilmaz, Nevin Uzuner and Ozkan Karaman. 2011. "Mesenchymal stem cells ameliorate the histopathological changes in a murine model of chronic asthma." *International immunopharmacology* 11(8):1120–1126.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Gelman, Andrew and Hal Stern. 2006. "The difference between "significant" and "not significant" is not itself statistically significant." *The American Statistician* 60(4):328–331.

Godlee, Fiona, Jane Smith and Harvey Marcovitch. 2011. "Wakefield's article linking MMR vaccine and autism was fraudulent." *BMJ* 342:c7452.

Ioannidis, John PA. 2005. "Why most published research findings are false." *PLoS Med* 2(8):e124.

Kochan, Carol Ann and John M Budd. 1992. "The persistence of fraud in the literature: the Darsee case." *Journal of the American Society for Information Science* 43(7):488.

Kosaka, Hiroshi, T Ichikawa, K Kurozumi, H Kambara, S Inoue, T Maruo, K Nakamura, H Hamada and I Date. 2012. "Therapeutic effect of suicide gene-transferred mesenchymal stem cells in a rat model of glioma." *Cancer gene therapy* 19(8):572.

Levelt, Willem JM, PJD Drenth and E Noort. 2012. "Flawed science: The fraudulent research practices of social psychologist Diederik Stapel.".

Lin, Jun, Lei Zhang and Huilin Yang. 2013. "Perioperative administration of selective cyclooxygenase-2 inhibitors for postoperative pain management in patients after total knee arthroplasty." *The Journal of arthroplasty* 28(2):207–213.

Martinson, Brian C, Melissa S Anderson and Raymond De Vries. 2005. "Scientists behaving badly." *Nature* 435(7043):737–738.

McNutt, Marcia. 2015. "Editorial retraction." *Science* p. aaa6638.

Nieuwenhuis, Sander, Birte U Forstmann and Eric-Jan Wagenmakers. 2011. "Erroneous analyses of interactions in neuroscience: a problem of significance." *Nature neuroscience* 14(9):1105–1107.

Paul, H Yi, Bryan D Haughom and Erik N Hansen. 2015. "Comment on" Perioperative Administration of Selective Cyclooxygenase-2 Inhibitors for Postoperative Pain Management in Patients After Total Knee Arthroplasty"." *The Journal of arthroplasty* 30(4):718.

Pfeifer, Mark P and Gwendolyn L Snodgrass. 1990. "The continued use of retracted, invalid scientific literature." *JAMA* 263(10):1420–1423.

Reuters, Thomson. 2012. "Web of Science.".

Rubio, Daniel, Javier Garcia-Castro, María C Martín, Ricardo de la Fuente, Juan C Cigudosa, Alison C Lloyd and Antonio Bernad. 2005. "Spontaneous human adult stem cell transformation." *Cancer research* 65(8):3035–3039.

Service, Robert F. 2003. "Scientific misconduct. More of Bell Labs physicist's papers retracted." *Science (New York, NY)* 299(5603):31.

Steen, R Grant. 2010. "Retractions in the scientific literature: is the incidence of research fraud increasing?" *Journal of medical ethics* pp. jme–2010.

Steen, R Grant, Arturo Casadevall and Ferric C Fang. 2013. "Why has the number of scientific retractions increased?" *PLoS One* 8(7):e68397.

Stewart, Walter W and Ned Feder. 1987. "The integrity of the scientific literature." *Nature* 325:207–14.

Titus, Sandra L, James A Wells and Lawrence J Rhoades. 2008. "Repairing research integrity." *Nature* 453(7198):980–982.

Torsvik, Anja, Gro V Røsland, Agnete Svendsen, Anders Molven, Heike Immervoll, Emmet McCormack, Per Eystein Lønning, Monika Primon, Ewa Sobala, Joerg-Christian Tonn et al. 2010. "Spontaneous malignant transformation of human mesenchymal stem cells reflects cross-contamination: putting the research field on track–letter." *Cancer research* 70(15):6393–6396.

Wakefield, Andrew J, Simon H Murch, Andrew Anthony, John Linnell, DM Casson, Mohsin Malik, Mark Berelowitz, Amar P Dhillon, Michael A Thomson, Peter Harvey et al. 1998. "RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children." *The Lancet* 351(9103):637–641.

Wallis, Claudia. 1983. "Fraud in a Harvard lab." *Time* 121(9):49–49.

Yong-Hak, Jo. 2013. "Web of Science." *Thomson Reuters* .

Supporting Information

SI 1 Creating the Retractions Dataset

To build a database of retracted articles, we started by creating a list of retraction notices. To do that, in August 2016, we searched WoS for titles containing the phrase "retraction of." The search yielded more than 14,000 records. Using the "corrections" filter in WoS—it is a WoS flag for retraction and correction notices—we filtered the list to 4,085 retraction notices.

Next, we wrote software to automatically search the WoS database for retracted articles using the information in the retraction notice records. Retraction notice records did not contain consistent titles to allow us to a simple search. But 99% of the retraction notices contained the year the original article was published, and 96% listed the authors of the original work. We used these two pieces of information along with the name of the publication to search the WoS for the original articles. The search resulted in a list of 3,776 articles. We could not locate the remaining 309 retracted articles.

Due to the variability in the information contained in the retraction notice records, the automated search process returned the wrong article in some cases. Our aim was to have zero false positives even at the risk of some false negatives. With that aim, we created rules to flag potential false positives. First, if the list of authors of the retracted article did not match the list of authors for the relevant retraction notice record, we flagged the record as a potential false positive. Second, if the title of the retracted article did not contain the words "retracted" or "retraction," we flagged it as a potential false positive. (It is standard practice for titles of original articles to be revised to indicate that the article has been retracted.) Third, we parsed the title of the retracted notices to extract the title of the original retracted article. And we flagged articles where the titles did not match as potential false positives. We then reviewed the potential false positives, filtering out all records where we could not verify the match. This resulted in a set of 3,084 articles. Finally,

we checked for duplicates. We found 55. This left us with 3,029 articles. And that served as our final sample.

As an additional robustness check, we manually checked a random sample of 100 retracted articles to confirm that the article had indeed been retracted. We found that all of them were.

To get a list of citations to these articles, we used the WoS functionality that allows users to access the list of citations to articles. We wrote software to automatically download citation records for each of the retracted articles. In total, we found 73,564 citations.

SI 2 Classifying Citations as Post-Retraction or Not

A few retraction notices, retracted articles, and articles citing the retracted articles have earlier online (or conference) publication dates than the print publication dates recorded by WoS. As a result, post-retraction citations can be classified as otherwise. Or vice versa. To determine the impact of this issue and issues like these on our estimate of the lower bound of the proportion of citations that are made a year or more after the publication of the retraction notice, we manually recorded the online publication dates for a random sample of 300 citations to retracted articles, the associated retraction notices, and retracted articles. We could not retrieve 20 articles citing a retracted article. Of the remaining 280 records, switching to online publication dates suggests that three articles were misclassified as post-retraction (2.2%) and four were misclassified as not post-retraction (3.2%). Taking these error rates at their face value, we re-calibrated our results. The recalibration results in an increase in the number of post-retraction citations, from 22,932 to 23,289. Or, the lower bound of the proportion of citations that happen the year after the retraction notice is published goes from 31.2% to 31.7%.

SI 3 Retracted Articles by Field

To understand the kinds of fields where retractions are more common, we used an augmented Web of Science research field categorization scheme to classify the articles. There is one caveat. Sometimes papers cover more than one topic. We just choose the first topic in these cases taking it to be the primary topic.

As Table SI 3.1 shows, 65% of the retracted articles were published in the Life Sciences and Biomedicine field. A distant second at 13% is Physical Sciences, followed by Technology at 10.7%. Social Sciences are at 5.5%. One reason why a large majority of the retractions are from the Life Sciences and Biomedicine field may be simply because the field has more publications. But we cannot say anything definitely.

Table SI 3.1: Retraction Notices By Field

Field	Number of Notices	Percentage of Total
Arts & Humanities	13	0.4
Life Sciences & Biomedicine	1974	65.2
Multidisciplinary	157	5.2
Physical Sciences	393	13.0
Social Sciences	165	5.5
Technology	325	10.7

SI 4 Coding Citations as Approving or Not

To code the citations, we downloaded citing articles and their associated retracted article. A research assistant then edited the citing article pdf to highlight where the retracted article was discussed in the citing article. The judgment of whether the article noted any concerns was made based on a review of the original retracted article pdf and the highlighted text.

If an article did not note any concerns with the cited article, it was coded as *approving*. Simply disagreeing with the conclusions of an article without noting any concern still meant that the article was being cited in a way that suggests that its findings are trustworthy and were also coded as *approving*. We code articles that note any concern with the citing article, even those unrelated to the cause of retraction, as *disapproving*.

In the Nieuwenhuis data, we could not locate one of the 100 articles, leaving us with 99 articles. Of the 99 articles, 2 were false positives—the articles did not cite erroneous research, but instead cited a paper with authors and title similar to published erroneous research. Of the 97 remaining articles, only one article noted concerns while citing an article making a mistake, citing Nieuwenhuis, Forstmann and Wagenmakers (2011) for support.

In the retracted article data, we could not locate 32 articles, leaving us with 343 articles. There were no false positives. Of the 87 articles citing retracted articles before or in the same year the retraction notice was published, 85 (97.7%) were approving. And of the 256 articles citing retracted articles the year after the retraction notice was published, 234 (91.4%) were approving.

We evaluated the reliability of the coding by having an independent rater code 50 randomly selected citing articles. The two sets of independent codes were found to agree in all 50 instances.

SI 5 Rate of Citations Before and After Publication of Nieuwenhuis et al.

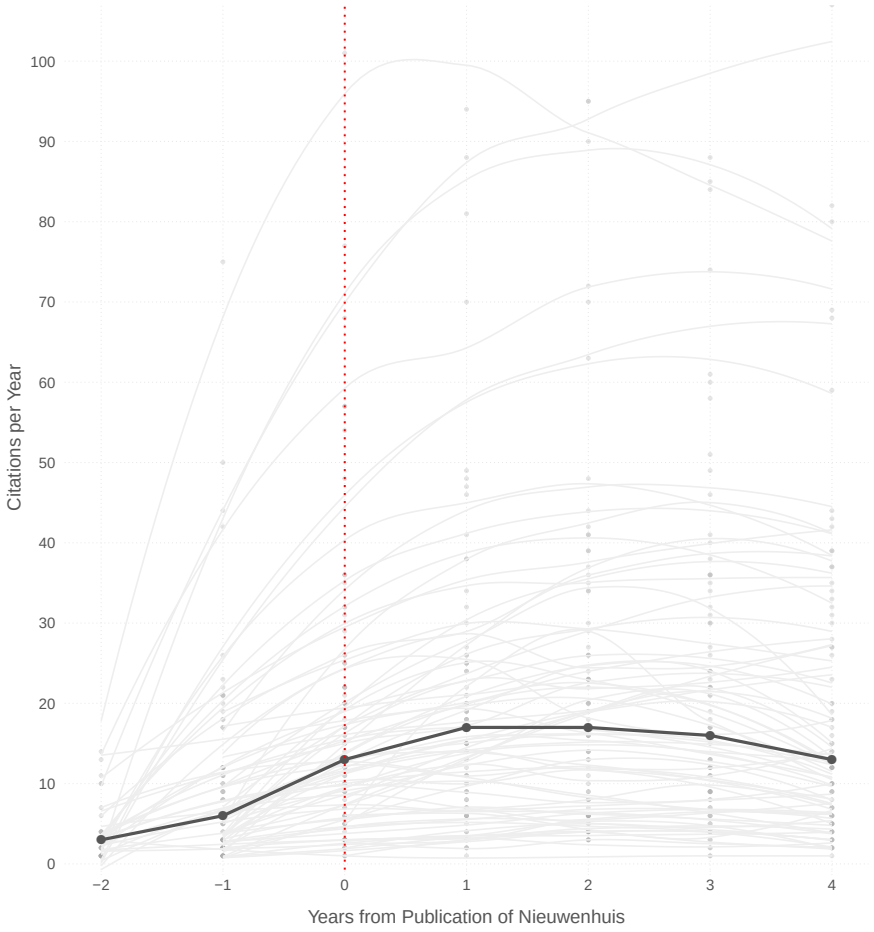

Figure SI 5.1: Total number of citations received per year by each of the papers making the mistake, and the median number of citations received per year by the articles.

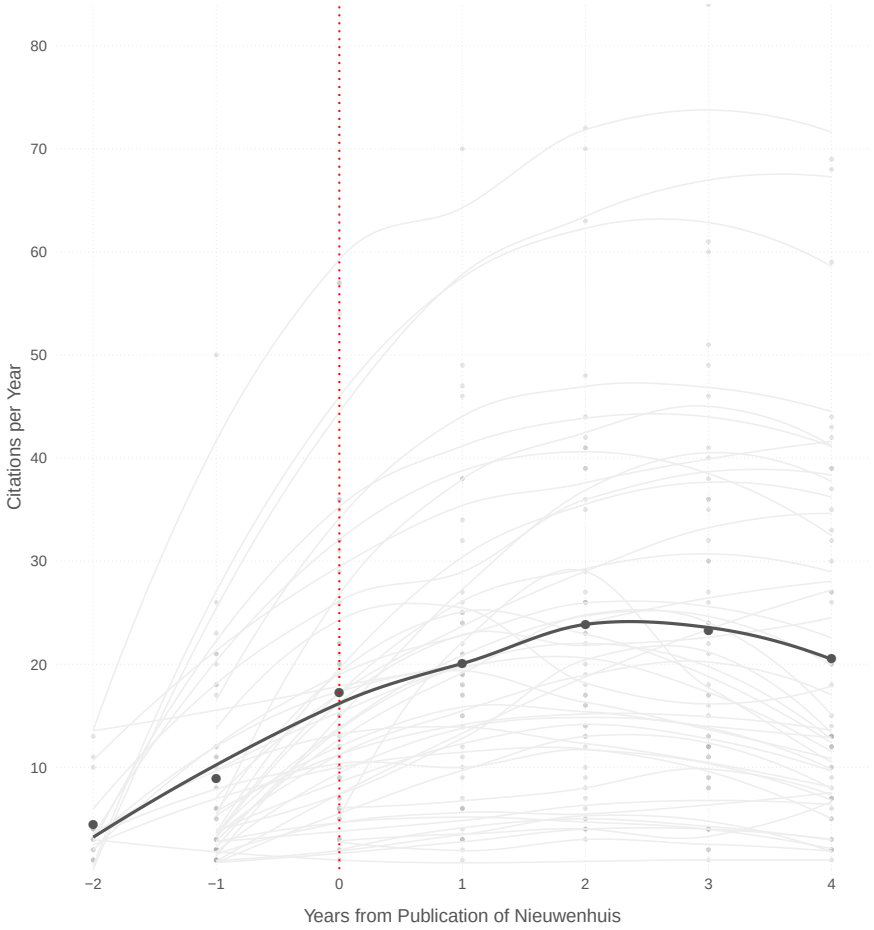

Figure SI 5.2: Total number of citations received per year by articles making the mistake with 'potentially serious' consequences for the results, and the average number of citations received per year by the articles.

Table SI 5.2: Change in the Number of Citations to Articles Containing the Error Per Year Before and After Publication of Nieuwenhuis

	Dependent variable:	
	Citations Per Year	
	All Articles with Mistakes	Articles with Potentially Serious Errors
	(1)	(2)
Transition Date	3.8**	5.0**
	(1.7)	(2.0)
Time	2.0***	2.1***
	(0.4)	(0.5)
Constant	12.4***	11.6***
	(1.9)	(2.1)
Observations	487	276
Akaike Inf. Crit.	3,818.0	2,095.8
Bayesian Inf. Crit.	3,838.9	2,113.9
Note:		*p<0.1; **p<0.05; ***p<0.01