Significant Error: Citations to Research With Publicized Statistical Errors*

Ken Cor Gaurav Sood

January 16, 2022

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

Summary. Many claims in a scientific article rest on research done by others. But when the claims are based on flawed research, scientific articles spread misinformation. Using 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 cited without noting concerns with the work *more frequently* after the problem was publicized. Our results have implications for the design of scholarship discovery systems and scientific practice more generally.

Keywords. Citation Behavior, Retractions, Scientific Integrity, Scientific Misconduct, Scientometrics

1 Introduction

Retractions are generally a result of serious scientific malpractice. Retractions are also easily identified. Because of these reasons, research on citations to problematic research focuses exclusively on citations to articles that have been officially retracted. However, by focusing on retractions alone, we miss the more common problem of citations to unretracted studies with major errors. For example, Gelman and Stern (2006) discuss a common statistical error where researchers treat differences between significant and non-significant results as significant without conducting the requisite statistical test. They describe a scenario where results from two independent studies report parameter estimates and standard errors of 25 ± 10 and 10 ± 10, respectively. The first result is significant at the 1% level, while the second is non-significant. The finding is interpreted by many as evidence of a significant difference, but a basic calculation of the difference in the effects and its standard error tells a different story, $15 \pm \sqrt{10^2 + 10^2}$, which is not significant at the conventional 95% level.

Nieuwenhuis et al. (2011) analyzed 157 behavioral, systems, and cognitive neuroscience articles that relied on such analysis and were published in journals like *Nature*, *Science*, *Neuron*, and *Journal of Neuroscience* between 2009 and 2010. They found that roughly half (79) of the articles made this error. Further, they found that the error had serious consequences for the results for approximately two-thirds of the studies that made the error. To date, none of the studies that Nieuwenhuis et al. (2011) identified as problematic have been retracted. We assess whether the citation rate changes after the publication of Nieuwenhuis et al. (2011).

Formally, we hypothesize that: Articles identified by Nieuwenhuis et al. (2011) as suffering from a general statistical error will receive fewer citations per year after the publication of Nieuwenhuis et al. (2011) vis-a-vis similar articles without the error.

2 Research Design

To estimate the impact of the publication of error on citation rates, we implement an event study design by tracking the citation rate a few years before and after 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 on citations one, two, and three years after the publication of the retraction notice. We also use a Difference-in-Differences estimator, exploiting the fact that roughly half of the articles published in the same journals did not make the same error.

3 Results

Prima facie evidence suggests little impact of the publication of Nieuwenhuis et al. (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 et al. (2011), and the year Nieuwenhuis et al. (2011) was published (2011), the 79 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 shows 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 1.1). Not all articles making the error, however, have results similarly affected by the error. Fortunately, Nieuwenhuis et al. (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

they are affected by the publication of the Nieuwenhuis et al. (2011). As Figure SI 1.2 shows, the median number of citations steadily and modestly increase over time with the publication of Nieuwenhuis et al. (2011).

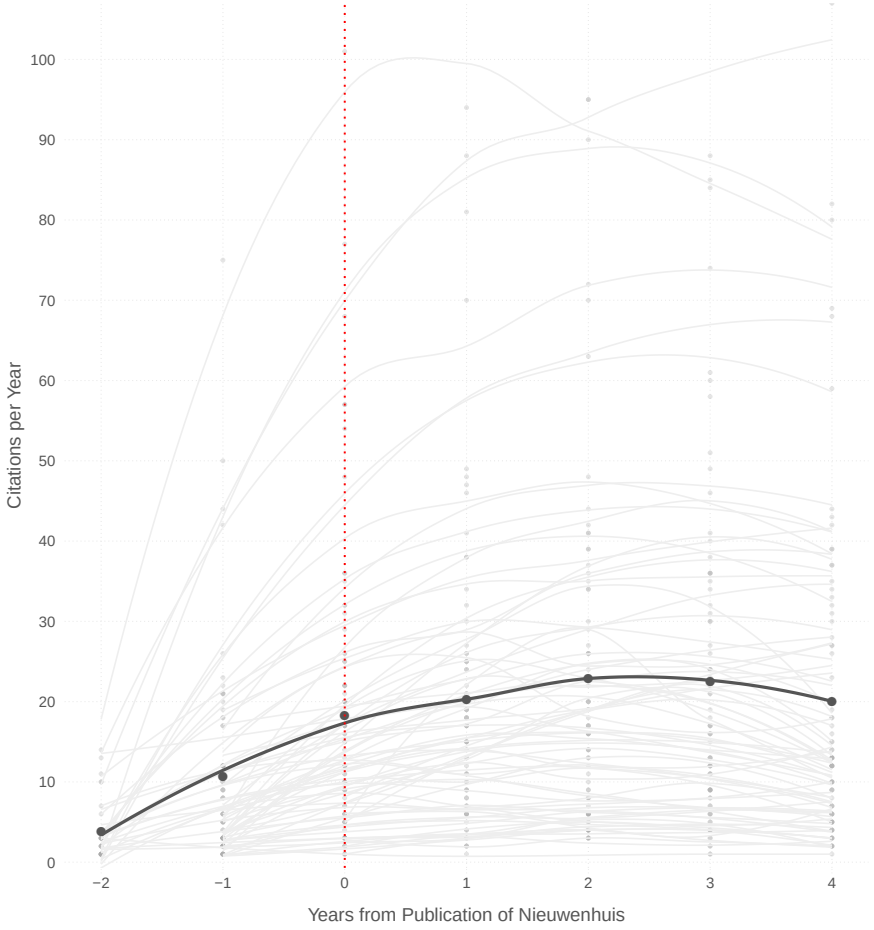

Figure 1: Number of Citations to Articles Containing the Error Per Year

To estimate the percentage of citations that do not acknowledge problems after the publication of Nieuwenhuis et al., we coded whether the citation acknowledged the problem or not in 100 randomly chosen articles citing articles with the mistake (see SI SI 2 for further details about how the citations were coded.) Of the 100, only one article noted concerns with the cited article, citing (Nieuwenhuis et al. 2011) for support.

To more formally explore the change in citation rate as a consequence of the publication of Nieuwenhuis et al. (2011), we regress citations per year on a dummy for the year Nieuwenhuis

et al. (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 estimating an average of within article changes after regressing out a linear time trend as above. Results show, if anything, a modest uptick in citations after Nieuwenhuis et al. (2011) is published—a year after the publication of Nieuwenhuis et al. (2011), articles containing the error get about four more citations per year compared to what they were getting before it (see Table SI 1.1).

Our main analysis for the Nieuwenhuis et al. data is a Difference-in-Differences (DID) analysis. DID gives us a better way to control for over time trends. 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 et al. (2011). In particular, letting i index articles and j index years, we regressed citations per year ($y_i j$) on whether or not the article makes the error (s_i), years to the publication of Nieuwenhuis et al. (n_i) and an interaction between the two. Again, we clustered the standard errors by article. In all, we estimated the following model:

$$y = \alpha + \beta_1 n_i + \beta_2 s_i + \beta_3 (n * s) + \epsilon \tag{1}$$

Table 1 tabulates the results. Models (1), (3), and (5) define error as all articles making the error. Models (2), (4), and (6) refer to error as articles making "potentially serious errors." As the table shows, 1 or 2 years after Nieuwenhuis et al. (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 one and two-year out estimates are closer to 3.5 additional citations per year than 3. 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 citations that do not acknowledge problems remain common

after the error is publicized. The publicity of Nieuwenhuis et al. has had little impact, with articles containing errors still being highly cited. These articles may be especially susceptible to continued citations because the nature of the error results in an inference of a difference in differences that is something other researchers are looking for evidence of in order to support an important point.

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."

References

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

Nieuwenhuis, S., Forstmann, B. U. and Wagenmakers, E.-J. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, **14**, 1105–1107.

Supporting Information

SI 1 Rate of Citations Before and After Publication of Nieuwenhuis et al.

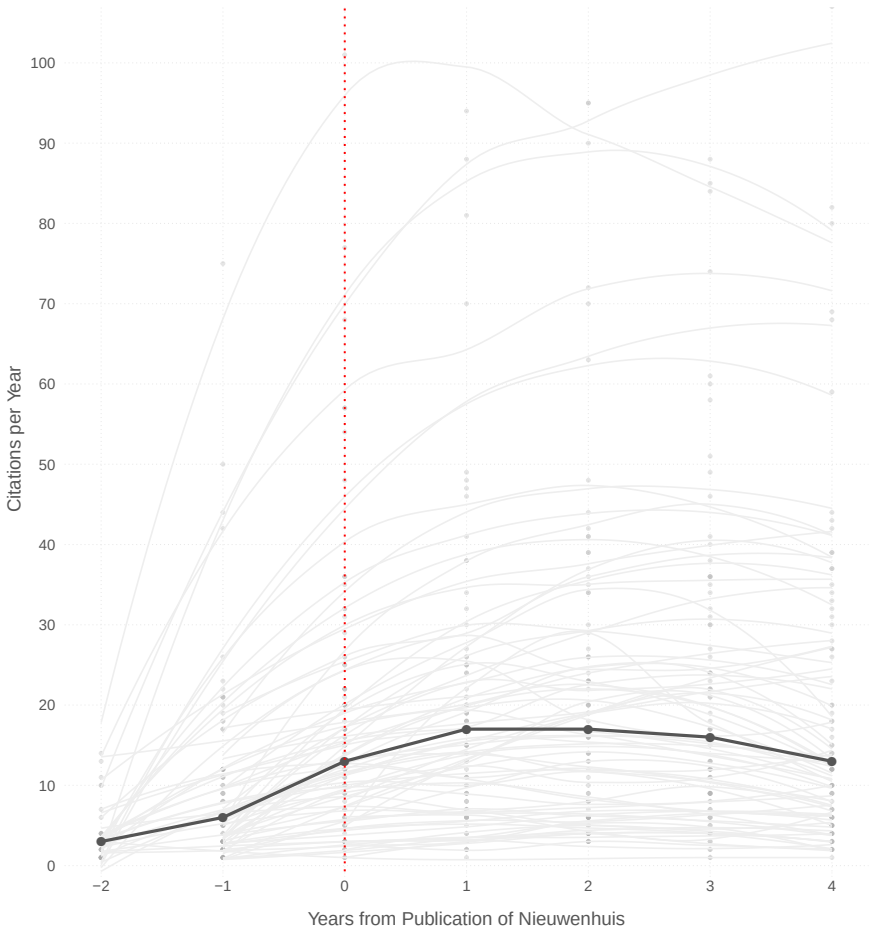

Figure SI 1.1: Total number of citations received per year by each of the papers making the mistake, and the median number of citations received per year by the articles.

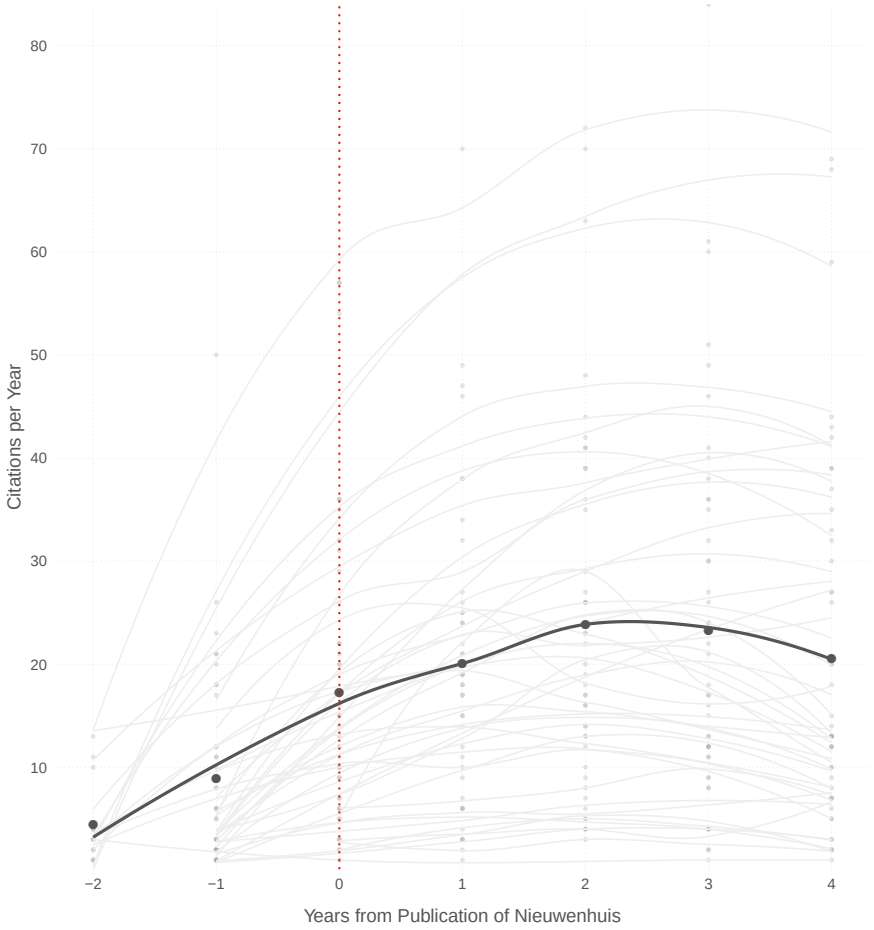

Figure SI 1.2: Total number of citations received per year by articles making the mistake with 'potentially serious' consequences for the results, and the average number of citations received per year by the articles.

Table SI 1.1: 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	20.8	19.8
	(15.6)	(28.6)
Observations	487	276
Akaike Inf. Crit.	3,322.0	1,827.1
Bayesian Inf. Crit.	3,665.4	2,004.5
Note:		*p<0.1; **p<0.05; ***p<0.01

SI 2 Coding Citations as Acknolwedging Problems 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 *not acknoledging problems*. 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 *not acknowledging problems*. We code articles that note any concern with the citing article, even those unrelated to the cause of retraction, as *acknowledging problems*.

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 et al. (2011) for support.

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.