The Career Effects of Scandal: Evidence from Scientific Retractions

Pierre Azoulay
MIT & NBER
pazoulay@mit.edu

Alessandro Bonatti
MIT
bonatti@mit.edu

Joshua L. Krieger
MIT
jkrieger@mit.edu

May 1, 2015

Abstract

Scandals permeate social and economic life, but their consequences have received scant attention in the economics literature. To shed empirical light on this phenomenon, we investigate how the scientific community's perception of a scientist's prior work changes when one of his articles is retracted. Relative to non-retracted control authors, faculty members who experience a retraction see the citation rate to their articles drop by 10% on average, consistent with the Bayesian intuition that the market inferred their work was mediocre all along. We then investigate whether the eminence of the retracted author, and the publicity surrounding the retraction, shape the magnitude of the penalty. We find that eminent scientists are more harshly penalized than their less-distinguished peers in the wake of a retraction, but only in cases involving fraud or misconduct. When the retraction event had it source in "honest mistakes," we find no evidence of differential stigma between high- and low-status faculty members.

*We gratefully acknowledge the financial support of the National Science Foundation through its SciSIP Program (Awards SBE-1460344) and the Sloan Foundation through its Research Program on the Economics of Knowledge Contribution and Distribution. We thank Ezra Zuckerman for insightful conversations. James Sappenfield provided excellent research assistance. The authors also express gratitude to the Association of American Medical Colleges for providing licensed access to the AAMC Faculty Roster, and acknowledge the stewardship of Dr. Hershel Alexander (AAMC Director of Medical School and Faculty Studies). The National Institutes of Health partially supports the AAMC Faculty Roster under contract HHSN263200900009C. All errors are our own.

1 Introduction

In July 1987 Charles Glueck, a leading scientist known for his investigations into the role of cholesterol in heart disease, was censured by the National Institutes of Health (NIH) for serious scientific misconduct in a study he published in *Pediatrics*, a major medical journal (Glueck et al. 1986). At the time the article was retracted, Dr. Glueck was the author of 200 publications that had garnered more than 10,000 citations. The scandal was well-publicized, including two articles in the New York Times calling into question the ability of peer reviewers to root out misconduct in scientific research more generally. Glueck's fall from grace was swift—he had to resign his post from the University of Cincinnati College of Medicine—but also far from complete: he found employment as the Medical Director of The Jewish Hospital Cholesterol Center in Cincinnati, and was still an active researcher as of 2014, though he never again received funding from NIH.

Reputation is a canonical concept in economics. Most of the existing research is concerned with how, and under what conditions, economic agents acquire a good one, separate themselves from other agents with a bad one, and more generally influence the beliefs of the market about their quality (Cabral 2005). The literature is mostly silent, however, about the conditions under which actors can lose their reputation.[1] Across many economic settings, including the realms of entertainment, sports, and the upper echelons of the corporate world, *scandal* looms as one of the primary mechanism through which the mighty are often brought low. Because scandal is at its core an informational phenomenon, this manuscript uses the applied economist's modern toolkit to study the professional fate of individuals whose transgressions are suddenly publicized — to paraphrase the succinct definition of the term provided by Adut (2005).

The reigning paradigm to assess the effects of the revelation of information is Bayesian updating. When the market observes the release of negative information, it might infer that the agent's quality was bad all along, therefore discounting the work that he produced in the past. But our understanding of the updating process itself is still rudimentary. How good is the market at parsing the "truth" from signals of varying informativeness? How does the extent of the penalty depend on the stock of reputation accumulated by the agent up until the time of the shock?

To answer these questions empirically, we turn to the setting of scientific retractions. We start from a list of 878 biomedical research articles retracted during a period that spans the years 1980 to 2009. We carefully match the authors of these publications to the Faculty Roster of the Association of American Medical Colleges (AAMC), a comprehensive panel dataset recording the career histories of U.S. academic biomedical researchers. This generates a list 376 US-based

[1] This omission stands in marked contrast to sociology, where this question occupies a central place in the disciplinary cannon (Goffman 1963; Fine 2001). A notable exception in economics is the model of Cripps, Mailath and Samuelson (2004); these authors demonstrate how reputation is necessarily a short-run phenomenon when actions are imperfectly monitored.

retracted faculty authors for whom we painstakingly assemble a curated list of publications, NIH grants, and citations received. We proceed in a symmetric fashion to produce a sample of 759 control authors in the same broad fields of the life sciences.

Armed with these data, we then analyze the impact of retraction events on the rate of citation received by non-retracted articles published prior to the retraction in a difference-in-differences framework. Perhaps unsurprisingly, we find that the pre-retraction work of retracted authors suffers a 10% citation penalty following a retraction event, relative to the fate of the articles published by non-retracted control authors.

We then investigate the impact of prior reputation (whether the authors belonged to the top quartile of the citation or funding distribution at the time of the retraction) and the informativeness of the signal contained in the retraction news, which we proxy by carefully separating instances of misconduct (from fraud to plagiarism) from instances of mistakes (stemming, for example, from contaminated biological samples or statistical errors). Our results indicate that informativeness and prior reputation interact in very specific ways to shape the magnitude of the audience's response. In particular, the work of eminent authors is not penalized more severely than that of less eminent ones in the case of honest mistakes. However, the difference in citation penalty is much more pronounced when retraction events stem from clear-cut cases of scientific misconduct. In these instances, the prior work of retracted authors sees its rate of citation fall by almost 20%.

Our study bears a resemblance with a recent paper by Jin et al. (2013). These authors also study the effect of retraction events on the citations received by prior work from retracted authors, but they focus on the differential penalty suffered by junior and senior authors on the same retracted paper. They find that the senior authors (those in last authorship position) escape mostly unscathed following a retraction, whereas their junior collaborators (typically graduate students of postdoctoral fellows) are often penalized severely, sometimes to the point of seeing their careers brought to an abrupt end. Their results are seemingly at odds with ours, but it is important to note that the variation we exploit exists between authorship teams, rather than within them.

The manuscript proceeds as follows. The next section summarizes the institutional context of retractions as part of the broader scientific peer review system. Section 3 introduces a simple Bayesian model to frame the empirical exercise. Section 4 describes the data and the process followed to assemble it. Section 5 presents our empirical strategy and results. Section 6 circles back to the model to discuss the extent to which the market's reaction is, in fact, consistent with Bayesian learning. Section 7 briefly concludes.

2 Institutional Setting

While the role of scientific research in enabling economic growth has become a truism among economists, scientific progress does not unfold in an institutional vacuum. Rather, the scientific enterprise relies on a set of reinforcing institutions that support individual accountability and reliable knowledge accumulation (Merton 1973; Dasgupta and David 1994). In the context of this manuscript, peer review, the allocation of credit through citation, and the retraction system are three fundamental practices worthy of discussion.

One of the central institution of science is the peer-review system. By submitting scientific articles for independent review by expert peers, the path to publication balances the integrity of published results with the desire to have an adequate pace of discovery. Similarly, the practice of citing relevant prior literature allows scientists to clearly and concisely communicate where there contributions fall within the scientific landscape, while allocating credit to the originators of particular ideas.

Retractions are often the culmination of a process used by journals to alert readers when articles they published in the past should be removed from the scientific literature. They are qualitatively different from simple corrections in that their intent is to strike the entire publication from the scientific record. Retraction notices may be initiated by the journal editors, by all or some of the authors of the original publication, or at the request of the authors' employer.

The informational content of retraction notices is highly variable. Some notices contain detailed explanations about the rationale for the decision to retract, while others are a single sentence long and leave the scientific community uncertain about (i) whether the results contained therein should be disregarded in part or in their entirety, and (ii) whether the retraction was due to fraud, more benign forms of scientific misconduct, or instead had its root in an "honest mistake."

In the recent past, specialized information resources, such as the popular blog *RetractionWatch*, have emerged to help scientists interpret the context surrounding specific retraction events. One aspect of a retraction's "back story" that often proves vexing to decipher pertains to the allocation of blame across members of the authorship team. Only in the most egregious and clear-cut instances of fraud would a retraction notice single out particular individuals. In the United States and for research supported by NIH, scientific misconduct is also policed by the Office of Research Integrity (ORI) within the Department of Health and Human Services. ORI is vested with broad investigative powers, and its reports are often the forerunners of retraction events, sometimes involving more than a single publication.

Retraction events are still rare (occurring at the rate of roughly one retraction per ten thousand scientific articles), but their frequency has been increasing steadily over the past 20 years (see Figure 1). This trend has been the cause of increasing concern in the media (e.g., Wade 2010; Van

Noorden 2011), and much hand-wringing within the scientific community (Fang et al. 2012), but its fundamental drivers remain an open question. While popular accounts espouse the view that heightened competition for funding leads to increased levels of sloppiness, scientists can also gain prominence by detecting instances of misconduct or error (Lacetera and Zirulia 2009). Moreover, the rise of the Internet and electronic resources has in all likelihood increased the speed at which peers can direct their attention to results that are both noteworthy and *ex-post* difficult to replicate. Much of the public attention to the retraction phenomenon can also be attributed to a handful of high-profile cases of scientific misconduct.[2]

3 Model

We introduce a simple model of Bayesian learning and scientific reputations as a guiding framework for our empirical results. In our model, a representative researcher (the *agent*) is continuously evaluated by the scientific community (the *market*). The agent has a fixed binary characteristic that denotes his "quality,"

$$\theta \in \{\theta_B, \theta_G\}.$$

Thus, the agent is either "good" or "bad." The market's prior belief that the agent is of the good type is $p^0 \triangleq \Pr(\theta = \theta_G)$.

The market learns about the agent's quality from observing his scientific output. For simplicity, we assume that the agent's output at each point in time is also binary,

$$y_t \in \{0, 1\}.$$

In particular, output at time t is given by $y_t = 1$, unless a *retraction event* occurs, in which case output is given by $y_t = 0$.

The market rewards the agent with citations based on his reputation. In other words, the flow of citations received by the agent's body of work is a function of the market's belief that his quality is high, based on his output history. Let p_t denote the market's posterior belief at time t. The flow of citations to any of the agent's papers at time t is given by $w(p_t)$, where w is a strictly increasing and twice differentiable function. Rewards for reputation are, in fact, highly nonlinear in our database (see Figure 6), where the distribution of citations is heavily skewed towards "superstar" agents.[3]

[2]Stem cell science has been rocked by two especially sensational scandals. The first was the case of Woo-suk Hwang—the South Korean scientists who fabricated experiments and claimed to have successfully cloned human embryonic stem cells. More recently, the media gave major coverage to the retraction of a stem cell paper that claimed to use acid baths to turn mature cells into stem cells. Tragically, one of the Japanese authors on the retracted paper, Yoshiki Sasai, committed suicide at his research lab.

[3]Both the agent's output and the reward function can be made endogenous by introducing an explicit choice of effort, as in the career concerns model of Holmström (1999).

3.1 Learning and Reputations

The market learns about the agent's quality through retractions that we model as a "bad news" Poisson process. The intensity of the Poisson process is higher for low-quality (bad) agents. Thus, retractions are rare, publicly observable events that reveal information about an agent's quality.[4]

More formally, let time be continuous and the horizon infinite. Retraction events for an agent of type θ are exponentially distributed with parameter λ_θ, where we assume that $\lambda_B \geq \lambda_G \geq 0$. Under this learning model, the agent's reputation (measured by the market's belief p_t) drifts upward over time, until a retraction event occurs, in which case it jumps down. Our empirical approach is focused on the drop in *citations* following a retraction, and does not aim to capture the more nuanced dynamics of reputations. Consequently, we now examine on the effect of a retraction on the market's *beliefs*.

Upon observing a retraction at time t, the market updates its beliefs from p_t to

$$p_{t+\mathrm{d}t} \triangleq \Pr\left[\theta = \theta_G \mid y_t = 0, t\right] = \frac{p_t \lambda_G}{p_t \lambda_G + (1 - p_t)\,\lambda_B}.$$

The change in the agent's reputation is then given by $\Delta\left(p_t\right) < 0$, where

$$\Delta\left(p_t\right) \triangleq p_{t+\mathrm{d}t} - p_t = -\frac{p_t\left(1 - p_t\right)\left(\lambda_B - \lambda_G\right)}{p_t \lambda_G + (1 - p_t)\,\lambda_B}.$$

If $\lambda_G = 0$, the expressions above yield $p_{t+\mathrm{d}t} = 0$ and $\Delta\left(p_t\right) = -p_t$. In other words, when the retraction event is fully revealing of a bad type, the agent loses his entire reputation, regardless of its initial level. Conversely, if $\lambda_G = \lambda_B$, then $\Delta\left(p_t\right) = 0$. Thus, when retraction events are uninformative, they cause no change in reputations.

We now introduce a measure of the informativeness of retractions, namely

$$\alpha \triangleq \frac{\lambda_B}{\lambda_G} \geq 1.$$

Let p denote the agent's current reputation level. We can rewrite the change in reputation as

$$\Delta\left(p, \alpha\right) = -\frac{p\left(1 - p\right)\left(\alpha - 1\right)}{p + (1 - p)\,\alpha}. \tag{1}$$

In particular, if signals are uninformative ($\alpha = 1$), then $\Delta\left(p, 1\right) = 0$ for all p. Conversely, if signals become arbitrarily informative ($\alpha \to \infty$), then $\Delta\left(p, \alpha\right) \to -p$. More generally, $\Delta\left(p, \alpha\right)$ is a non-monotone function of an agent's reputation.[5] In particular, $\Delta(p, \alpha)$ attains a minimum (corresponding to the greatest loss in reputation) for $p = 1 - 1/(\sqrt{\alpha} + 1)$. Figure 2 illustrates $\Delta\left(\cdot, \alpha\right)$ for several values of α, and Proposition 1 collects our comparative statics results.

[4]The model can easily be extended to simultaneously account for several informative signals, e.g., retractions due to misconduct and to mistakes. We do so in Appendix A. It is also possible to include both positive and negative events, such as scientific breakthroughs that augment the agent's reputation.

[5]In our Bayesian model, if the market holds beliefs $p = 1$, the agent's reputation is unaffected by retractions. We assume that some arbitrarily small amount of uncertainty persists for all scientists.

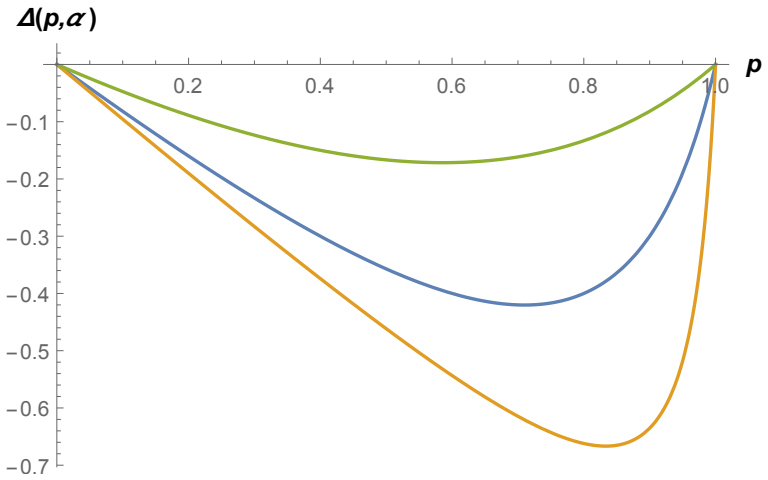


Figure 2: Reputation Losses $\Delta(p, \alpha)$, for $\alpha \in \{2, 6, 25\}$

Proposition 1 (Signal Informativeness) *As the signal informativeness α increases:*

1. $\partial\Delta(p,\alpha)/\partial\alpha < 0$ *for all p, i.e. retractions yield greater losses of reputation.*
2. $\partial^2\Delta(p,\alpha)/\partial\alpha\partial p < 0$ *for all $p \in [0, \alpha/(\alpha+1)]$, i.e., over that range, reputation losses are increasing in the prior reputation p.*

The bound $\alpha/(\alpha+1)$ in part (2.) approaches 1 as $\alpha \to \infty$. Thus, as signals become arbitrarily precise, the negative effect of an increase in signal precision is greatest for the agents with the highest reputation.

3.2 Implications for Citations

We now turn to the average effect of a retraction on agents with similar reputation levels. We consider a population of agents whose reputations p are uniformly distributed.[6] We then aggregate the reputation losses of the top quartile and of the bottom three quartiles in the population. Finally, we compare the effect of a retraction across different levels of signal informativeness. Figure 3 illustrates the aggregate implications of our model when $\alpha \in \{6, 25\}$.

The example in Figure 3 is broadly consistent with our main empirical finding – that high-status agents after a retraction even due to misconduct suffer the sharpest drop in reputation, while the three other reputation losses are of comparable magnitude to one another.

In Appendix A, we formalize the intuition that an increase in signal informativeness amplifies the difference in the reputation losses of high- and low-status agents. In particular, we consider the average reputation drop for agents with initial levels of reputation $p \in [0, p^*]$ and $p \in [p^*, 1]$,

[6]This is a useful special case because in our empirical approach we will use quantiles of the citations and funding distributions as proxies for reputation.

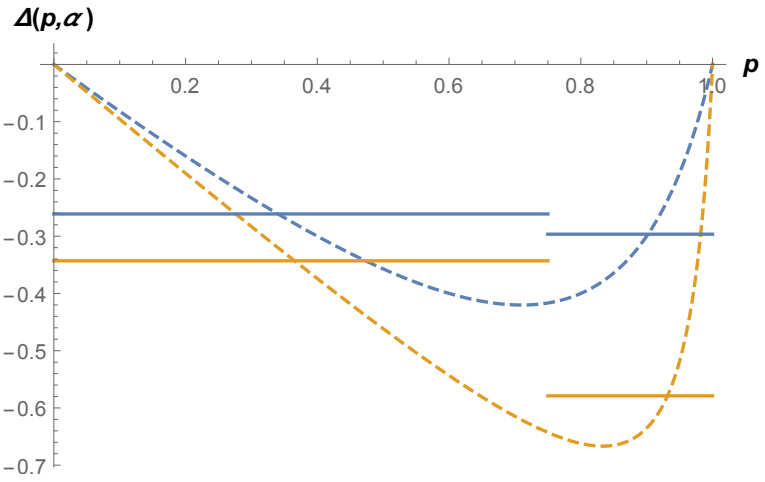


Figure 3: Average Reputation Loss by Group, $\alpha \in \{6, 25\}$

respectively. We show that, if one considers a large enough subset of high-status agents (low p^*), then an increase in the informativeness of the retraction signal α widens the gap between the reputation losses of high- and low-status agents.

We conclude this section by analyzing the implications of Bayesian learning for the drop in citations following a retraction by an agent with initial reputation p. In order to correctly capture the effect of a retraction, we must consider two elements: the shape of the rewards for reputation $w(p)$; and the drop in the market's beliefs $\Delta(p)$. The change in the flow of citations is given by

$$\Delta\left(w\left(p_t\right)\right) \triangleq w\left(p_{t+1}\right) - w\left(p_t\right).$$

Consider, for example, an exponential reward function $w(p) = e^p$. We can then write the percentage drop in citations as

$$\frac{d \ln w\left(p\right)}{dp} = \Delta\left(p\right).$$

Thus, under an exponential reward function, the results of Proposition 1 that relate the dynamics of reputation p to the signal informativeness α also apply to the *relative drop* in citations $w(p)$.

The exponential rewards function is a reasonable approximation to the distribution of citations and funding at baseline in our data. Consequently, in our empirical analysis, we shall report regression results in *logs*, and apply the insights derived earlier for the reputation *levels*.

4 Data Construction

This section details the construction of our multilevel, panel dataset. We begin by describing the criteria used to select the sample of retracted scientists and how we identified their career and

publication histories. Next, we present the outcome variables used in the study, as well as our measures of publicity and author prestige. The last step is to explicate the process through which a sample of control authors—faculty members who did not experience a retraction event, but are otherwise similar to the retracted authors—was selected.

Retractions, retracted authors, and career histories. In order to build our sample of retracted authors and their publication histories, we begin with a set of 1,129 retractions published in the period 1977-2007, and retracted prior to 2009. The source of these retractions is *PubMed*, the United States National Library of Medicine's (NLM) primary database for biomedical and life science publications. *PubMed* contains more than 24 million citations and indexes articles along a number of dimensions, including retraction status.

To analyze the impact of retraction events on scientist citation trajectories, we carefully matched retracted authors to the Faculty Roster of the Association of American Medical Colleges (AAMC), to which we secured licensed access for the years 1975 through 2006, and which we augmented using NIH grantee information (cf. Azoulay et al. [2010] for more details).[7] Whenever the authors we identify in this way were trainees (graduate students or postdoctoral fellows) at the time of the retraction event, we exclude them from the sample.[8]

We were able to match at least one author on 43% of the retracted publications to scientists in the AAMC Faculty Roster. While this figure may seem low, it is a reflection of the fact that the majority of the retractions are authored by non-US scientists who would not, by definition, be expected to appear in the AAMC Faculty Roster. The match rate for American scientists is much higher. Of the 488 retractions with US reprint addresses, we matched at least one author on 412 (84%) of the publications. The matching process yielded 195 retractions with one author matched, 148 retractions with two authors matched, and 146 retractions with three or more authors matched. Since many of these authors are involved in multiple retractions, matched authors have an average of 1.5 retracted publications in the sample. Our analyses focused on 878 retractions where retraction at least partially invalidated the original paper's claims.[9] From this sample of retractions, we matched a total of 376 retracted faculty authors. For more information on the author matching process, see Appendix B.

[7] An important implication of our reliance on these source of data is that we can only identify authors who are faculty members in U.S. medical schools, or recipient of NIH funding. In particular, we cannot systematically identify trainees, staff scientists without a faculty position, scientists working for industrial firms, or scientists employed in foreign academic institutions. The great benefit of using these data, however, is that they ensure we know quite a bit about the individuals we are able to identify: their (career) age, type of degree awarded, place of employment, gender, and research output, whether measured by publications or NIH grants.

[8] We do so because these trainees-turned-faculty members are selected in a non-random fashion from the entire population of trainees which we cannot get systematic data about.

[9] As in Azoulay et al. (2015), our analyses exclude the 202 retraction cases where the retracted paper's claims remain valid after the retraction event (i.e., most—but not all—cases of plagiarism, duplication of publications, faulty IRB approval, etc.) We verified that including these retractions in the sample does not materially affect our conclusions.

Once matched to the AAMC Faculty Roster, we linked authors to their publication histories by developing detailed *PubMed* search queries that return the author's entire body of work. Figure 4 illustrates this process for the case of one faculty member, Kirk E. Sperber, MD. More details regarding the procedure used to link authors with their publication histories can be found in Appendix C.

Citation data. The primary outcome in the analyses presented below is the annual flow of citations to authors' publications in the sample. Citations are both a measure of intellectual credit and professional attention. Scientists cite prior work in order to communicate where their contributions fall within their field or subfield, and to bestow credit to the research they are building upon. Citations also serve as the currency that is essential to maintaining the incentives and norms that compel honest work and competition in science (Merton 1957).[10] We follow in the footsteps of prior scholarship in the economics of science in using an information shock to trace out the effect of this shock on the citation trajectories of scientific articles published before the shock (e.g., Furman and Stern 2011; Azoulay et al. 2015).

Since *PubMed* does not provide citation data, we use Thomson-Reuters' *Web of Science* (WoS) to obtain citations for publications in *PubMed*. We match *PubMed* with *WoS* to generate a dataset with 190 million cited-to-citing paper pairs. This resulting dataset contains cited-to-citing pairs for all *PubMed*-indexed articles that cite other *PubMed*-indexed articles.[11] Our analyses exclude all self-citations from any member of the authorship team.

Measuring the public nature of retraction events: misconduct vs. "honest mistakes." An important implication of our model is that the informativeness of the signal contained in a retraction event should influence the extent to which the market updates on the quality of a scientist's prior work. It is therefore essential to develop a way of capturing systematically the extent to which retraction events are publicized beyond the confines of a scientist's narrow intellectual community. In the post-Internet era, it is possible to develop direct measures of publicity using the electronic trails as meaningful "breadcrumbs." But this approach is not practical for our sample, which includes data from an earlier period. We settle on the distinction between misconduct and

[10] Citations can also be used for less noble purposes such as appeasing editors and reviewers by adding citations, or making larger claims by reducing the number of citations. It is a limitation of our study that we do not have the ability to determine which cites are "strategic" rather than "substantive" (cf. Lampe [2012] for examples of such strategic citation in the case of patents).

[11] In a separate analysis, available from the authors, we found that citations from *PubMed*-indexed articles to *PubMed*-indexed articles that are also in the Web of Science account for 86% of the total number of citations that are received by these articles in a sample of 320,000 articles carefully matched between the two sources of data. The correlation between *PubMed*-to-*PubMed* citations and *WoS*-to-*PubMed* citations is higher than .99. We conclude that our decision to focus on the *PubMed*-to-*PubMed* citation information for the analyses presented in this paper is innocuous.

"honest mistakes" as a pragmatic solution to the challenge of proxying for the extent to which retraction events enter the public sphere.[12]

In order to differentiate between retractions due to misconduct and retractions due to mistakes, we used the misconduct codes assigned to retractions in Azoulay et al. (2015). These codes required manual review of every retraction and their associated public documents (Appendix D provides more details on the assignment of these codes). The difference between retractions due to misconduct and mistakes is often quite stark. The case of anesthesiologist Scott Reuben is a clear-cut example of retractions due to misconduct. As a professor at Tufts University purportedly running clinical trials on the effectiveness of painkillers, Reuben fabricated data and conclusions. He was charged with and found guilty of health care fraud, resulting in a sentence of six months in federal prison and over $400,000 in fines and restitution. Our retractions data set contains 15 of his publications, many of which were simultaneously retracted.

Retractions due to mistakes tend to be less sensational. Contaminated samples and reagents are the most frequent reasons for mistake retractions. Occasionally, authors also retract papers due to flawed interpretation of results, or conclusions nullified by subsequent studies. In one case, the authors retracted a publication after realizing that they mistakenly analyzed the genetic code of a butterfly rather than a dragonfly (Arikawa et al. 1996).

Measures of author prestige. The seminal work of Merton (1968) alerted scholars that recognition and rewards for a given level of achievement are more likely to accrue to scientists whose reputation was already established, a phenomenon known as the "Matthew Effect." As pointed out by Jin et al. (2013), the retraction phenomenon presents an opportunity to ask whether the Matthew Effect also operates in reverse, that is, whether more prominent are penalized more harshly by the scientific community in the wake of a retraction than their less-distinguished peers. In their work, Jin et al. (2013) choose to operationalize prior prestige using authorship position on the retracted article. Given the prevailing authorship norms in most of natural and life sciences, this approach effectively distinguishes between high and low-status scientists within a research team (i.e., graduate student or postdoctoral fellow vs. faculty member or principal investigator).

Because we have at our disposal detailed career and publication histories for each of the scientists in our sample, we adopt a strategy to measure variation in prior prestige that is more global in nature. In a first step, we compute each matched author's cumulative citation count, across all of their publications, through the year before their first retraction. We define "high-status" scientists as those authors who belong in the top quartile of this distribution at baseline, and those whose cumulative citations place them in the bottom three quartiles as "low-status." Using this measure,

[12] Certainly, instances of fraud and misconduct attract much more attention in the comment sections of specialized blogs such as *RetractionWatch*. Very few, if any, instances of retraction due to mere error lead to editorials, pontification, or hand-wringing in scientific journals or the national press.

high-status scientists account for 58% of all of the articles published by retracted authors up to the year of their first retraction.

In a second step, we also compute cumulative funding from the National Institutes of Health (NIH). Again, we defined high-status authors (resp. low-status) as those in the top quartile (resp. bottom three quartiles) of the corresponding distribution at baseline. The high-funding group accounts for 47% of all the articles published by retracted authors up to the year of their first retraction.[13]

Identifying and selecting control authors. To shed light on the counterfactual citation trajectories of retracted authors' pre-retraction publications, we looked to the authors of the articles immediately preceding and following the retracted publication in the same journal/issue. Using adjacent articles to construct a control group for a set of treated articles is an approach pioneered by Furman and Stern (2011), and adopted by Furman et al. (2012), and Azoulay et al. (2015).[14]

To identify control authors, we follow a procedure which mirrors in all respects the process we adopted to identify treated authors in the sample of retracted articles in all respects. The final analytic sample of faculty members includes only retracted authors for whom we have located at least one matched control author. In total, we have 759 such control authors.

Descriptive statistics. Our sample includes 23,620 publications by 376 retracted authors and 46,538 by the 759 control authors.[15] Since each control faculty member entered the dataset because it is the author of a paper that appeared in the same journal and issue as a retracted paper, we can assign to them a counterfactual date of retraction, which is the year in which the retracted author to which they are indirectly paired experienced a retraction event. Table 1 compares treated and control authors along demographic dimensions, such as gender, degree, career age, and eminence (measured as cumulative citations as well as cumulative funding). Retracted authors are slightly more likely to be male, and also have slightly higher cumulative funding and citation impact as of one year before the earliest associated retraction event, relative to control authors. Below, we will show that these small differences in baseline achievement levels do not translate into differences in achievement *trends* before the treatment.

[13] We also used average citations per publication and average yearly funding as measures of prestige, and the results were similar to those we present below. We considered using membership in the National Academy of Sciences (NAS) as an additional measure of author prestige. However, this measure did not give us enough power to perform our analysis as only 3.6% of the authors in our sample were members of the NAS at baseline.

[14] One can think of different choices to identify a set of potential control authors, including choosing a random article in the same journal/issue as the treated article, or all non-retracted articles in the same journal/issue. In past work, we showed that there is very little difference between choosing a "random neighbor" as opposed to a "nearest neighbor" (Azoulay et al. 2015). The second strategy (use all non-retracted articles in the journal/issue as a source of control authors) would have been prohibitively time-consuming: the process followed to match authors to the AAMC Faculty Roster and link them with their publication output is very labor intensive.

[15] The publications we considered for inclusion in the sample include only original research articles, and exclude reviews, editorials, comments, etc.

Appendix D provides details regarding the extent to which specific authors were singled out as particularly blameworthy. The assignment of blame was unambiguous for only 24 out of the 376 retracted authors in the sample (6.38%). The majority of blamed authors are precisely the types of scientists that would be less likely to ever appear in the AAMC Faculty Roster: graduate students, postdoctoral fellows, or technicians.[16] Moreover, the set of blamed authors is a proper subset of authors whose work was retracted because of misconduct; in our data, there is not a single example of an article retracted because of a mistake which laid blame for the event at the feet of a specific member of the research team. As a result, while the "blamed" indicator variable is interesting from a descriptive standpoint, we will not use it in the rest of the analysis.

Table 2 presents descriptive statistics at the level of the author/article pair, which is also the level of analysis in the econometric exercise. The stock of citations received up to the year of retraction is well balanced between treated and control articles. This is the case not simply for the mean and median of these distributions, but for other quantiles as well (see Figure 5). Figure 6 provides evidence of the skew in the distribution of eminence at baseline, measured in terms of cumulative citations (Panel A) and cumulative NIH funding (Panel B). These quantile plots provide some empirical justification for splitting our sample along the top quartile of these distributions to distinguish the effect of retractions on eminent (top quartile) and less distinguished (bottom three quartiles) scholars.

5 Methodological Considerations and Results

5.1 Identification Strategy

To identify the impact of retractions on author reputations, we examine citations to the authors' pre-retraction work, before and after the retraction event, and relative to the corresponding change for control authors. Retraction events may influence a number of subsequent research inputs, including effort, flow of funding, referee beliefs, and collaborator behavior. Since our goal is to measure the scientific community's response to sudden changes in the reputation of individual faculty members embroiled in retraction cases, we focus on pre-retraction publications only. The quality of these publications is not affected by subsequent changes to the research environment. The difference-in-differences research design allows us to measure the impact of retractions, while accounting for life-cycle and time-period effects that might be shared by retracted and non-retracted authors.

A maintained assumption in this approach is the absence of citation trends that might affect the pre-retracted articles of retracted authors, relative to control authors. Preexisting trends loom

[16] Retraction events at such an early stage of one's career would certainly decrease the likelihood of ever holding a faculty position in the future.

especially large as a concern because prior research has demonstrated that retracted articles exhibit a pronounced citation uptick (relative to articles published in the same issue) in the months and years immediately leading up to the retraction event (Furman et al. 2012). Fortunately, we can evaluate the validity of the control group *ex post*, by flexibly interacting the treatment effect with a full series of indicator variables corresponding to years before and after the retraction date. This is a common diagnostic test with a difference-in-differences research design, and its result will be reported below.

An additional issue could confound the interpretation of the results. We have modeled the process through which the scientific community updates its beliefs regarding the reputation of individual scientists following a retraction. Empirically, this response might be commingled with learning about the foundations of the intellectual area to which the retraction contributed. Indeed, prior work has shown that non-retracted articles related to the same line of scientific inquiry see their rate of citation drop in the wake of a retraction (Azoulay et al. 2015). To filter out this aspect of the learning process, we focus on pre-retracted work by the retracted authors that does not belong to the same narrow subfield as the underlying retraction.

In practice, we use the topic-based *PubMed* Related Citations Algorithm (PMRA) to define intellectual fields (see Appendix E). We remove all publications that are related (in the sense that PMRA lists them as a related citation) to the source article. These deletions are performed in a parallel fashion for both treated and control authors. In total, we remove 12.2% of retracted authors' pre-retraction publications that were in the same PMRA field as one of their retracted articles, and 9.2% of control authors pre-retraction publications that were in the same PMRA field as their source publications (i.e., the article adjacent to the retraction in the same journal/issue). The descriptive statistics above, and the econometric analyses below refer only to this sample of author/publication pairs without the set of in-field publications.

5.2 Econometric Considerations

Our econometric model relates the number of citations to author i's pre-retraction article j received in year t to characteristics of both i and j:

$$E\left[y_{ijt}|X_{it}\right] = exp\left[\beta_0 + \beta_1 RETRACTED_i \times AFTER_{jt} + \phi(AGE_{it}) + \psi(AGE_{jt}) + \delta_t + \gamma_{ij}\right]$$

where $AFTER$ is an indicator variable that switches to one in the year during which author i's experiences his first retraction, $RETRACTED$ is equal to one for retracted authors and zero for control authors, the age functions ϕ and ψ are flexible functions of author age and article age consisting of 50 and 33 indicator variables (respectively), the δ_t's represent a full set of calendar year indicator variables, and the γ_{ij}'s are fixed effects corresponding to author-publications pairs.

The dependent variable y_{ijt} is the number of forward citations received by author i's article j in year t (excluding self-citations). About 44% of all observations in the sample correspond to years in which the article received exactly zero citations. We follow the long-standing practice in the analysis of bibliometric data to use the conditional fixed-effect Poisson model due to Hall et al. (1984), which we estimate by quasi-maximum likelihood (Gouriéroux et al. 1984; Wooldridge 1997). The standard errors are robust, and clustered at the level of individual authors.

5.3 Econometric Results

We report the results of the simple difference-in-differences specification in Table 3, column 1. The coefficient estimate implies that, following a retraction event, the rate of citation to retracted author's unrelated work published before the retraction drops by 10.7% relative to the citation trajectories of articles published by control authors.

Figure 7 displays the results of the dynamic version of the model estimated in column 1. We interact the treatment effect variable with indicator variables for number of years until (respectively after) the author's earliest retraction event. We graph the estimates corresponding to these interaction terms along with the associated 95% confidence intervals. Relative to control authors, the retracted authors' pre-retraction publications receive slightly more citations in the pre-retraction period; however, this difference appears to be roughly constant in the years leading up to retraction—there is no evidence of a pre-trend, validating ex post our research design and control group. Figure 6 also shows that the citation penalty appears to increase over time; it appears to be a permanent, and not merely transitory, phenomenon.

Exploring heterogeneity in the retraction effect. We begin by splitting the sample into high- and low-status subgroups, first using cumulative citations as a marker of eminence (Table 3, columns 2a and 2b), second using cumulative funding (Table 3, columns 3a and 3b). Since high-status authors tend to produce more publications, splitting the sample by separating the top quartile of each status metric from its bottom three quartiles yields subsamples of approximately equivalent size. We cannot detect large differences in the magnitude of the treatment effects across these groupings. Even in the case of funding, where there is a slightly larger difference in the post-retraction penalty for low-status faculty members (7.6% vs. 12.2% decrease), this difference is in itself not statistically significant.

The next step is to split the sample by separating instances of misconduct from instances of mere error (see Appendix D). The estimates reported in columns (4a) and (4b) do suggest a much stronger market response when the retraction event comes closer to our definition of scandal, that is when the event is more publicized, as is usually the case when misconduct or fraud are alleged (17.6% vs. 8.2% decrease).

Interaction between prior eminence and the informativeness of the retraction event. Table 4 splits the sample into four subgroups, corresponding to both the status and misconduct dimensions. One result stands out qualitatively: the high-status authors are more harshly penalized than their less-distinguished peers, but only in instances of misconduct (columns 1b and 2b). In all other subgroups, the differences in the magnitude of the treatment effect are modest at best.

6 Discussion

Three different comparisons bear directly on the suitability of our simple Bayesian framework to explain the empirical patterns that emerge from the econometric analysis.

First, for authors of any status, the effect of a retraction due to misconduct is larger than the effect of a retraction due to mistake (Table 3, columns 4a and 4b). This result is consistent with a model where a misconduct retraction is a more informative signal of an author's bad quality, *i.e.*, a higher α in Proposition 1. See Figure 2 for the intuition behind this result.

Second, the most significant effect of retractions occurs after a misconduct event for authors in the top status quartile. Furthermore, citation penalties for all other event type/author status combinations have a lower and relatively homogeneous effect (Table 4). The aggregate implications of our model match these regression results (see Figure 3 for a simple illustration). When a signal is very informative, it has a large impact on an author's reputation, independently of its initial level. The resulting loss of reputation is therefore largest for high-status authors. Conversely, when the signal is not particularly informative, the reputation loss is mostly tied to the initial level of uncertainty. This is highest for agents with intermediate reputations, which means very high- and very-low status authors experience similar drops in reputation.

Third, we can go one step beyond the binary distinction between high- and low-status authors. We do not have sufficient statistical power to recover the full shapes of the reputation loss as characterized in our model, for example in Figure 2. Instead, to generate the coefficients graphed in Figure 8, we partition authors into quintiles of the status distribution.[17] We then contrast the effects of different types of retraction events for each of five status grouping. Figure 7, Panel A suggests that the largest drop in citations following a mistake occurs for scientists with intermediate reputation levels (the third quintile). Conversely, the drop in citations following misconduct is largest for the highest-status scientists (fourth and fifth quintiles in Figure 7, Panel B).[18]

Together, these results suggest that the market response to a retraction event is consistent with Bayesian learning about the author's quality. In particular, the distinct responses to mistakes and

[17] In this case, status is only measured by cumulative citation count at the time of the retraction.

[18] These statements must be interpreted with a great deal of caution, since the sample size is too small for these differences between coefficient estimates to be statistically significant. We only mean to suggest that their overall pattern is consistent with the more nuanced implications of our model.

misconduct indicate that the market considers misconduct events as more precisely revealing the (low) quality of an individual scientist, relative to instances of "honest mistake."

From this standpoint, the fact that the ratio of misconduct and mistake retractions is about the same for both high and low-status authors (Table 5) is an anomaly. While high-status scientists experience fewer retractions overall, observing a mistake vs. misconduct retraction is not particularly helpful to predict the eminence of a retracted author. If misconduct is a more informative signal, and high-status scientists are, in fact, of higher average quality, we would expect them to exhibit a lower *misconduct-to-mistake* ratio.

There are two distinct explanations for the discrepancy between the empirical distribution of retraction events and the theory consistent with an equilibrium market response. Both explanations point to forces that lie outside the relatively simple model presented in Section 3: the market "overreacts" to the misconduct information; or the authors adjust their behavior in response to the market incentives. We examine both possibilities.

Market overreaction. It is possible that the market simply overestimates the informativeness of misconduct. However, the citation penalty may represent more than just the market's response to an information shock. For instance, it may be part of an implicit incentive scheme that sees ordinary scientists recoil from the prior work of scientists embroiled in scandal, particularly if they have achieved great fame. That part of the punishment is carried out by giving less credit to the author's earlier work makes sense especially if some of the citations accruing to these scientists were "ceremonial" in nature. If principal investigators can control the likelihood of their team making a mistake or explicitly cheating, then this stigmatization (whether understood as a deterrent or as pure sociological mechanism à la Adut [2005]) could discourage scientific misconduct.

Scientist behavior. The pattern may be consistent with the distribution of outcomes in a richer *signal-jamming* model where rewards are solely based on the market's beliefs, but scientists can exert effort to reduce the likelihood of scientific misconduct. In the equilibrium of a model in that vein, very high-status agents exert low effort because they deem the probability of misconduct to be quite small.[19] Thus, even if misconduct were, in fact, more likely to occur for bad agents, the shares of mistakes and misconduct would be more similar for high- and low-status agents due to the effort component. This would yield patterns consistent with the cross-tabulation results in Table 5.

[19] See Board and Meyer-ter-Vehn (2013) or Bonatti and Hörner (2015) for models with this feature.

7 Concluding Remarks

The distribution of scientific recognition is a complex phenomenon. Disproportionate amounts of credit are given to the very best authors in a field (Merton 1968), but these authors must maintain their reputation at a high level through consistent performance. We have documented the scientific community's response to negative information shocks about a scientist's past output. The flow of credit (in the form of citations) responds to scandal (*i.e.*, retractions involving misconduct), all the more sharply when bad news involve an established member of the profession. Overall, the community's response is consistent with Bayesian learning under the assumptions that high-status scientists have a better initial reputation, and that misconduct is a more revealing signal, compared to an honest mistake.

In our current approach, we have taken the retraction-generating process as given. In other words, we do not attempt to construct and test a model of scientist behavior and market response to scandal, where the frequency and the consequences of a retraction are jointly determined in equilibrium. With endogenous effort choices, incorporating drivers of incentives such as punishment schemes and career concerns would enhance our understanding of the scientific reward system. The data currently available do not allow us to distinguish the effects of pure learning from those of more elaborate incentive schemes. However, developing empirical tests capable of adjudicating their relative salience is a valuable objective for future research in this area.

One limitation of looking at the retraction phenomenon through the prism of information revelation is that it sheds light on only a fraction of the private costs of false science — those narrowly associated with the prior work of the scientists embroiled in scandal. But these scientists bear additional costs in the form of foregone future funding, collaboration, and publication opportunities. Moreover, we cannot say anything definitive regarding the private *benefits* of fraud or sloppiness, because we only observe their consequences conditional on detection by the scientific community. Furman et al. (2012) have shown that retracted articles exhibit "excess" citations prior to retraction. Therefore, it is reasonable to infer that undetected instances of false science confer on their authors enhanced prestige, as well as privileged access to tangible resources, such as editorial goodwill, better trainees, or state-of-the-art laboratory equipment. These benefits are extremely difficult to assess without making a host of untestable assumptions.

Paradoxically, it is somewhat easier to quantify the externalities that retraction events generate on the vitality of the scientific fields they afflict. Azoulay et al. (2015) provide evidence that these external effects are quantitatively large, and that they arise mostly because scientists tend to stay away from these fields lest their own reputation suffers through mere intellectual proximity.

References

Adut, Ari. 2005. "A Theory of Scandal: Victorians, Homosexuality, and the Fall of Oscar Wilde." *American Journal of Sociology* **111**(1): 213-248.

Arikawa, Kentaro, Koichi Ozaki, Takanari Tsuda, Junko Kitamoto, and Yuji Mishina. 1996. "Retraction of paper: Two visual pigment opsins, one expressed in the dorsal region and another in the dorsal and the ventral regions, of the compound eye of a dragonfly, Sympetrum frequens." *Invertebrate Neuroscience* **2**(3): 209.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2014. "Matthew: Effect or Fable?" *Management Science* **60**(1): 92-109.

Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics,* Forthcoming.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* **125**(2): 549-589.

Board, Simon, and Moritz Meyer-ter-Vehn. 2013. "Reputation for Quality." *Econometrica* **81**(6): 2381-2462.

Bonatti, Alessandro, and Johannes Hörner. 2015. "Career Concerns with Exponential Learning." Working Paper, MIT and Yale.

Cabral, Luís M. B. 2005. "The Economics of Trust and Reputation: A Primer." Working Paper, New York University.

Cripps, Martin W., George J. Mailath, and Larry Samuelson. 2004. "Imperfect Monitoring and Impermanent Reputations." *Econometrica* **72**(2): 407-432.

Dasgupta, Partha, and Paul A. David. 1994. "Toward a New Economics of Science." *Research Policy* **23**(5): 487-521.

Fang, Ferric C., R. Grant Steen, and Arturo Casadevall. 2012. "Misconduct Accounts for the Majority of Retracted Scientific Publications." *Proceedings of the National Academy of Science* **109**(42): 17028-17033.

Fine, Gary Alan. 2001. *Difficult Reputations: Collective Memories of the Evil, Inept, and Controversial*. Chicago, IL: University of Chicago Press.

Furman, Jeffrey L., Kyle Jensen, and Fiona Murray. 2012. "Governing Knowledge in the Scientific Community: Exploring the Role of Retractions in Biomedicine." *Research Policy* **41**(2): 276-290.

Furman, Jeffrey L., and Scott Stern. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production." *American Economic Review* **101**(5): 1933-1963.

Glueck, Charles J., Margot J. Mellies, Mark Dine, Tammy Perry, and Peter Laskarzewski. 1986. "Safety and Efficacy of Long-Term Diet Plus Bile Acid-Binding Resin Cholesterol-Lowering Therapy in 73 Children Heterozygous for Familial Hypercholesterolemia." *Pediatrics* **78**(2): 338-348.

Goffman, Erving. 1963. *Stigma: Notes on the Management of Spoiled Identity*. New York, NY: Simon & Schuster.

Gouriéroux, Christian, Alain Montfort, and Alain Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models." *Econometrica* **53**(3): 701-720.

Holmström, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." *The Review of Economic Studies* **66**(1): 169-182.

Jin, Ginger Zhe, Benjamin Jones, Susan Feng Lu, and Brian Uzzi. 2013. "The Reverse Matthew Effect: Catastrophe and Consequence in Scientific Teams." NBER Working Paper #19489.

Lacetera, Nicola, and Lorenzo Zirulia. 2009. "The Economics of Scientific Misconduct." *The Journal of Law, Economics, & Organization* **27**(3): 568-603.

Lampe, Ryan. 2012. "Strategic Citation." *Review of Economics and Statistics* **94**(1): 320-333.

Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* **8**(423), doi:10.1186/1471-2105-8-423.

Merton, Robert K. 1957. "Priorities in Scientific Discovery: A Chapter in the Sociology of Science." *American Sociological Review* **22**(6): 635-659.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* **159**(3810): 56-63.

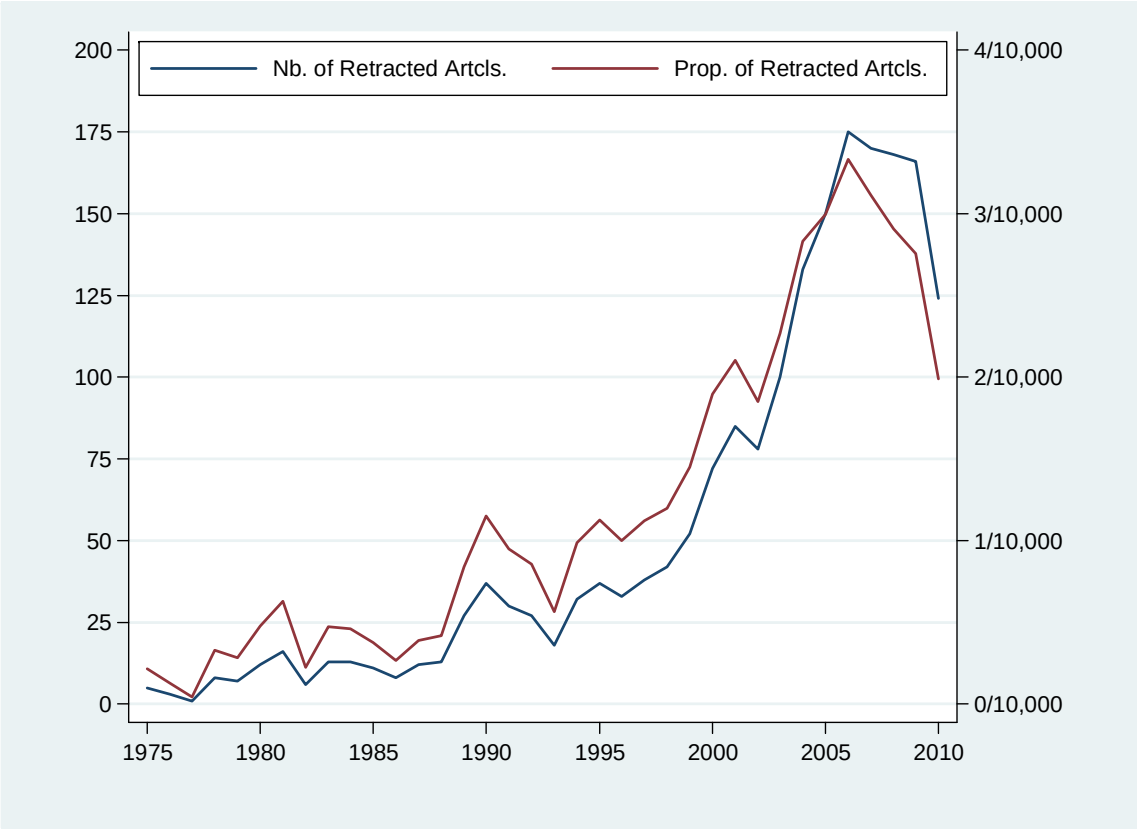
Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigation.* Chicago, IL: University of Chicago Press.

Van Noorden, Richard. 2011. "The Trouble with Retractions." *Nature* **478**(7367): 26-28.

Wade, Nicholas. 2010. "Inquiry on Harvard Lab Threatens Ripple Effect." *The New York Times*, August 12, 2010.

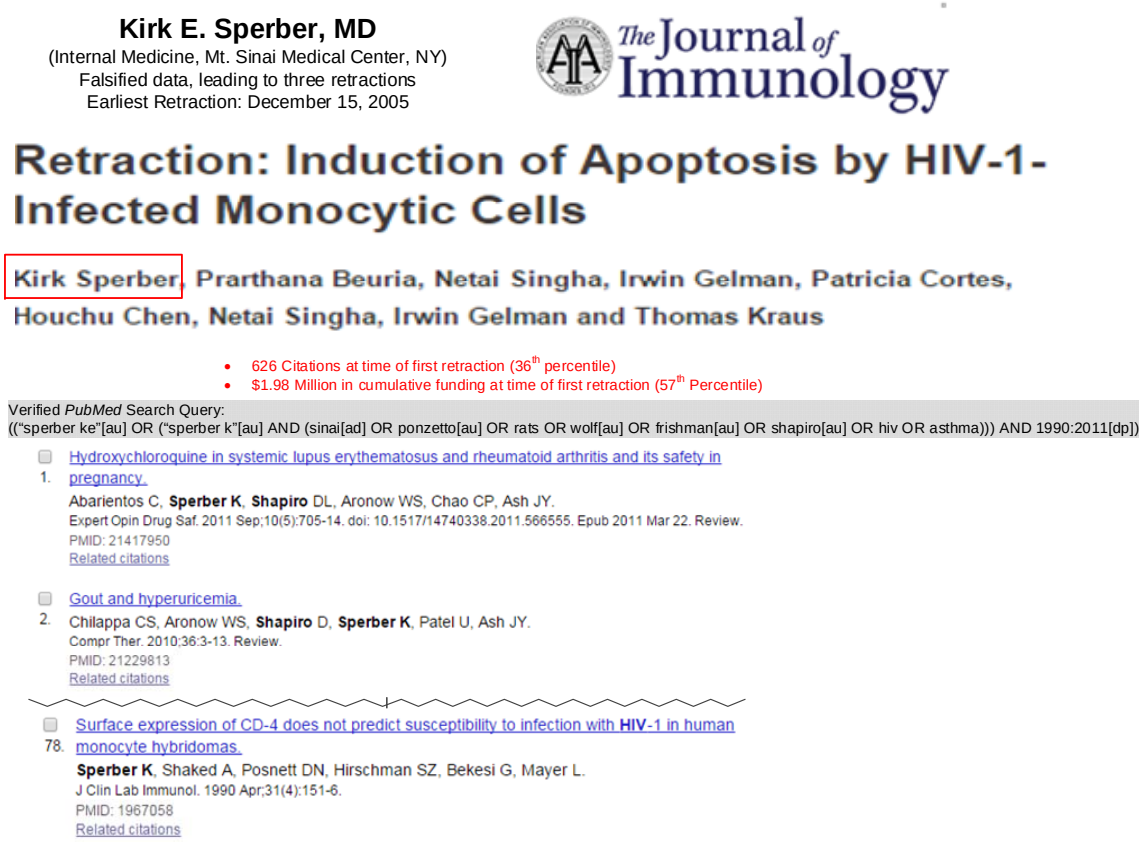
Wooldridge, Jeffrey M. 1997. "Quasi-Likelihood Methods for Count Data." In M. Hashem Pesaran and Peter Schmidt (Eds.), *Handbook of Applied Econometrics*, pp. 352-406. Oxford: Blackwell.

Figure 1: Incidence of PubMed-Indexed Retractions



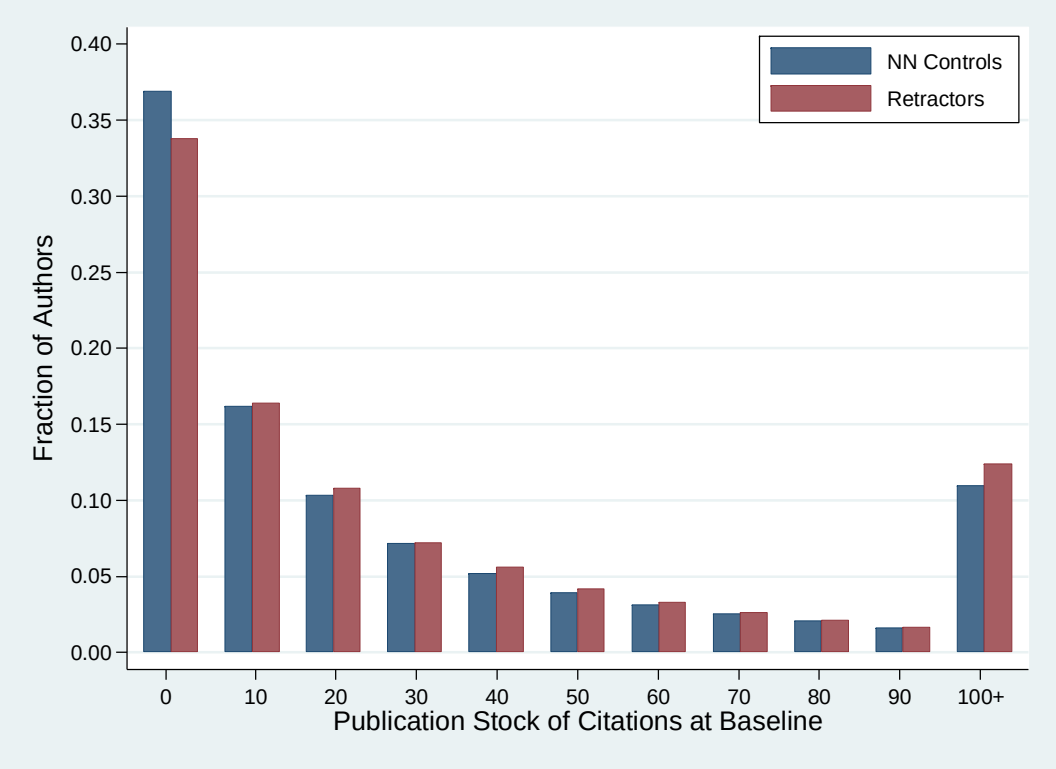
Note: The solid blue line displays the yearly frequency of retraction events in PubMed as a whole, all retraction reasons included. The solid red line displays the yearly retraction rate, where the denominator excludes PubMed-indexed articles that are not original journal articles (e.g., comments, editorials, reviews, etc.)

Figure 4: Matching Authors to their Bibliomes



Note: The example above illustrates the matching procedure employed to identify the career publication histories of faculty authors. In the example, Kirk Sperber is an author on three publications retracted due to fabricated data (he was later barred from receiving grants and contracts for four years by the Department of Health and Human Services' Office of Research Integrity). His earliest retraction came in the Journal of Immunology in December 2005. Our hand-curated *PubMed* query for Dr. Sperber utilizes common coauthors and research topics for his publications, as well as the relevant date range; it also addresses his lack of consistency in using a middle initial as an author. The query results in 78 publications, which we verified as his complete body of work. 60 of these articles were published prior to his earliest retraction (2005), and 7 publications were intellectually related (via the PMRA algorithm) to a retracted paper.

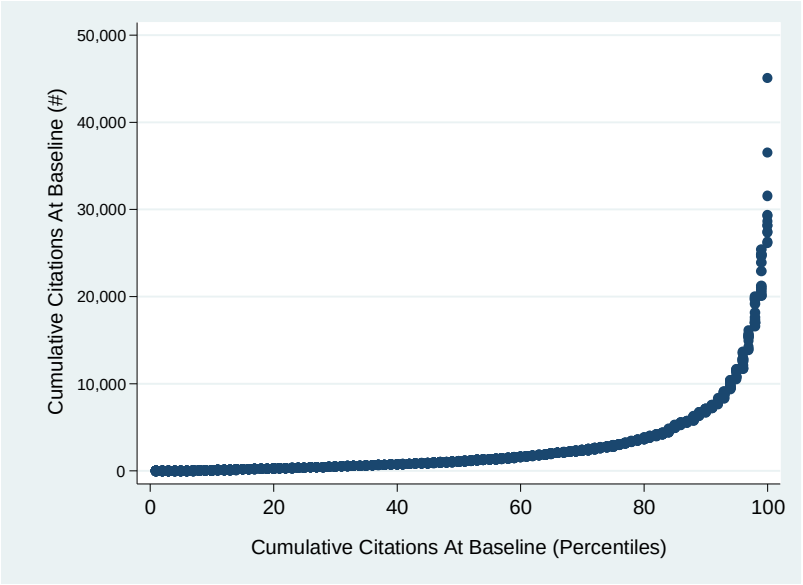
Figure 5: Cumulative Citations to Pre-Retraction Publications



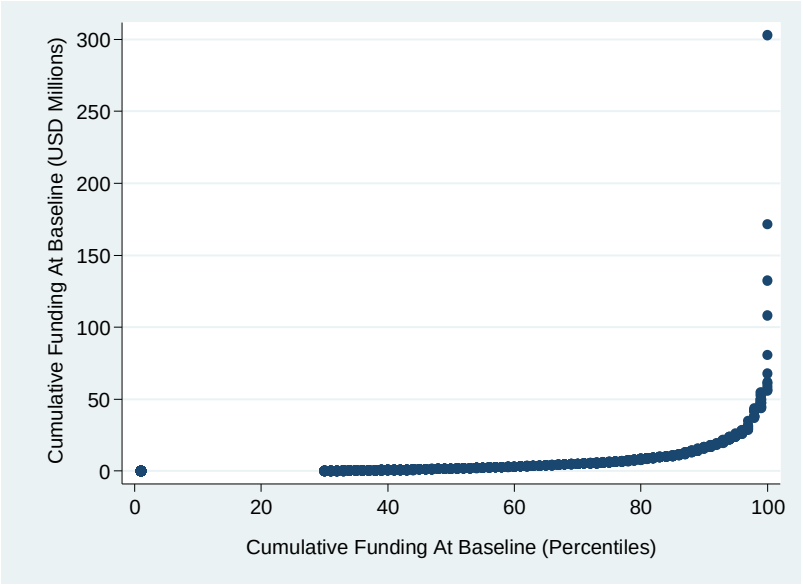
Note: Cumulative number of citations up to the year preceding the corresponding earliest retraction event for 21,103 retracted authors' publications and 42,115 control authors' publications.

Figure 6: Cumulative Citations and Funding at Baseline

A. Citations

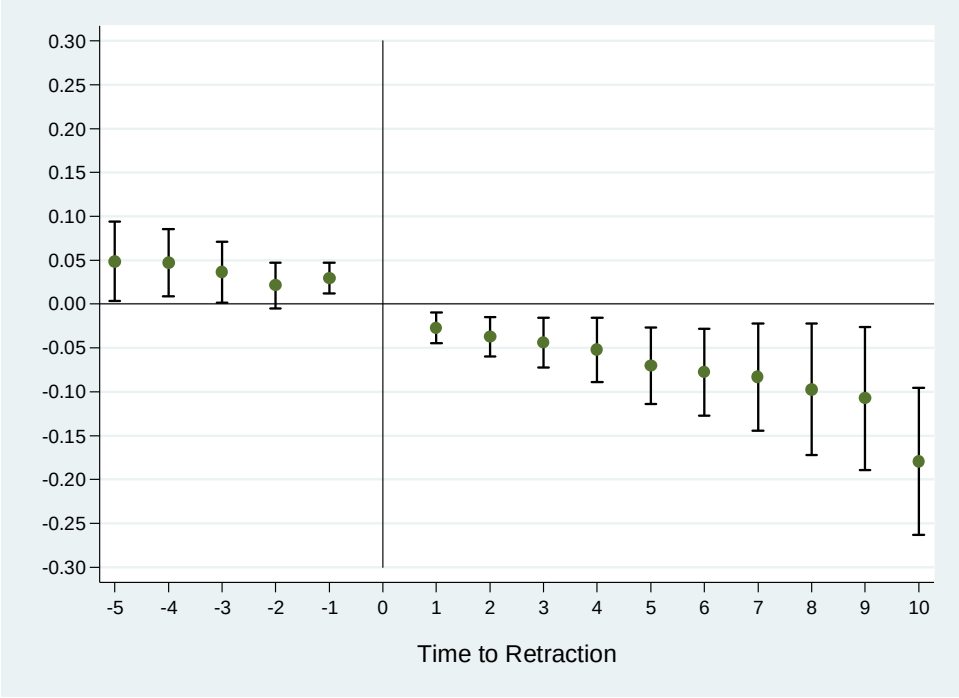


B. Funding



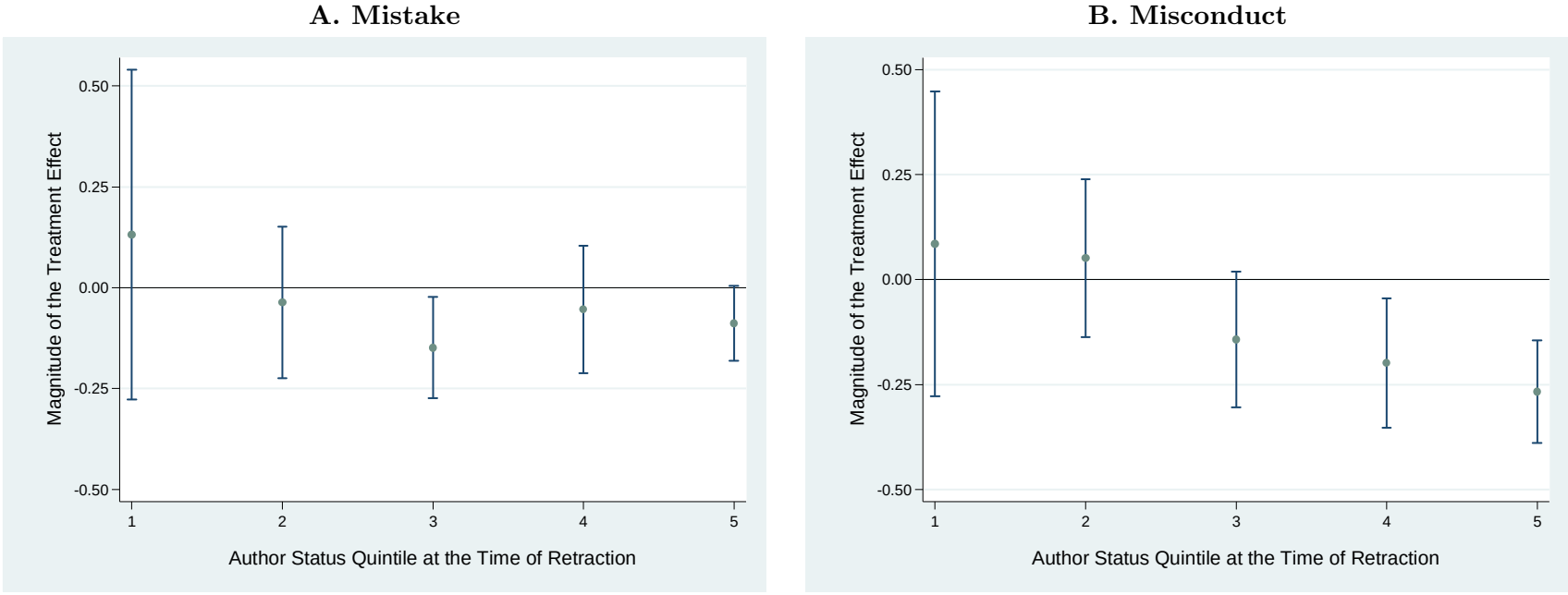
Note: We compute the cumulative number of citations (Panel A), and cumulative amount of funding (Panel B) up to the year preceding the corresponding earliest retraction event for all 376 retracted and 759 control authors, and plot it against 100 percentiles of the corresponding distribution.

Figure 7: Dynamics of Retraction Effect on Citations to Pre-Retraction Publications



Note: The green dots in the above plot correspond to coefficient estimates stemming from conditional fixed effects quasi-maximum likelihood Poisson specifications in which the citation rates for articles published by retracted and control authors prior to their first associated retraction event are regressed onto year effects, article age indicator variables, as well as 15 interaction terms between treatment status and the number of years before/after the retraction event (the indicator variable for treatment status interacted with the year of retraction is omitted). The 95% confidence intervals (corresponding to robust standard errors, clustered around unique author identifiers) around these estimates is plotted with the help of capped spikes.

Figure 8: Retraction Effect on Citations to Pre-Retraction Publications, by Status Quintile



Note: The green dots in the above plots correspond to coefficient estimates stemming from conditional fixed effects quasi-maximum likelihood Poisson specifications in which the citation rates for articles published by retracted and control authors prior to their first associated retraction event are regressed onto year effects, article age indicator variables, as well as five interaction terms between the treatment indicator variable and five indicator variables corresponding to each quintile of the author status distribution. Status is defined by the author's cumulative citations at baseline. Panel A limits the sample to publications of retracted authors and their controls who were associated with a retraction event stemming from mistake, while Panel B includes only publications for retracted authors and their controls that were associated with a retraction event stemming from misconduct. The 95% confidence intervals (corresponding to robust standard errors, clustered around unique author identifiers) around these estimates is plotted with the help of capped spikes.

Table 1: Baseline Descriptive Statistics for Retracted and Control Authors

	Mean	Std. Dev.	Median	Min	Max
Control Authors (n=759)					
Female	0.19	0.39	0	0	1
Degree Year	1975.13	10.85	1975	1941	1999
MD	0.41	0.49	0	0	1
PhD	0.48	0.50	0	0	1
MD/PhD	0.10	0.30	0	0	1
Misconduct – Earliest Associated Retraction	0.53	0.50	1	0	1
Blamed Author – Earliest Associated Retraction	0.00	0.00	0	0	0
Cumulative Citations	2,776	4,961.30	1,091	0	45,077
Cumulative Funding ($1000s)	5,971	15,884.90	1,362	0	302,862
Retracted Authors (n=376)					
Female	0.16	0.36	0	0	1
Degree Year	1976.95	10.64	1978	1938	1998
MD	0.39	0.49	0	0	1
PhD	0.49	0.50	0	0	1
MD/PhD	0.11	0.31	0	0	1
Misconduct – Earliest Associated Retraction	0.40	0.49	0	0	1
Blamed Author – Earliest Associated Retraction	0.06	0.24	0	0	1
Cumulative Citations	2,994	4,543	1,267	1	28,633
Cumulative Funding ($1000s)	6,373	12,619	2,191	0	132,403

Note: The set of 376 retracted authors consist of authors from 412 retracted papers for which we matched at least one author to the Faculty Roster of the Association of American Medical Colleges (AAMC). The 759 control authors are authors from adjacent articles in the same journal and issue as their retracted counterpart, and matched to the AAMC in the same fashion. The 12 retracted and 24 control authors who were NIH intramural researchers are excluded from the cumulative funding calculations, because their research funded through a very different system. The percentage of authors affiliated with misconduct cases for their earliest retractions is different between the two groups because the number of control authors varies by retraction case.

Table 2: Baseline Descriptive Statistics for Author-Publication Pairs

	Mean	Std. Dev.	Median	Min	Max
Control Author-Publications (n=46,538)					
Article Age (years)	9.34	7.42	8	0	32
First Author	0.16	0.37	0	0	1
Middle Author	0.41	0.49	0	0	1
Last Author	0.43	0.49	0	0	1
Cumulative Article Citations	42.25	136.29	16	0	18,301
Retracted Author-Publications (n=23,620)					
Article Age (years)	9.43	7.30	8	0	32
First Author	0.16	0.37	0	0	1
Middle Author	0.39	0.49	0	0	1
Last Author	0.44	0.50	0	0	1
Cumulative Article Citations	44.35	89.22	18	0	3,430

Note: The retracted- and control author/publications pairs in the sample correspond to articles published by the retracted and control authors prior to their affiliated retraction event. Authors were matched to their pre-retraction publications by developing *PubMed* search queries for each scientist using relevant keywords, names of frequent collaborators, journal names and institutional affiliations. The publication information for each paper, including publication date and authorship list, was gathered using the *PubHarvester* open source software tool [`http://www.stellman-greene.com/PublicationHarvester/`]. In biomedical journal publications, the last author is usually the primary investigator (lab director), and junior investigators (e.g. post-docs, graduate students, junior faculty) are listed as first or middle authors. Citation data was obtained through Thomson-Reuters' *Web of Science* (WoS) database, and we excluded all self-citations. We defined within-field and outside-field citations based on the *PubMed* Related Citations Algorithm (PMRA), such that citations from other publications within the same PMRA field were considered within-field. We removed all articles that were in the same PMRA field as original retracted or control publications.

Table 3: Citations to Pre-Retraction Articles, by Author Prominence and Misconduct

	Full Sample	Author Status				Retraction Type	
		Citations		Funding			
	(1)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
		High	Low	High	Low	Mistake	Misconduct
After Retraction	-0.113**	-0.100*	-0.105**	-0.070	-0.130**	-0.086*	-0.193**
	(0.033)	(0.042)	(0.039)	(0.050)	(0.046)	(0.038)	(0.053)
Nb. Authors	1,130	286	844	277	829	577	553
Nb. of Author-Publications	70,158	40,665	29,493	32,265	35,671	38,204	31,954
Nb. of Author-Paper-Year Obs.	1,736,319	979,230	757,089	802,765	878,238	888,557	847,762

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The unit of analysis is Author-Publication-Year, and the dependent variable is number of citations. All models incorporate a full suite of calendar year effects as well as indicator variables for the age of the publication and age of the author. High-status authors are those in the top quartile in terms of the author's cumulative citations (column 2a) or funding (column 3a) at baseline. Low-status authors are those in the bottom three quartiles of the same measures (columns 2b and 3b). NIH intramural scientists are excluded in the funding status models. The retraction type is defined by the earliest retraction for a given retracted author (controls retain the code associated with the retracted author with whom they are affiliated).

Exponentiating the coefficients and differencing from one yields estimates interpretable as elasticities. For example, the estimates in column (1a) imply that related articles suffer on average a statistically significant (1-exp[-0.113])=10.7% yearly decrease in citation rate after the retraction event. QML (robust) standard errors in parentheses, clustered around author identifiers. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 4: Citations to Pre-Retraction Articles, by Author Prominence and Misconduct Interactions

	Status: Citations				Status: Funding			
	High		Low		High		Low	
	(1a)	(1b)	(1c)	(1d)	(2a)	(2b)	(2c)	(2d)
	Mistake	Misconduct	Mistake	Misconduct	Mistake	Misconduct	Mistake	Misconduct
After Retraction	-0.079†	-0.191**	-0.097†	-0.100*	-0.027	-0.200*	-0.135*	-0.152*
	(0.047)	(0.066)	(0.051)	(0.051)	(0.055)	(0.079)	(0.055)	(0.070)
Nb. Authors	160	126	417	427	149	128	411	418
Nb. of Author-Publications	22,635	18,030	15,569	13,924	18,256	14,009	18,369	17,302
Nb. of Author-Paper-Year Obs.	528,115	451,115	360,442	396,647	439,066	363,699	410,869	467,369

Note: Estimates stem from conditional quasi-maximum likelihood Poisson specifications. The unit of analysis is Author-Publication-Year, and the dependent variable is number of citations. All models incorporate a full suite of calendar year effects as well as indicator variables for the age of the publication and age of the author. High-status authors are those in the top quartile in terms of the author's cumulative citations (columns 1a, 1b) or funding (columns 2a, 2b) at baseline. Low-status authors are those in the bottom three quartiles of the same measures (columns 1c, 1d, 2c, and 2d). NIH intramural scientists are excluded in the funding status models. The retraction type is defined by the earliest retraction for a given retracted author (controls retain the code associated with the retracted author with whom they are affiliated).

Exponentiating the coefficients and differencing from one yields estimates interpretable as elasticities. For example, the estimates in column (1b) imply that related articles suffer on average a statistically significant (1-exp[-0.191])=17.4% yearly decrease in citation rate after the retraction event. QML (robust) standard errors in parentheses, clustered around author identifiers. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$.

Table 5: Status and Retraction Type – Two-way Table of Frequencies

	Low Status	**High Status**	
Mistake	152 [58.02%]	68 [61.82%]	220
Misconduct	110 [41.98%]	42 [38.18%]	152
Total	262	110	

Note: Total author counts for each status group are displayed in the bottom row, while total author counts for each retraction type are displayed on the far right column. The percentages refer to the fraction of events in that cell for the corresponding column (status grouping). As can be readily observed, the ratio of misconduct and mistake retractions is about the same for both high and low-status authors.

Appendix A: Model Extensions

Two Signals

Consider a model where retractions can occur due to two different processes. In particular, a "mistake" retraction follows a Poisson process with parameter λ_θ, and a "misconduct " retraction arrives according to an independent process with parameter μ_θ. When a retraction event occurs, its type is publicly observed.

Because information arrives continuously to the market, when a retraction occurs, the drop in the agent's reputation depends on the probability distribution of that retraction type only. Therefore, let

$$\beta \triangleq \frac{\mu_B}{\mu_G} > 1$$

denote the relative informativeness of the misconduct signal. The resulting drop in reputation is given by $\Delta(p, \alpha)$ following a mistake event and by $\Delta(p, \beta)$ following a misconduct event.

We assume that the misconduct signal is more informative of the agent's low quality, i.e. $\beta > \alpha$. Our earlier Proposition 1 states that reputations suffer a larger drop following a retraction due to misconduct than after a mistake.

Finally, Bayesian updating and rational expectations have testable implications for the distribution of retractions in a population of high-and low-reputation agents. In particular, if the market holds correct beliefs p_t at each point in time, the arrival rate of a retraction for agents with reputation p is given by

$$p(\lambda_G + \mu_B) + (1 - p)(\lambda_B + \mu_B).$$

It then follows that the distribution of retractions of different kinds is related to the current reputation level of an agent.

Proposition 2 (Relative Frequency) *The fraction of misconduct events is decreasing in the agent's reputation p.*

Similarly, the distribution of retracted authors' reputations for each kind of retraction should differ in a systematic way: high-reputation agents should be relatively more frequent among authors with a retraction due to mistake.

Aggregate Implications

We define the average reputation drop for agents with initial levels of reputation $p \in [0, p^*]$ and $p \in [p^*, 1]$ respectively as follows:

$$\begin{aligned} L(p^*, \alpha) &= \frac{1}{p^*} \int_0^{p^*} \Delta(p, \alpha)\, \mathrm{d}p \\ H(p^*, \alpha) &= \frac{1}{1 - p^*} \int_{p^*}^1 \Delta(p, \alpha)\, \mathrm{d}p. \end{aligned}$$

We study the gap in reputation losses as a function of signal informativeness α, which we define as $|H(p^*, \alpha)| - |L(p^*, \alpha)|$. We then have the following result.

Proposition 3 (Critical Partition) *For each α, there exists a critical p such that the gap in reputation losses is increasing in α for all $p^* \leq p$.*

Changing Types

Suppose the agent's type follows a continuous-time Markov chain with transition rate matrix

$$Q = \begin{bmatrix} -\gamma & \gamma \\ \beta & -\beta \end{bmatrix}.$$

That is, it switches from good to bad at rate γ and from bad to good at rate β. The probability matrix $P(t)$ with entries $P_{ij} = \Pr[\theta_t = j \mid \theta_0 = i]$ is given by

$$P(t) = \begin{bmatrix} \frac{\beta}{\gamma+\beta} + \frac{\gamma}{\gamma+\beta}e^{-(\gamma+\beta)t} & \frac{\gamma}{\gamma+\beta} - \frac{\gamma}{\gamma+\beta}e^{-(\gamma+\beta)t} \\ \frac{\beta}{\gamma+\beta} - \frac{\beta}{\gamma+\beta}e^{-(\gamma+\beta)t} & \frac{\gamma}{\gamma+\beta} + \frac{\beta}{\gamma+\beta}e^{-(\gamma+\beta)t} \end{bmatrix}.$$

This means intuitively that the effect of—even arbitrarily precise—signals fades away as time passes (because the underlying fundamental is likely to have changed).

In our context, we can compute this "depreciation effect" backwards. In particular, if the market assigns probability p to $\theta = \theta_G$ after a retraction, it will assign probability

$$\pi(p,t) = p\left(\frac{\beta}{\gamma+\beta} + \frac{\gamma}{\gamma+\beta}e^{-(\gamma+\beta)t}\right) + (1-p)\left(\frac{\beta}{\gamma+\beta} - \frac{\beta}{\gamma+\beta}e^{-(\gamma+\beta)t}\right)$$

to the agent's type being good t periods ago. While $\pi(p,t)$ will be increasing or decreasing depending on the comparison of p with its long-run mean, it will always move in the direction of dampening the most recent change, *i.e.*, the retraction.

Proofs

Proof of Proposition 1 (1.) Differentiating $\Delta(p,\alpha)$ with respect to α yields

$$\frac{\partial \Delta(p,\alpha)}{\partial \alpha} = -\frac{p(1-p)}{(p+(1-p)\alpha)^2} < 0.$$

(2.) Similarly, the cross-partial is given by

$$\frac{\partial^2 \Delta(p,\alpha)}{\partial \alpha \partial p} = \frac{p - \alpha(1-p)}{(p+(1-p)\alpha)^3},$$

which is negative over the range $p \in \left[0, 1/\left(1+\alpha^{-1}\right)\right]$. □

Proof of Proposition 2 The relative frequency of misconduct events is given by

$$\frac{p\mu_G + (1-p)\mu_B}{p(\lambda_G+\mu_B) + (1-p)(\lambda_B+\mu_B)},$$

whose derivative is

$$\frac{\lambda_B\mu_G - \mu_B\lambda_G}{(p(\lambda_G+\mu_B) + (1-p)(\lambda_B+\mu_B))^2},$$

which is *negative* because $\alpha < \beta$. □

Proof of Proposition 3 Use the definition of $\Delta(p,\alpha)$ given in (1) to compute $|H(p^*,\alpha)| - |L(p^*,\alpha)|$. The result then follows directly. □

Appendix B: Author Matching

This appendix describes the method used to match retraction and control article authors to the augmented Association of American Medical Colleges (AAMC) Faculty Roster (cf. Azoulay et al. [2010] for more details on the AAMC Faculty Roster). Our process involved two main steps, using different pieces of available information about authors, publications, and grants. We have checked that our matching criteria of both steps is reliable and conservative, such that we are very confident in the accuracy of our final set of matched authors.

As a first step, we matched all authors for whom we already had a confirmed AAMC Faculty Roster match and full career publication histories from prior work (see Azoulay et al. 2012). We determined this set of pre-matched authors by identifying any relevant source publications (retracted or control articles) in the validated career publications for our set of previously matched authors.

For the remaining unmatched retraction or control authors, we undertook an iterative process to determine accurate matches in the augmented AAMC Faculty Roster. As a first pass, we identified potential matches using author names, and confirmed and matched those with only one possible match. For those with common names or multiple potential name matches, we used additional observable characteristics such as institution, department, and degree to remove erroneous potential matches. When multiple potential matches remained, we compared the topic area of the retracted/control paper to the grant titles, *PubMed* publication titles and abstracts associated with author name and the AAMC Faculty Roster entry. In these cases, we only declared a match when the additional information made the choice clear.

Appendix C: Linking Scientists with their Journal Articles

The next step in data construction is to link each matched author to their publications. The source of our publication data is *PubMed*, a bibliographic database maintained by the U.S. National Library of Medicine that is searchable on the web at no cost.[i] *PubMed* contains over 24.6 million citations from 23,000 journals published in the United States and more than 70 other countries from 1966 to the present. The subject scope of this database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. In order to effectively mine this publicly-available data source, we used PubHarvester[ii], an open-source software tool that automates the process of gathering publication information for individual life scientists (see Azoulay et al. 2006 for a complete description of the software). PubHarvester is fast, simple to use, and reliable. Its output consists of a series of reports that can be easily imported by statistical software packages.

This software tool does not obviate the two challenges faced by empirical researchers when attempting to link accurately individual scientists with their published output. The first relates to what one might term "Type I Error," whereby we mistakenly attribute to a scientist a journal article actually authored by a namesake; The second relates to "Type II error," whereby we conservatively exclude from a scientist's bibliome legitimate articles:

Namesakes and popular names. *PubMed* does not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials, and an optional suffix. This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Inconsistent publication names. The opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffixes (Jr., Sr., 2nd, etc.), or from multiple patronyms due to changes in spousal status.

[i] `http://www.pubmed.gov/`

[ii] The software can be downloaded at `http://www.stellman-greene.com/PublicationHarvester/`

To deal with these measurement problems, we opted for a labor-intensive approach: the design of individual search queries that relies on relevant scientific keywords, the names of frequent collaborators, journal names, as well as institutional affiliations. We are aided in the time-consuming process of query design by the availability of a reliable archival data source, namely, these scientists' CVs and biosketches. PubHarvester provides the option to use such custom queries in lieu of a completely generic query (e.g, `"azoulay p"[au]` or `"krieger jl"[au]`). For authors with uncommon names and distinct areas of study, a customized query may simply require a name and date range. For example, scientist Wilfred A. van der Donk required a simple *PubMed* search query: `("van der donk wa"[au] AND 1989:2012[dp])`. On the other hand, more common names required very detailed queries that focus on coauthor patterns, topics of research, and institution locations. An example of this type of detailed query is that of author John L. Cleveland in our data: `(("cleveland jl"[au] OR ("cleveland j" AND (rapp or hiebert))) NOT (oral OR diabetes OR disease[ad]) AND 1985:2012[dp])`.

As an additional tool, we also employed the Author Identifier feature of Elsevier's Scopus database to help link authors to their correct publication histories. This feature assigns author identification numbers using names, name variants, institutional affiliations, addresses, subject areas, publication titles, publication dates and coauthor networks.[iii] We compared the publication histories compiled by the Scopus system to our our detailed *PubMed* queries and found greater than 90% concordance, and extremely few "Type I" errors in either system. Our systematic comparisons led us to believe that the Scopus system provides an accurate set of career publications.

Appendix D: Measuring Misconduct and Blame

In order to distinguish between instances of misconduct and instances of "honest mistakes," we relied on the coding scheme developed in Azoulay et al. (2015). These authors developed a procedure to capture whether intentional deception was involved in the events that led to a specific article being retracted. They investigated each retraction by sifting through publicly available information, ranging from the retraction notice itself, Google searches, the news media, and blog entries in *RetractionWatch*.

The "intent" coding scheme divide retractions into three categories :

1. ***No Sign of Intentional Deception*** for cases where the authors did not appear to intentionally deceive the audience (i.e., "honest mistakes").
2. ***Uncertain Intent*** when negligence or unsubstantiated claims were present, but an investigation of the public documents did not hint at malice on the authors' part.
3. ***Intentional Deception*** is reserved for retractions due to falsification, intentional misconduct, or willful acts of plagiarism.

There is of course an element of subjectivity in the assignment of these codes, but the third category can be distinguished from the first two unambiguously.[iv]

For the empirical exercise performed in this manuscript, we lumped the "No Sign of Intentional Deception" and "Uncertain Intent" categories into a single "honest mistake" grouping. This coding choice ensures that retracted authors associated with a misconduct retraction have been linked unambiguously to a case of intentional deception. In robustness checks, we also replicated the results presented in Table 4 while (a) lumping the uncertain cases with the clear-cut cases of misconduct; and (b) dropping from the sample all the retractions that belong to the "uncertain Intent" category. These tweaks had an impact on the precision of some of the estimates presented in Table 5, but did not change its take-away message.

[iii] described at `http://help.scopus.com/Content/h_autsrch_intro.htm`

[iv] The codes for each retraction, together with a rationale for the category chosen, can be downloaded at `http://jkrieger.scripts.mit.edu/retractions_index.html`.

We evaluated the assignment of blame among the authors of each retracted publication, and coded which authors were deemed at-fault for the events that led to retraction. On occasion, the retraction notice singles out particular authors. In other cases, the notice itself might be silent on the topic of blame, but other publicly available sources of information (e.g., newspaper articles, press releases, blog posts, ORI investigation reports) enable us to pinpoint the individual deemed responsible. Additionally, authors are occasionally blamed by omission, such as when an author name is conspicuously absent from a series of retractions or related documents, or the retracted publication has a sole author.

In the full sample of 1,129 retractions, 565 had at least one "blameworthy" author according to our definition. However, the majority of blamed authors are precisely the kinds of scientists less likely to ever appear in the AAMC Faculty Roster (e.g. graduate students, postdoctoral fellows, and technicians). Only 24 out of the 376 retracted authors we could match to the AAMC Faculty Roster qualified as blameworthy using the criteria above.

Appendix E: In-Field and Out-of-Field Publications

This appendix describes our method of identifying "related" publications for all of the retracted/control publications in our sample. In the econometric analyses, we separated publications that were in the same line of scientific inquiry as the retracted or control source article. We treated these closely related papers separately because in prior work (Azoulay et al., 2015), we found that papers in the same field as a retraction experience citation declines due to their intellectual association with the retracted piece. Therefore, we wanted to remove such papers to avoid contaminating our measurement of individual reputation effects with the field-level effects found in this prior work. Furthermore, by identifying the entire set of related papers, we can also differentiate between citations coming from within vs. outside a particular field.

The data challenge in the paper is to separate, in the body of published work for a given scientist that predates a retraction, the set of articles that belong to the same narrow intellectual subfield as the retraction from the set of articles that lies outside the retracted article's narrow subfield. This challenge is met by the use of the PubMed Related Citations Algorithm [PMRA], a probabilistic, topic-based model for content similarity that underlies the "related articles" search feature in PubMed. This database feature is designed to aid a typical user search through the literature by presenting a set of records topically related to any article returned by a PubMed search query.[v] To assess the degree of intellectual similarity between any two PubMed records, PMRA relies crucially on MeSH keywords. MeSH is the National Library of Medicine's [NLM] controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. There are 27,149 descriptors in the 2013 MeSH edition. Almost every publication in PubMed is tagged with a set of MeSH terms (between 1 and 103 in the current edition of PubMed, with both the mean and median approximately equal to 11). NLM's professional indexers are trained to select indexing terms from MeSH according to a specific protocol, and consider each article in the context of the entire collection (Bachrach and Charen 1978; Névéol et al. 2010). What is key for our purposes is that the subjectivity inherent in any indexing task is confined to the MeSH term assignment process and does not involve the articles' authors.

Using the MeSH keywords as input, PMRA essentially defines a distance concept in idea space such that the proximity between a source article and any other PubMed-indexed publication can be assessed. The algorithm focuses on the smallest neighborhood in this space that includes 100 related records.[vi] The following paragraphs were extracted from a brief description of PMRA:

> *The neighbors of a document are those documents in the database that are the most similar to it. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths.*

[v] Lin and Wilbur (2007) report that one fifth of "non-trivial" browser sessions in PubMed involve at least one invocation of PMRA.

[vi] However, the algorithm embodies a transitivity rule as well as a minimum distance cutoff rule, such that the effective number of related articles returned by PMRA varies between 58 and 2,097 in the larger sample of 3,071 source articles published by the 451 star scientists in the five years preceding their death. The mean is 185 related articles, and the median 141.

To carry out such a program, one must first define what a word is. For us, a word is basically an unbroken string of letters and numerals with at least one letter of the alphabet in it. Words end at hyphens, spaces, new lines, and punctuation. A list of 310 common, but uninformative, words (also known as stopwords) are eliminated from processing at this stage. Next, a limited amount of stemming of words is done, but no thesaurus is used in processing. Words from the abstract of a document are classified as text words. Words from titles are also classified as text words, but words from titles are added in a second time to give them a small advantage in the local weighting scheme. MeSH terms are placed in a third category, and a MeSH term with a subheading qualifier is entered twice, once without the qualifier and once with it. If a MeSH term is starred (indicating a major concept in a document), the star is ignored. These three categories of words (or phrases in the case of MeSH) comprise the representation of a document. No other fields, such as Author or Journal, enter into the calculations.

Having obtained the set of terms that represent each document, the next step is to recognize that not all words are of equal value. Each time a word is used, it is assigned a numerical weight. This numerical weight is based on information that the computer can obtain by automatic processing. Automatic processing is important because the number of different terms that have to be assigned weights is close to two million for this system. The weight or value of a term is dependent on three types of information: 1) the number of different documents in the database that contain the term; 2) the number of times the term occurs in a particular document; and 3) the number of term occurrences in the document. The first of these pieces of information is used to produce a number called the global weight of the term. The global weight is used in weighting the term throughout the database. The second and third pieces of information pertain only to a particular document and are used to produce a number called the local weight of the term in that specific document. When a word occurs in two documents, its weight is computed as the product of the global weight times the two local weights (one pertaining to each of the documents).

The global weight of a term is greater for the less frequent terms. This is reasonable because the presence of a term that occurred in most of the documents would really tell one very little about a document. On the other hand, a term that occurred in only 100 documents of one million would be very helpful in limiting the set of documents of interest. A word that occurred in only 10 documents is likely to be even more informative and will receive an even higher weight.

The local weight of a term is the measure of its importance in a particular document. Generally, the more frequent a term is within a document, the more important it is in representing the content of that document. However, this relationship is saturating, i.e., as the frequency continues to go up, the importance of the word increases less rapidly and finally comes to a finite limit. In addition, we do not want a longer document to be considered more important just because it is longer; therefore, a length correction is applied.

The similarity between two documents is computed by adding up the weights of all of the terms the two documents have in common. Once the similarity score of a document in relation to each of the other documents in the database has been computed, that document's neighbors are identified as the most similar (highest scoring) documents found. These closely related documents are pre-computed for each document in PubMed so that when one selects Related Articles, the system has only to retrieve this list. This enables a fast response time for such queries.[vii]

To summarize, PMRA is a modern implementation of *co-word analysis*, a content analysis technique that uses patterns of co-occurrence of pairs of items (i.e., title words or phrases, or keywords) in a corpus of texts to identify the relationships between ideas within the subject areas presented in these text (Callon et al. 1989; He 1999). One long-standing concern among practitioners of this technique has been the "indexer effect" (Whittaker 1989). Clustering algorithm such as PMRA assume that the scientific corpus has been correctly indexed. But what if the indexers who chose the keywords brought their own "conceptual baggage" to the indexing task, so that the pictures that emerge from this process are more akin to their conceptualization than to those of the scientists whose work it was intended to study?

Indexer effects could manifest themselves in three distinct ways. First, indexers may have available a lexicon of permitted keywords which is itself out of date. Second, there is an inevitable delay between the publication of an article and the appearance of an entry in PubMed. Third, indexers, in their efforts to be helpful to users of the database, may use combinations of keywords which reflect the conventional views of the field. The first two concerns are legitimate, but probably have only a limited impact on the accuracy of the relationships between articles which PMRA deems related. This is because the NLM continually revises and updates the MeSH vocabulary, precisely in an attempt to neutralize keyword vintage effects. Moreover, the time elapsed between an article's publication and the indexing task has shrunk dramatically, though time lag issues might have been a first-order challenge when MeSH

[vii] Available at `http://ii.nlm.nih.gov/MTI/related.shtml`

was crated, back in 1963. The last concern strikes us as being potentially more serious; a few studies have asked authors to validate ex post the quality of the keywords selected by independent indexers, with generally encouraging results (Law and Whittaker 1992). Inter-indexer reliability is also very high (Wilbur 1998).

References

Azoulay, Pierre, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics* Forthcoming.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* **125**(2): 549-589.

Azoulay, Pierre, Andrew Stellman, and Joshua Graff Zivin. 2006. "PublicationHarvester: An Open-source Software Tool for Science Policy Research." *Research Policy* **35**(7): 970-974.

Azoulay, Pierre, Joshua S. Graff Zivin, and Bhaven N. Sampat. 2012. "The Diffusion of Scientific Knowledge Across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine." In Josh Lerner, and Scott Stern (Eds.), *The Rate & Direction of Inventive Activity Revisited*, pp. 107-155. Chicago, IL: University of Chicago Press.

Bachrach, C. A., and Thelma Charen. 1978. "Selection of MEDLINE Contents, the Development of its Thesaurus, and the Indexing Process." *Medical Informatics (London)* **3**(3): 237-254.

Callon, Michel, Jean-Philippe Courtial, and F. Laville. 1991. "Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry." *Scientometrics* **22**(1): 155-205.

He, Qin. 1999. "Knowledge Discovery Through Co-Word Analysis." *Library Trends* **48**(1): 133-159.

Law, John, and John Whittaker. 1992. "Mapping Acidification Research: A Test of the Co-word Method." *Scientometrics* **23**(3): 417-461.

Lin, Jimmy, and W. John Wilbur. 2007. "PubMed Related Articles: A Probabilistic Topic-based Model for Content Similarity." *BMC Bioinformatics* **8**(423).

Whittaker, John. 1989. "Creativity and Conformity in Science: Titles, Keywords and Co-Word Analysis." *Social Studies of Science* **19**(3): 473-496.

Wilbur, W. John. 1998. "The Knowledge in Multiple Human Relevance Judgments." *ACM Transactions on Information Systems* **16**(2): 101-126.