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
Harvard Business School
jkrieger@hbs.edu

July 7, 2017

Abstract

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 earlier, non-retracted 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 cause of the retraction (fraud vs. mistake) 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 its source in "honest mistakes," we find no evidence of differential stigma between high- and low-status faculty members.

Keywords: retractions, fraud, reputation, scandal, status, scientists

*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 At the time the article was retracted, Dr.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.

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 mechanisms through which the mighty are often brought low. The consequences of scandalous revelations are especially important in the scientific community, where reputation functions like a currency (Dasgupta & David 1994). However, the efficiency of the scientific reward system is predicated upon the community's ability to separate truth from falsehood, to strike inaccuracies from the scientific record, and to dole out reputational punishment in the wake of errors or misconduct (Budd et al. 1998; Lacetera & Zirulia 2011; Fang et al. 2012; Furman et al. 2012; Azoulay et al. 2015). Which scientists are most vulnerable to these punishments? How does the nature of an infraction and the prominence of the scientist moderate the effect of scandal on scientific reputation? Because scandal is at its core an informational phenomenon, we study the professional fate of scientists whose transgressions are suddenly publicized—to paraphrase the succinct definition of scandal provided by Adut (2005).

The reigning theoretical paradigm to assess the effects of the revelation of information is Bayesian updating. When the "market" (the scientific community) observes the release of negative information, it might infer that the agent (in this case, a scientist) was mediocre all along, therefore discounting the work that he produced in the past. In line with this paradigm, we develop a theoretical model that incorporates two key factors in the community's assessment of a scandal: (i) the agent's prominence at the time of the negative revelation, and (ii) the informational content of the disclosure itself. Our model predicts

that more prominent scientists will suffer greater reputation loss than less prominent authors following disclosures of misconduct, but not following disclosures of "honest mistakes."

To address these issues empirically, we turn to the setting of scientific retractions. We start from a list of 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 faculty with at least one retracted publication (retracted authors) for whom we assemble a curated history of publications, NIH grants, and citations. Our novel multi-level panel dataset links individual faculty members associated with retraction events together with their prior, un-retracted publication output. We proceed in a symmetric fashion to produce a sample of articles linked to 759 control authors who were not embroiled in retraction scandals, but published articles in the same journals where the retraction events occurred.[1]

Armed with these data, we 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. Analyzing citations to prior work, rather than citations to articles published after the retraction event, is a key feature of our empirical strategy. Following a negative reputation shock, scientists might adjust their level of effort or face new production constraints (*e.g.*, reduced funding opportunities). Focusing on unretracted prior work allows us to attribute any shift in citation patterns to the negative reputation shock, rather than to changes in production inputs.

This type of analysis may, however, confound any citation penalty suffered by a specific retracted author with the broader consequences of the scientific community abandoning a research field altogether. Significant spillover effects of retractions on the evolution of research fields were documented by Azoulay et al. (2015), who examined the impact of retractions on the citation of papers in the same field by non-overlapping authors. In order to isolate the effects of retractions on individuals' reputations and avoid the field-level spillover effects, we focus exclusively on publications by the retracted authors in a different research subfield than the retracted paper. Having filtered out the research field-specific effects, we find that the pre-retraction work of retracted authors suffers a 10% average annual citation

[1]We focus on faculty members and exclude technicians, graduate students and postdocs in order to avoid confounding differences in prominence with differences in career stage.

penalty following a retraction event, relative to the fate of the articles published by non-retracted control authors.

We then investigate the impact of the authors' reputation at the time of the retraction (whether they belonged to the top quartile of the citation or funding distribution) and of the reasons for the retraction by carefully separating instances of misconduct (including fraud and plagiarism) from instances of mistakes (stemming, for example, from contaminated biological samples or statistical errors). Our results indicate that the cause of the retraction (mistake vs. misconduct) and the scientist's prior reputation interact in very specific ways to shape the magnitude of the community'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%.

Jointly, these results show that the penalty levied by the scientific community on a retracted author matches the response of a Bayesian decision maker who holds prior beliefs correlated with the author's prominence in the profession and perceives misconduct cases as more informative signals than honest mistakes. To then assess how well the market is able to parse the "truth" in signals of varying informativeness, we circle back to the joint distribution of author reputations and retraction events. Consistent with the scientific community's beliefs, we find that prior reputation levels are negatively correlated with the incidence of retractions (as it should be if reputation is informative of the true quality of a scientist). Surprisingly, however, cases of misconduct are not relatively more prevalent among low-reputation authors and should not, therefore, carry statistical information. Among possible explanations, this discrepancy in the market's reaction may suggest either an information-processing problem (*i.e.*, the market is unable to filter truth from noise), or an information-acquisition problem (*i.e.*, misconduct cases involving famous authors are much more publicized than all others).

Our study is related to a recent paper by Jin et al. (2017). 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. In other words, for each retracted article, we usually focus on a single author, typically the principal investigator. In contrast, Jin et al. (2017) compare the citation trajectories of scientists who appeared on the authorship roster of the same retracted publication.[2] Additionally, our study directly investigates how the type of retraction signal (mistake vs. misconduct) moderates reputation penalties, while Jin et al. (2017) aim to remove such variation by discarding self-reported errors from their sample of retraction events.

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 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 revisits 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 institutions 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.

[2]These authors might be graduate students, postdoctoral fellows, staff scientists, or heads of laboratory, though they cannot be separately identified within the constraints of the Jin et al. (2017) empirical exercise. In contrast, we have gathered extensive information about the scientists in our sample, such as demographic characteristics and past productivity. At the time of the retraction event, all of the scientists in our sample are faculty members in a U.S. Medical School.

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 *Retraction-Watch*, 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.

3 Theoretical Framework

Reputation is a canonical concept across the social sciences. Within the economics literature, the existing theoretical research is concerned with how, and under what conditions, economic agents acquire a good (or a bad) one, and more generally, how they influence the beliefs of the market about their innate type (Mailath & Samuelson 2006). In contrast, few empirical studies document the events that can lead actors to lose their reputation, or quantify the consequences of this loss. Some notable exceptions include studies that assess how product recalls (Jarrell & Peltzman 1985), product liability lawsuits (Prince & Rubin 2002), and medical malpractice (Dranove et al. 2012) impact the producer's subsequent market valuation and demand. Egan et al. (2016) provides evidence of how financial advisors' misconduct records impact their careers and unemployment experiences. The corporate finance and accounting literature also addresses the career effects of financial fraud by evaluating how financial fraud and earnings restatements impact the reputations of board members (Srinivasan 2005; Fich & Shivdasani 2007). However, reputation loss occupies a more central place in sociology (Goffman 1963; Fine 2001).

Our theoretical framework formalizes the dynamics of reputation in a model of Bayesian learning. In particular, we explore how scandal impacts individual reputations, as a function of the informational content of the scandalous revelation and of the prominence of the individual scientists involved. The model yields insights that guide the interpretation of the empirical results.

We begin with a single, representative researcher (the *agent*) who is continuously evaluated by the scientific community (the *market*). The agent has a fixed binary characteristic that denotes his *reliability*

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

Thus, the agent is either good or bad. We let $p_0 \triangleq \Pr(\theta = \theta_G)$ denote the market's prior belief that the agent is of the good type.[3]

The agent's output at each point in time is also binary,

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

[3]In practice, a researcher's quality is definitely multi-dimensional, and the shocks we observe (e.g., a retraction due to mistake or misconduct) lead to updating about different dimensions (e.g., the author's carefulness or honesty). Ultimately though, we are interested in the market's ability to trust the author's results. Put differently, reliability requires the author to be both careful and honest, so that every retraction leads to negative updating about his quality.

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 learns about the agent's reliability from observing his scientific output, and rewards the agent with citations based on his reputation. Let p_t denote the market's posterior belief that the agent is good, conditional on the output produced up to that time.[4] 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. In other words, the citations received by the agent's body of work are a function of the market's belief that his reliability is high, based on his output history. Rewards for reputation are highly nonlinear in our database (see Figure 7), where the distribution of citations is heavily skewed towards "superstar" agents.[5]

3.1 Learning and Reputation

The market learns about the agent's reliability through retractions that we model as a Poisson process. The intensity of the Poisson process is higher for low-reliability (bad) agents.[6] Thus, retractions are rare, publicly observable events that reveal information about an agent's reliability. As our interest lies in the comparison across rather than within retracted papers, we assume that a retraction is an equally informative signal for every identifiable author of a retracted paper. The consequences of this signal, however, vary with each author's prior reputation, as described below.

More formally, 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 at each time t is measured by the market's belief p_t. Figure 2 illustrates the dynamics of reputation through a sample path generated by our model.

As Figure 2 shows, the market's posterior belief p_t drifts upward in the absence of a retraction. Upon observing a retraction, the market belief jumps downward. The magnitude of this jump is related to the agent's reputation level p_t at the time t of the retraction.

[4] A sample path that illustrates the dynamics of posterior beliefs in our model is shown in Figure 2.

[5] We have chosen not to model the agent's actions explicitly, as our data is not sufficiently rich to identify a model with both incomplete information and moral hazard. However, both the agent's output and the market's reward can be endogenized in the model through a choice of (unobserved) retraction-reducing effort. This suggests a version of the career concerns model of Holmström (1999) that allows for lumpy output and coarse signals, similar to Board and Meyer-ter-Vehn (2013) or Bonatti and Hörner (2017a).

[6] This type of Poisson process is analogous to the models in Bonatti and Hörner (2017b) and Halac and Kremer (2017), which analyze how the arrival of bad news impacts stopping decisions and strategic interactions.

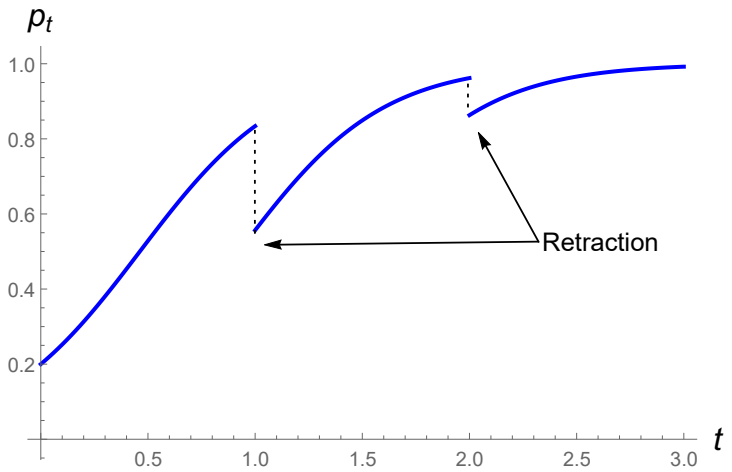

Figure 2: Reputation Dynamics ($\lambda_B = 4, \lambda_G = 1, p_0 = 1/4$)

Specifically, when an agent with reputation p_t retracts an article, his reputation drops 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

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

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 $\Delta\left(p_t\right) = 0$. Clearly, when retraction events are uninformative, they cause no change in reputations.

Furthermore, equation (1) shows that the reputation loss depends only on the market's beliefs prior to the retraction and on the *relative occurrence* of retractions for high- vs. low-reliability scientists. Consequently, the following measure of the *informativeness* of retractions is sufficient for the market's belief updating process:

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

Letting p denote the agent's current reputation level, we may then rewrite the change in reputation as

$$\Delta\left(p, \alpha\right) = -\frac{p\left(1 - p\right)\left(\alpha - 1\right)}{p + (1 - p)\,\alpha}. \tag{2}$$

Figure 3 illustrates the change in reputation $\Delta\left(\cdot, \alpha\right)$ for several values of α.

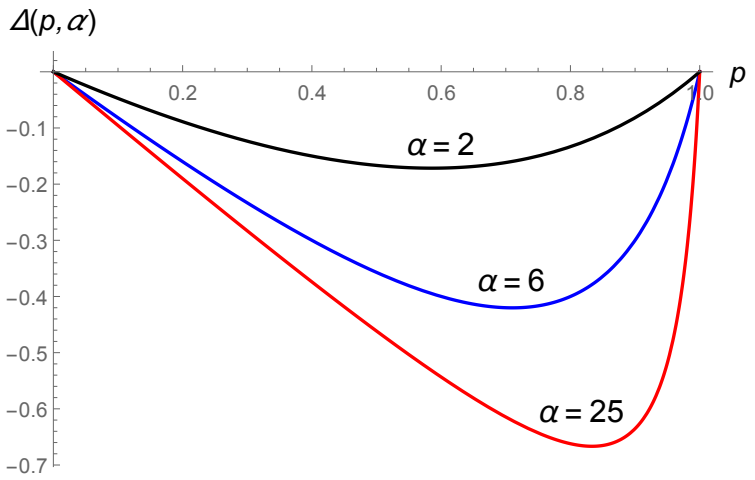

Figure 3: Reputation Losses ($\alpha \in \{2, 6, 25\}$)

As Figure 3 shows, the negative effect of a retraction is a nonlinear function of the agent's prior reputation: for $p = 0$ and $p = 1$, the market's prior belief is so strong that no signal can affect it. In contrast, when the market is very uncertain about the agent, the reputation change is large: the loss $-\Delta(p, \alpha)$ is greatest for an agent with an intermediate reputation.

We now turn to the comparative statics of reputation losses with respect to the informativeness of the signal. We are particularly interested in whether a more informative signal has a larger effect on agents with higher prior reputations. Proposition 1 collects our comparative statics results.

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

1. *retractions yield greater reputation losses for all values of p;*
2. *reputation losses are increasing in the prior reputation only for low values of p.*

Part (1) establishes that $\partial\Delta(p, \alpha)/\partial\alpha < 0$ for all p. This result is intuitive: if signals are uninformative ($\alpha = 1$), then $\Delta(p, 1) = 0$ for all p. Conversely, if signals become arbitrarily informative ($\alpha \to \infty$), then $\Delta(p, \alpha) \to -p$. Part (2) shows that the interaction effect $\partial^2\Delta(p, \alpha)/\partial\alpha\partial p < 0$ if and only if $p < \alpha/(1 + \alpha)$. This effect follows a similar logic to the *level* of reputation losses. For agents with sufficiently high reputation levels, the market essentially attributes a retraction to "chance," which also dampens the negative effect of greater signal precision.As signals become arbitrarily informative, the negative effect of a retraction becomes increasing in p on $(0, 1)$. Moreover, an increase in signal precision is most damaging to the agents with the highest reputation. However, for any finite level of

informativeness, the effects of prior reputation and signal informativeness on the retraction penalty remain an empirical question.

3.2 Implications for Citations

We now turn to the average effect of retractions across a population of heterogeneous agents. We consider a population of agents i whose reputations p^i are uniformly distributed, i.e., $F(p^i) = p^i$. This is consistent with our empirical approach in Section 5, where we use the quantiles of the distribution of citations and funding as proxies for a scientist's reputation.[7]

In order to compare the average effect of a retraction on the citations of scientists with high and low initial reputations, we partition the population of agents i in two groups, and we aggregate the reputation losses at the group level. We assume that the parameters (λ_B, λ_G) of the Poisson process governing the occurrence of retractions are common to all agents i. We then compare the average effect of a retraction in each group. In particular, for a given quantile p^*, we define the average reputation drop for agents with initial levels of reputation $p^i \in [0, p^*]$ and $p^i \in [p^*, 1]$, 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 now study the gap in the reputation losses of the two groups as a function of the signal's informativeness. We define the gap in reputation losses as follows:

$$G(p^*, \alpha) \triangleq |H(p^*, \alpha)| - |L(p^*, \alpha)|.$$

We then obtain the result in Proposition 2, which is illustrated in Figure 4.

Proposition 2 (Critical Partition) *For each α, there exists $\hat{p}$ such that the gap $G(p^*, \alpha)$ is increasing in α for all $p^* \leq \hat{p}$.*

As Figure 4 shows, an increase in the informativeness of the retraction signal α may amplify the difference in the reputation losses of high- and low-status agents. In particular, if one considers the average reputation drop across a large enough set of high-status agents

[7] Under this assumption, the initial reputation levels p^i are then uniformly distributed by construction.

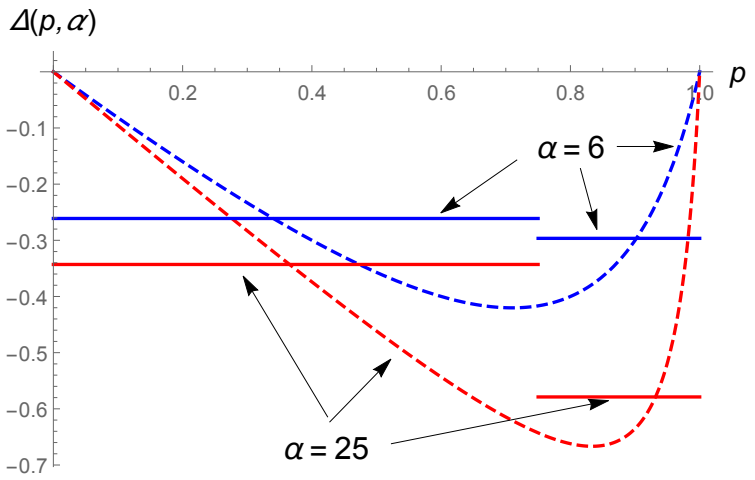

Figure 4: Average Reputation Losses ($p^* = 3/4, \alpha \in \{6, 25\}$)

(i.e., a sufficiently low p^*), then the gap between the reputation losses of high- and low-status agents is wider for $\alpha = 25$ than for $\alpha = 6$.

Viewed in the light of this result, our main empirical findings—that high-status agents after a retraction due to misconduct suffer the sharpest drop in reputation, while the three other reputation losses are of comparable magnitude to one another—are consistent with a Bayesian model where misconduct signals are more informative of the scientist's reliability than honest mistakes.[8]

It may be useful to illustrate the differences between our paper and the one in Jin et al. (2013) through the lens of our model. For instance, the role of prior reputation can explain the finding of Jin et al. (2013) that established authors (with a very high p) face a smaller retraction penalty, relative to their less prominent co-authors (for whom uncertainty still looms large). However, our model also implies that more drastic events, such as simultaneous retractions of multiple papers due to a case of misconduct, would cause more severe reputation losses for more established coauthors, reversing the above result. Unfortunately, such major events are excluded from the sample in Jin et al. (2013). In this sense, their paper focuses on retraction events of limited informativeness.

We conclude this section by deriving implications for the effect of retractions on the flow of citations. Consider an agent with initial reputation p^i. In order to correctly capture the effect of a retraction, we must consider two elements: the shape of the rewards for reputation

[8]For ease of exposition, we interpret the comparative statics with respect to the signal informativeness α as a comparison of different signals. In Appendix A, we extend the model to simultaneously account for several informative signals and obtain analogous results.

$w(p^i)$; and the drop in the market's beliefs $\Delta(p^i)$. The change in citations is given by

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

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

$$\frac{d \ln w(p^i)}{dp^i} = \Delta(p^i).$$

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

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 report regression results in *logs* and apply the insights derived earlier for 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 classification of retraction type 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.

The critical ingredient in the construction of our dataset is the *bibliome* for each retracted author, *i.e.*, an exhaustive and accurate list of articles published by these authors. A perennial challenge in collecting author-specific publication data is name disambiguation, since bibliographic databases such as *PubMed* and *Web of Science* do not typically include

individual author identifiers. A related paper by Lu et al. (2013) uses self-citation linkages (starting with the retracted paper) to build author publication histories; this approach has the advantage that it is automated, scalable, and effectively deals with errors of *commission*—mistakenly attributing publications authored by a namesake to the focal author. However, it is much less effective in warding off errors of *omission*, especially when scientists have multiple streams of work that are not connected through self-citation. In our view, this limitation argues against the use of an automated approach: using authors' prior work that falls outside the line of research culminating in a retraction can help us distinguish between the punishment meted out to individual scientists from the loss of intellectual credibility suffered by the specific ideas associated with the retraction. This concern leads us to invest in the labor-intensive process of disambiguating publication histories manually. This is a rate-limiting step for our empirical approach, and it has implications for the way in which we select control authors (see below).

First, 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).[9] Whenever the authors we identified in this way were trainees (graduate students or postdoctoral fellows) at the time of the retraction event, we excluded them from the sample.[10] See Appendix B for a full description of the process of matching author names to the Faculty Roster.

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

[9] 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. Unlike Lu et al. (2013), we cannot 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.

[10] 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.

in multiple retractions, matched authors have an average of 1.5 retracted publications in the sample. 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.).[11] 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.

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 5 illustrates this process for the case of one faculty member, Kirk E. Sperber, MD. This process allowed us to identify author publication histories while carefully removing papers belonging to other authors with similar or identical names, and reliably capturing the full set of a scientist's publications.[12]

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).[13] 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 con-

[11] We verified that including these retractions in the sample does not materially affect our conclusions.

[12] This manual process is made possible by the construction of a dossier on each author, based on a combination of curriculum vitae, NIH biosketches, *Who's Who* profiles, accolades/obituaries in medical journals, National Academy of Sciences biographical memoirs, and Google searches. More details regarding the procedure used to link authors with their publication histories can be found in Appendix C.

[13] 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).

tains cited-to-citing pairs for all *PubMed*-indexed articles that cite other *PubMed*-indexed articles.[14] Our analyses exclude all self-citations from any member of the authorship team.

Nature of retraction events: misconduct vs. "honest mistake." An important implication of our model is that different types of news should trigger different responses. In particular, the informativeness of the signal contained in a retraction event determines the extent to which the market updates on the reliability of a scientist's prior work. We propose a distinction between misconduct and "honest mistakes" as a pragmatic solution to the challenge of identifying retraction events that may be perceived quite differently by the scientific community.

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 to separate misconduct retractions from retractions due to mistakes (Appendix D provides more details on the assignment of these codes.) The difference between retractions due to misconduct and mistakes is often quite stark: misconduct retractions include cases of fabricated data and conclusions, while contaminated samples and reagents are the most frequent reasons for mistake retractions.[15]

Importantly, we do not assume that misconduct events provide stronger evidence regarding an author's reliability, relative to mistakes. Instead, our empirical specifications allow the market response to depend on the type of event in a flexible fashion. Certainly, instances of fraud and misconduct attract much more attention in the comment sections of specialized blogs such as *RetractionWatch*, while retractions due to mistakes tend to be less sensational. We comment on the relationship between misconduct and publicity when discussing our results in Section 6.

[14] 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.

[15] 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 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. Instead, an example of a "honest mistake" consists of the authors retracting a publication after realizing that they mistakenly analyzed the genetic code of a butterfly rather than a dragonfly (Arikawa et al. 1996). Occasionally, authors also retract papers due to flawed interpretation of results, or conclusions nullified by subsequent studies.

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 analysis, 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 citation distribution at baseline, and those whose cumulative citations place them in the bottom three quartiles as "low-status." Using this measure, 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 analysis, we also compute cumulative funding from the National Institutes of Health (NIH). Again, we defined high-status authors as those in the top quartile of the corresponding distribution at baseline, and low-status authors as those in the bottom three quartiles. The high-funding group accounts for 47% of all the articles published by retracted authors up to the year of their first retraction.[16]

Identifying and selecting control authors. To shed light on the counterfactual citation trajectories of retracted authors' pre-retraction publications, we need to assemble a set of control authors. The most direct approach to identifying controls would be to select from the population of scientists those whose flows and stocks of publications best mirror the corresponding flows and stocks for retracted authors, as in Jin et al. (2013). However,

[16]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.

this direct approach is infeasible, since we do not have at our disposal name-disambiguated bibliomes for every individual in the AAMC Faculty Roster. Instead, we follow an indirect approach that enables us to delineate, *ex ante*, a much smaller set of potential control authors that we expect to exhibit productivity profiles comparable to that of the retracted authors, at least on average. The onus will be on us to demonstrate, *ex post*, that treated and control authors are well-balanced along demographic characteristics and output measures.

Specifically, we focus on the authorship roster 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).[17] The procedure we follow mirrors in all respects the process we adopted to identify treated authors in the sample of retracted articles: matching the authors to the faculty roster, then assembling detailed publication histories (see Appendix B and C). The final analytic sample includes only retracted authors for whom we have located at least one matched control author. In total, we have 759 such control authors. Tables 1 and 2 demonstrate that the control authors selected by our procedure are very similar to the retracted authors along multiple dimensions, a point to which we return in more detail below.

One legitimate concern with this indirect approach to selecting control authors is that of contamination: these immediately adjacent publications could be intellectually related, or their authors might have been competing for funding during the period leading up to the retraction. If this were the case, then it is possible that the retraction event also affected the control author and the market's perception of her work. Fortunately, the data allows us to do more than speculate about the potential for contamination: we can assess empirically the extent to which treated and control authors are related. First, we use the PubMed Related Citation Algorithm (see Appendix E for more details) to ascertain whether the retracted articles and their journal/issue neighbors are intellectually related. We find this to be the case in only three instances.[18] Second, we check in NIH's Compound Grant Applicant File whether treated/control author pairs compete directly for funding. We found no instances of author pairs who applied for funding from the same component institute within NIH and whose work was evaluated by the same review committee in a window of five years before

[17] 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).

[18] We select the articles twice-removed from the retracted publication in the table of contents in these three instances.

the retraction event. Despite publishing in the same journal at the same time, we conclude that treated and control authors' scientific trajectories are sufficiently distinct in intellectual space to ward off the specter of contamination between the treated and control groups. At the same time, the fact that they are part of the same broad labor market (faculty members in US Medical Schools), participate in the same broad scientific fields, and face a similar institutional environment entails that the comparison between their publications and the citations they garner over time is substantively meaningful.

Descriptive statistics. Our sample includes 23,620 publications by 376 retracted authors and 46,538 by 759 control authors.[19] 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.

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.[20] 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.

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

[20]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.

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 6). Figure 7 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 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 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 8 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 8 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 for the process of assigning misconduct and mistake coding). The estimates reported in columns (4a) and (4b) of Table 3 do suggest a much stronger market response 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.[21] But are the differences in treatment effect across subgroups themselves statistically significant? This is less clear, since our strategy of splitting the overall data into four subgroups results in relatively noisy estimates for some of the subgroups. An alternative is to pool the entire data and focus on the coefficients for the interaction effects corresponding to each subgroup. Appendix G discusses these comparison challenges and deploys two different approaches to comparing magnitudes statistically. Regardless of the approach chosen, the statistical tests support the main qualitative conclusion: high-status authors embroiled in misconduct cases are punished significantly more severely than high-status authors guilty of making a mistake resulting in a retraction. The claim that the high-status misconduct group's penalty is greater than that of all the other subgroups is statistically more tenuous.

6 Discussion

Bayesian Learning. 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 case of misconduct is more informative about an author's reliability, *i.e.*, a higher α in Proposition 1. See Figure 3 for the intuition behind this result.

Second, the most significant effect of retractions occurs only after a misconduct event for authors in the top status quartile. 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 4 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

[21] Appendix F shows that even after removing from the sample authors involved in multiple retractions across multiple years, these patterns continue to hold, at least qualitatively.

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 explains why very high- and very-low status authors may 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 3. Instead, to generate the coefficients graphed in Figure 9, we partition authors into quintiles of the status distribution.[22] We then contrast the effects of different types of retraction events for each of five status grouping. Figure 9, 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 9, Panel B).[23]

Together, these results suggest that the market response to a retraction event is consistent with Bayesian learning about the author's reliability. In particular, the distinct responses to mistakes and 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, more reliable on average, we would expect them to exhibit a lower *misconduct-to-mistake* ratio.

Market overreaction. We now explore three distinct explanations for the discrepancy between the empirical distribution of retraction events and the theory consistent with an equilibrium market response, i.e., for why the market "overreacts" to instances of misconduct by high-status authors.

It is possible that the market overestimates the informativeness of misconduct events. Under this interpretation, the outlook on the market's ability to correctly "parse the truth"

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

[23] 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.

is quite bleak. Quite simply, the scientific community may perceive the misconduct (vs. mistake) signal as more revealing while, in fact, high-reliability authors cheat at a similar rate as low-reliability authors—they just retract fewer papers.

However, misconduct retractions come closer to our definition of scandal—a suddenly publicized transgression. Very few, if any, instances of retraction due to mere error lead to editorials, pontificating, or hand-wringing in scientific journals or the national press. Instead, much of the public attention to the retraction phenomenon can be attributed to a handful of high-profile cases of scientific misconduct.[24] Thus, an equally plausible explanation for the discrepancy in responses is based on *rational inattention*: acquiring information about the validity of scientific results is costly, but it is relatively cheaper to learn about highly-publicized retractions. This mechanism introduces a scale dimension to the market response, whereby a larger number of researchers are aware of fraud by famous authors, which leads to a proportionally larger drop in citations.

Finally, 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.

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

[24]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 usse 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.

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.[25] 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.

A troubling narrative is that prominent and powerful scientists escape major reputational damage following scientific misconduct scandals, while their junior colleagues shoulder most of blame. For example, a March 2017 New York Times article titled, "Years of Ethics Charges, but Star Cancer Researcher Gets a Pass," reported that cancer biologist Carlo Croce of Ohio State University had avoided any official sanctions from the university or federal agencies, despite multiple accusations of academic fraud. Croce "largely placed the

[25]When false science persists in the literature without retraction, the authors reputations benefit from the additional credit for both productivity (one additional item on their CVs) and their abnormally large citation counts.

blame for any problems with figures or text on junior researchers or collaborators at other labs" (Glanz & Armendariz 2017). This report fits the more general storyline of established professionals and executives skirting blame following scandals.

Jin et al. (2017) show that scandals impact the reputations of early career scientists more than that of senior authors. But our evidence speaks to reputation penalties across researchers who are in a similar career stage. Among scientists with established track records, we find that retraction events involving misconduct disproportionately hurts the reputation of the most prominent authors.

Our findings do not imply that the scientific community currently has the optimal incentive system, but by showing the additional punishments for misconduct and revealing that senior scientists cannot escape blame, our results do speak against the jaundiced narrative that regards peer review as fundamentally undermined by hypercompetitiveness, fraud, and other forms of misconduct (Fang & Casadevall 2015). Furthermore, the results highlight the importance of transparency in the retraction process itself. Retraction notices often obfuscate the difference between instances of "honest mistake" and scientific misconduct in order to avoid litigation risk or more rigorous fact-finding responsibilities. In spite of this garbled information, our study reveals that the content and context of retraction events influences their fallout. We surmise that more straightforward "findings of fact" published concurrently with a retraction notice would allow the scientific community to mete out punishment more effectively, thus buttressing the norms that govern the Republic of Science.

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, Jeffrey L. Furman, Joshua L. Krieger, and Fiona Murray. 2015. "Retractions." *Review of Economics and Statistics* **97**(5): 1118-1136.

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. 2017a. "Career Concerns with Exponential Learning." *Theoretical Economics*, **12**(1): 425–475.

Bonatti, Alessandro, and Johannes Hörner. 2017b. "Learning to disagree in a game of experimentation." *Journal of Economic Theory*, **169**: 234–269.

Budd JM, Sievert M, and Schultz TR. 1998. "Phenomena of Retraction: Reasons for Retraction and Citations to the Publications." *JAMA* **280(**3): 296–97.

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

Dranove, David, Subramaniam Ramanarayanan, and Yasutora Watanabe. 2012. "Delivering Bad News: Market Responses to Negligence." The Journal of Law & Economics **55**(1): 1–25.

Egan, Mark, Gregor Matvos, and Amit Seru. "The Market for Financial Adviser Misconduct." NBER Working Paper #22050.

Fang, Ferric C., and Arturo Casadeval. 2015. "Competitive Science: Is Competition Ruining Science?" *Infection and Immunity* **83**(4): 1229-1233.

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.

Fich, Eliezer M. and Anil Shivdasani. 2007. ″Financial Fraud, Director Reputation, and Shareholder Wealth.″ Journal of Financial Economics, **86**(2), 306-36.

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.

Glanz, James and Agustin Armendariz. "Years of Ethics Charges, but Star Cancer Researcher Gets a Pass" *The New York Times*, March 8, 2017.

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.

Halac, Marina and Ilan Kremer. 2017. "Experimenting with Career Concerns." Working Paper.

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

Jarrell, Gregg, and Sam Peltzman. 1985. "The Impact of Product Recalls on the Wealth of Sellers." Journal of Political Economy **93**(3): 512–36.

Jin, Ginger Zhe, Benjamin Jones, Susan Feng Lu, and Brian Uzzi. 2017. "The Reverse Matthew Effect: Catastrophe and Consequence in Scientific Teams." Also NBER Working Paper #19489 (2013).

Lacetera, Nicola, and Lorenzo Zirulia. 2011. "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.

Lu, Susan Feng, Ginger Zhe Jin, Brian Uzzi, and Benjamin Jones. 2013. "The Retraction Penalty: Evidence from the Web of Science." *Scientific Reports* **3**: 3146.

Mailath, George J. and Larry Samuelson. 2006. *Repeated Games and Reputations: Long-Run Relationships*. New York, NY: Oxford University Press.

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.

Prince, David W., and Paul H. Rubin. 2002. "The Effects of Product Liability Litigation on the Value of Firms." American Law and Economics Review **4**(1): 44–87.

Srinivasan, Suraj. 2005. ″Consequences of Financial Reporting Failure for Outside Directors: Evidence from Accounting Restatements and Audit Committee Members.″ Journal of Accounting Research, 43(2), 291-334.

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.

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 reliability, 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 3 (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.

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 3 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 2 Use the definition of $\Delta(p,\alpha)$ given in (2) 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:

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

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

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.

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.

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

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.

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

[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`.

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. 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.*

[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.

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 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).

Appendix F: The Impact of Authors with Multiple Retractions

In the analyses presented in the main body of the manuscript, we use the retracted authors' earliest retraction event as their year of "treatment." One problem with this approach is the possibility that scientists are associated with retraction events across multiple years over the course of their careers. This is not idle speculation: 71 of the 376 authors (18.88%) in the sample have retraction events that occur in more than one calendar year (with a maximum of nine different years with retractions). If we also count authors who retract multiple articles *in the same year*, the number of multiple retractors grows to 115 (30.59%). This appendix investigates the role that these "multiple retractors" have on the manuscript's key findings. First, we provide descriptive summaries regarding the nature of the retracted authors with retractions that span more than one year. We focus on this group, rather than additionally including authors with multiple retractions in the same year, because the paper's theoretical and empirical approach rely heavily on the timing of the earliest retractions as representing one-time shocks to scientist reputation. Retractions across multiple years are a threat to this approach because, in those cases, the timing of the reputation shock is spread out such that our empirical estimation could be picking up reactions to the subsequent retractions

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

and assigning those responses to the earlier events. Next, we evaluate the impact of the multiple-retraction cases by running the regression analyses with the multiple retractors removed.

As one might expect, multiple retractions authors are more likely to be affiliated with misconduct retractions than the singleton retractors. Table F2 shows that of the 71 scientists in our sample with retractions in more than one year, 69.01% have earliest retractions associated with misconduct. Among the 305 single retractions authors, only 33.77% have misconduct earliest retractions. This disproportionate number of misconduct cases among the multiple retractors raises the possibility that more drawn out (and potentially more severe and higher-profile) "retraction episodes" drive the negative citation impact for prior work. This concern motivates our supplemental regression analysis that excludes the multiple retractors group (see below).[viii] In contrast to the misconduct split, the proportion of high-status authors (as measured by cumulative citations) is fairly similar for singleton and multiple retractors. Table F3 shows that 30.49% and 23.94% of single and multiple retraction authors, respectively, are high status.

To evaluate the impact of multiple retractors on our results, we ran our primary sets of regressions (Tables 3 and 4, as well as the event-study graph) excluding the authors with multiple retraction events spanning multiple years. We also excluded their associated control authors, so that the sample contains the articles of authors with retraction events occurring in the same year, together with the corresponding control authors identified by looking at articles in the same/journal issues as the retraction events. Naturally, excluding these authors means that the coefficients of interest are less precisely estimated, especially when splitting the sample using the misconduct and status covariates.

Column 1 of Table F5 shows that despite the more limited sample, we still see that retraction authors experienced a statistically significant drop in citations to prior work after the retraction event. The magnitude of this effect is slightly smaller than in the main analyses (8.1% yearly decrease in citation rate, as opposed to 10.7% in the full sample version). Figure F2 displays the interactions between the treatment effect variable with indicator variables for number of years until the retraction event. The general pattern is the same as the full sample version (Figure 8), but there are two notable differences. First, the confidence intervals are slightly larger in Figure F2. Second, the magnitude of the treatment effect is more steady over time than in the full sample analysis, with very little change between year three and year nine after treatment. One explanation for this difference is that the steady negative slope in the full analysis graph may be influenced by the multiple retraction authors' later retraction events, which serve as additional shocks to their reputation.

The results in Table F6 (which is analogous to Table 4 in the paper) show that the subgroup comparisons remain directionally the same with multiple retractors excluded. We still find the most stark difference between mistake and misconduct authors within the high-status category (as measured either by citations or funding). We also continue to observe similar magnitudes in the citation decline for low-status authors, whether or not misconduct is involved in the corresponding retraction event.

We highlight two important findings contained in these supplemental analyses. First, multiple-retraction authors are more likely to be associated with misconduct cases than are singleton retractors. Removing the multiple-year retraction authors from the analysis sample reduces statistical power, but our main results continue to hold, at least qualitatively. A single retraction event is enough to damage the reputation of an author's prior work, and the punishment meted out by failing to cite these authors still varies across misconduct and status subgroups in ways that are consistent with our theoretical predictions.

[viii] Jin et al. (2013) and Lu et al. (2013) remove authors who retracted publications multiple times from their analyses. This data construction choive makes the difference-in-differences analysis more straightforward, but also entails that the misconduct/mistake split in their sample of authors is less representative of the split in the overall population of retracted authors.

Appendix G: Comparing Magnitudes for the Treatment Effects Across Author Groups

In our analysis of heterogeneous treatment effects, we split the sample by retraction type interacted with status cells (Table 4). The advantage of the "split-sample" approach is that the corresponding treatment effects are estimated off control groups that correspond exactly to the treatment group (e.g., high-status retracted authors are only compared to high-status control authors). A downside of this approach, however, is that it becomes more challenging to compare the magnitudes of the treatment effects, since they are estimated from separate samples that are not randomly selected from the overall data.

An obvious alternative is to run a regression on the entire sample, interacting the "after retraction" treatment variable (as well as all the other covariates) with indicator variables for each of the four subgroups. Unfortunately, we cannot get this fully saturated specification to converge.

We use two different approaches to statistically compare the magnitudes of the treatment effects across subgroups. First, under the assumption that, conditional on the included covariates, the four subsamples are randomly drawn from the overall data, then we can compare the coefficients' magnitudes using a Z-test in the spirit of Clogg et al. (1995):

$$Z = \frac{\beta_{1a} - \beta_{1b}}{\sqrt{SE_{1a}^2 + SE_{1b}^2}}$$

Using this approach, the p-value for the one-tailed test of equality between the treatment effects in columns (1a) and (1b) is 0.08. But the p-value for the one-tailed test of equality between the treatment effects in columns (1b) and (1d) is 0.13: the difference between these coefficient estimates is not statistically significant at conventional levels.

Since the Z-test relies on a problematic assumption (that the subsamples are randomly drawn from the overall data), we also pool the data in a slightly simplified version of the fully-saturated model. In particular, our pooled specification does not include a full suite of interaction terms between the year effects, age effects, and the subgroup indicator variables. The results are displayed in Table G1. Columns (1) and (2) utilize the entire analysis sample, and columns (3) and (4) split the sample by author citation status. We use Wald hypothesis tests to compare the magnitudes of the treatment effect variables. In column (1), we observe a statistically significant ($p < 0.05$) difference between the misconduct and mistake groups. In column (2), we compare the treatment effects for each of the four subgroups. The Wald test statistic allows us to reject the null hypothesis that the coefficient on the treatment variable "After Retraction × High Status × Misconduct" is equal to any of the other treatment variables (when the Wald test is performed in a pairwise manner). We cannot reject the null hypothesis that any of the other three treatment variables are equal at conventional levels of statistical significance. Column (3) also offers support for the claim that the high-status misconduct group suffers the largest citation penalty; conversely, in column (4), the two low-status subgroups experience penalties that are statistically indistinguishable from one another.