RESEARCH ARTICLE

Is "the time ripe" for quantitative research on misconduct in science?

Harriet Zuckerman

Columbia University, 450 Riverside Drive, New York. New York 10027, U.S.A.

Keywords: bias, deviant behavior, research misconduct, trust in data

ABSTRACT

Misconduct in science is a timely and substantively important problem in the social study of science. But in the absence of comprehensive and reliable data needed for analysis, formidable obstacles stand in the way of its being studied quantitively. Accessible databases, including government data, are flawed, while undertaking new data collection presents its own problems. First, little is known about biases in official government reports. Second, official reports exclude classes of malfeasance other than fabrication, falsification, and plagiarism of evidence (FFP). Third, while drawing on official data is expedient, available official information is thin; it tells little about miscreants and fails to identify potential causes of their actions and the environments in which misconduct occurred. Fourth, it also fails the test of permitting estimates to be made of populations at risk, making it impossible to calculate incidence. A healthy dose of skepticism is in order in evaluating both the findings of current quantitative studies and of proposals for remediation.

1. INTRODUCTION

Misconduct in science has attracted considerable interest among economists, political scientists, policy analysts, sociologists, some researchers in science studies[1], and, of course, scientists of all stripes. It is also a thriving area of research, according to Google Scholar's database, a rough but ready indicator of research activity. It lists some 1,310 papers, books, conference presentations, and other items on some aspect of "scientific misconduct" in its database in 2019. Indeed, items on "scientific misconduct" grew sharply in Google Scholar: In 1981 there were fewer than 10 such entries, while in 2011 this number jumped to 725 (Ben-Yehuda & Oliver-Lumerman, 2017, Figure 4.1, p. 96).[2] The absolute number of entries and steepness of the associated growth curve were surprising to this observer, who had written on the subject in 1977 and 1984 (Zuckerman, 1977, 1984) when the subject was far from a "growth industry." One had to wonder whether the apparently escalating interest in misconduct signaled in Google Scholar 's database in

[1] In general, empirical research in science studies is skewed toward qualitative inquiries and smallbore case studies. At least this seems to be so, based on recent book length studies published in the field and the contents of journals such as *Social Studies of Science* and *Science, Technology and Human Values*. This tendency is consistent with the view that some have that quantitative evidence and the analysis that goes with it are inevitably misleading when trying to investigate what scientists do, and indeed, a lot of other social phenomena.

[2] Google Scholar was launched in 2004, but lists publications relevant to this inquiry since 1980. Its coverage of the research literature includes all manner of publications, but is limited to those in English. The number given here departs slightly from the one shown in the Ben-Yehuda and Oliver-Lumerman's graph (2017, Figure 4.1, p. 96).

QSS

an open access journal

Check for updates

Citation: Zuckerman, H. (2020). Is "the time ripe" for quantitative research on misconduct in science? *Quantitative Science Studies*, *1*(3), 945–958. https://doi.org/10.1162/qss_a_00065

DOI: https://doi.org/10.1162/qss_a_00065

Corresponding Author: Harriet Zuckerman
haz1@columbia.edu

Handling Editors: Loet Leydesdorff, Ismael Rafols, and Staša Milojević

The MIT Press

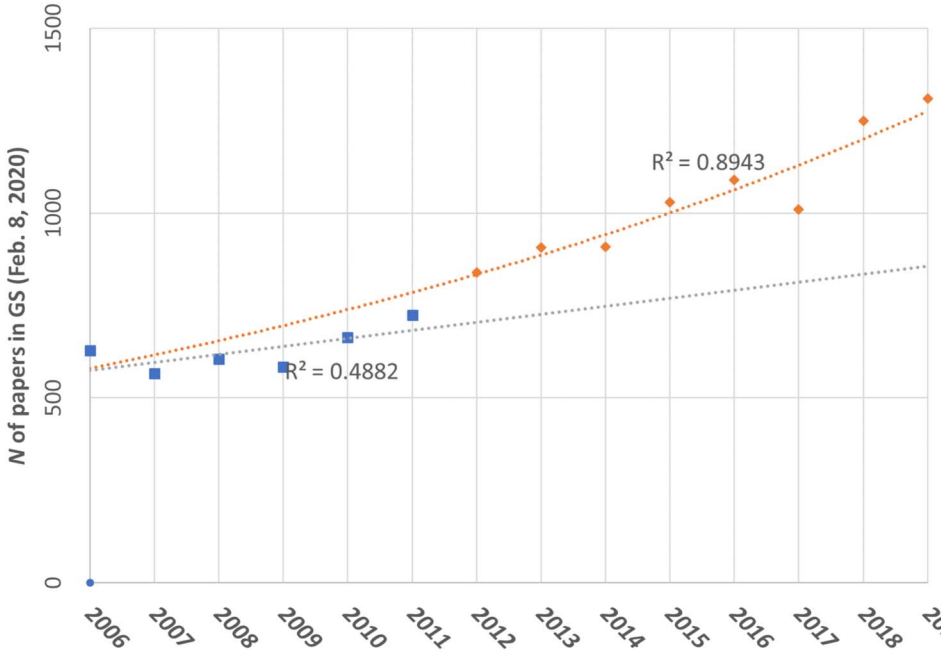

Figure 1. Listings in Google Scholar of scientific misconduct 2006–2020. Source: The figure is based on Ben-Yehuda and Oliver-Lumerman (2017, Figure 4.1, p. 96), but was amended and redesigned for this paper by Loet Leydesdorff, February 8, 2020.

the first 30 years had been sustained or had fizzled out, as is so common in empirical studies of research attention. Figure 1 compares the growth rate in Google Scholar mentions for 2006–2011 with those for 2012–2019 and answers this question.

Interest in research misconduct has not only continued but, as Figure 1 shows, has escalated. The increase in absolute numbers is still striking, while the annual rate of increase in the later years is even faster than it was earlier. It rose 5.89 times from 2012 to 2019 and 1.61 times from 2006 to 2011. Without putting a great deal of emphasis on the precise numbers shown here, it is fair to conclude that the phenomenon of scientific misconduct has clearly captured a good deal of scholarly attention.

It remains to be seen, however, what a content analysis of these papers would show: about the kind of scholarship they represent, what subjects they address, the extent to which these papers report empirical research, and what methods and data they use, much less the disciplinary origins of the authors; and if they are journal articles, how they are distributed among journals and what share of them measure up to the kind of quantitative analysis that *Quantitative Science Studies* seeks to publish. Does all this activity indicate that a solid body of quantitative studies of misconduct exists or is in the making? I confess that I have not done the substantial work that answering these questions requires.

When the chance arose to return to the subject of research misconduct, especially to the question of its incidence, a review of the current literature was clearly in order. The more I read, the more it seemed that misconduct was far more often "talked about" than studied quantitatively, and when quantitative studies were done, they faced substantial methodological obstacles. These derived in large part from inadequacies in the available data and also from the formidable difficulties of launching new data gathering. As a substitute, then, for the more orthodox research paper I originally intended write, the remainder of these comments address the significant problems that arise in doing quantitative research on misconduct in science and what steps might be taken to improve the current circumstances. As will become evident, these comments draw far

more often on the evolution of the subject of research misconduct in the United States than is desirable[3]. This is likely the self-exemplifying outcome of the greater amount of U.S. data available and the greater amount of research done on the U.S. experience than elsewhere. It does not show whether misconduct is more frequent in the U.S. than elsewhere.

My comments address the following: the shortcomings of available data; how definitions of misconduct have changed in the course of its becoming a matter of official oversight and the complexities these events have introduced into research potentials; current foci of attention on the "replication crisis" and rising numbers of journal retractions; the introduction of technologies for detection of misconduct; the consequences of press attention on misconduct for public confidence in science; and repeated observations, especially the need for an epidemiology of research misconduct.

2. SHORTCOMINGS OF AVAILABLE DATA

As my paper published over 40 years ago (Zuckerman, 1977) observed, it is odd that nothing like an epidemiology of misconduct exists in science, a domain in which data are considered central, careful record keeping is obligatory, and statistical sophistication is common. It seems likely that the absence of good evidence derives from the long-held belief among scientists that misconduct is exceedingly rare and that it is unproblematic for ongoing scientific practice, and thus needs no further attention.

In that same paper, I also observed that the study of misconduct in science presents many of the same problems that researchers on deviant or criminal behavior have faced for decades. For example, official data on incidence are very likely biased, because bias exists in methods of detecting deviance and thus in reports based on them. To what extent is this so for misconduct in science? While there have been some studies done using official data from the U.S. databases, they are limited and, in any event, there is no way of knowing about the biases inherent in them. We do know that official data, that is, data collected by government agencies charged with overseeing the most consequential instances of misconduct, are limited only to acts of "proven" violations. There is no information on charges brought but that failed to be substantiated or the characteristics of complainants and in what circumstances they came forward; there is no counterpart to officially stipulated penalties for particular acts; and, very importantly, there are no efforts made to identify the populations at risk for engaging in misconduct. Without major attempts to identify the populations from which miscreants are drawn, it is impossible to calculate their incidence. The problems multiply if one is interested in determining whether changes have occurred in the incidence of particular acts of misconduct or in other attributes of misconduct. In the absence of such information, assessing the effectiveness of interventions to reduce misconduct is not possible.

The list of shortcomings also includes the greatly compressed official definition of misconduct that now includes fabrication, falsification, and plagiarism of evidence (or the frequently used acronym, FFP)[4]. These three are undeniably significant violations of the standards that scientists

[3] Research misconduct and concern about it are far from U.S.-specific phenomena. Agencies aimed at overseeing and curbing research misconduct have been established in Canada, China, Denmark, Israel, Norway, and the United Kingdom, among other nations. They differ greatly in their purviews and operations (Steneck & Mayer, 2017). Comparative cross-national research on misconduct and responses to it is therefore especially challenging.

[4] Definitions used by the National Institutes of Health (NIH) and National Science Foundation (NSF) differ but not a great deal. According to the latter: Misconduct is "a significant departure form accepted practices of the scientific community for maintaining the integrity of the research record. The misconduct must be committed intentionally, or knowingly, or in reckless disregard of accepted practices. The allegation be proven by a preponderance of the evidence." Later in these comments, I will have more to say about this language. https://grants.nih.gov/policy/research_integrity/overview.htm.

believe are required for scientific contributions to be considered credible. But are they the only actions worthy of inquiry? They are not. (I will come back to this question shortly.)

Other sources of information are also problematic. Self-reports of violations are considered better indicators than official reports in the criminological literature. But they too are suspect for a variety of reasons. How solid is the evidence gathered in surveys from those who report they have engaged in research misconduct? Intersubjectivity is always a problem in survey research, and this is surely so here. What acts do respondents' have in mind when they say they have fabricated or falsified data? It would be illuminating to know in greater detail what counts as falsification in the minds of survey respondents. For example, where is the line to be drawn between reporting all the evidence investigators have assembled and only those findings they believe are valid? Careful study of historical evidence of alleged falsification demonstrates (see Holton, 1994) that this line is blurred and that investigators are reasonably clear about which observations are likely to be correct and which are not.

Fabrication of evidence is more straightforward, but questions, in surveys for example, should address how central the admitted fabrication was to reported research outcomes, the frequency of occurrence of fabrication, whether others knew, and whether they were ever detected[5]. Based on the near-absence of information on respondents' interpretations of the phenomena addressed in such questionnaires, one can, at best, guess about biases in self-reports, including whether they undercount (as many believe they do) or overcount the infractions they are designed to measure.

Similarly, the reliability of first-hand witness accounts reported in surveys or made public by whistleblowers' reports is also not self-evident. Despite official protections provided to whistleblowers, it is likely that they underestimate the incidence of misconduct. It is said, for example, that whistleblowing is a luxury few working scientists can afford—and this seems obvious for those who are young and those of relatively low standing. Although senior scientists may be less vulnerable than their lower ranked coworkers, and while it is ultimately in their interest to report violations they know to be true, in the short term they may also be reluctant to report having witnessed incidents of misconduct "on their watch" or because they are loath to damage the careers of students or colleagues.

Furthermore, some acts of misconduct are more visible than others "to the naked eye." How frequent are cases of miscreants being caught in the act, as William Summerlin was some decades ago? His undoing occurred when a skeptical technician discovered that a black patch of fur on a white mouse could be easily removed with alcohol. This black patch was fundamental to Summerlin's claim that the spotted mouse was evidence that he had successfully transplanted skin between genetically unrelated animals (Hixon, 1976).

More likely, misconduct is detected by intentional reviews of research notebooks being undertaken, although even these data can be, and have been, skillfully doctored by miscreants or because the evidence itself is itself not dispositive.

Other means of detection are now in use, including software designed to capture plagiarism of text and images and the fabrication and "fixing" of images taken from others' publications to make them seem original contributions to the fabricators' research. Statistical tests are also used to gauge whether reported statistics are so out of bounds that they are likely to be fraudulent, although here too, drawing on research by historians of science, there is the example of the classic claim that Gregor Mendel's statistical observations of genes and sweet peas were "too good to be true," which was shown to be unfounded; and furthermore, Ronald Fisher, often held responsible for

[5] See https://grants.nih.gov/policy/research_integrity/overview.htm.

the accusation, was not Mendel's accuser, but rather another statistician was answerable for the charge (Radick, 2015). In short, these new interventions may introduce biases into research on misconduct, not least by increasing chances of detection of misconduct in the kinds of misconduct they were designed to reveal, while leaving others untouched other forms of deviance. Account should also be taken of the likelihood that all such technologies are susceptible to manipulation by sophisticated miscreants and that keeping ahead of them will not be easy.

Available data on retractions of papers also present problems for research on misconduct. What does the avalanche of papers listed in the Retraction Watch reveal? Retraction Watch, both blog and database, was established in 2002 as a means of making published science "more transparent." By assembling retractions of articles published not just in a handful of journals but across fields and over time, the founders' intention was to make retracted articles more visible and thereby to reduce their use by others who were unaware they might be using flawed or erroneous research papers in their own research. As one of its founders later put it, they aimed to highlight "fictional science" for otherwise uninformed investigators and also to become a "window on the self-correcting nature of science, … [that] can provide insight into cases of scientific fraud" (Oransky & Marcus, 2010).

As of January 2020, the Retraction Watch database listed as many as 21,792 papers, almost all withdrawn by authors or editors. This is so large a number that researchers seeking promising databases for use in studies of misconduct have used it. Yet there are also reasons to be skeptical about the reliability of these data and the studies using them (Brainard & You, 2018).

As those who have examined published lists of retractions know, the details of retracted papers and the reasons given for their withdrawal are highly variable. Some are richer than others, but most are thin, at best. Furthermore, the reasons given for retractions are often ambiguous and thus difficult to interpret and classify. Others are sufficiently clear-cut to make it possible to distinguish between those reporting evidence of misconduct and of questionable research practices, and those reporting errors of various kinds. (I will return to errors later in the text.) Still other papers, one finds, have been retracted because the reported evidence was less persuasive than later results assembled after the submission of papers. Retracted papers do offer data for examining the effects of retraction on citation before and after retractions occurred (Garfield, 1987b) and on the effects of retractions on their coauthors' citations, both revealing consequences of misconduct, an aspect of misconduct not pursued often enough.

Another strand of research on misconduct draws on surveys of scientists' views about the relative seriousness of "questionable research practices" (QRPs). These studies broaden perspectives on deviance in science and variations in the perceived gravity of particular acts. Judging from the data reported in these papers, QRPs are numerous. Among others, they include depriving coworkers of authorship rights, violations of rules protecting human and animal subjects, failure to disclose sources of financial interest that might bias research, and "self-plagiarism" or "text recycling," labels for republishing one's own prior work (Horbach & Halffman, 2019). Some of these clearly violate rules promulgated by funding agencies or by journals seeking to discourage such practices in papers they publish. Others, such as "self plagiarism," (a seeming oxymoron) seem less pernicious than FFP, but journals have been known to refuse to publish papers containing earlier published texts by authors unless ample notice is given that the text has appeared before and that the submission contains new material worthy of publication. Despite clear statements in at least one publication of the NIH's Statement of Policy on Plagiarism[6] that self-plagiarism is not considered research misconduct, it has been labeled as unacceptable and some

[6] https://grants.nih.gov/policy/research_integrity/overview.htm; and the equivalent from the National Science Foundation, https://www.nsf.gov/oig/regulations/.

scientists, with substantial records of research and multiple publications, to have been subjected to public attacks for reproducing their earlier work. It remains quite literally a "questionable research practice."

In principle, data collected in high-quality surveys of the significance accorded to QRPs would add to our understanding of misconduct and the practice of science more generally. I might add that further inquiry into scientists' reasons for believing that some practices are unacceptable while others are trivial is in order, and coupling this with evidence on the incidence of QRPs would be even more valuable. And to raise the ante even further, focusing research attention on the consequences of QRPs—those they entail for the reliability of published science and for miscreants and victims—would be even more illuminating. Considering the consequences of misconduct has sometimes figured in discussions about its costs, but not often. The price that misconduct imposes on the well-being of science is a central question in the epidemiology of misconduct, just as the epidemiology of disease is central to public well-being.

Finally, a new line of research on misconduct has focused on the influence of the organizational contexts in which it occurs, and provokes a set of questions entirely consistent with the creation of an epidemiology of research misconduct. Promising studies have focused, for example, on the influence of variables such as the degree of social integration in organizations and the willingness of those who work in them to skirt rules and to report misconduct when they believe it has occurred (Ben-Yehuda & Oliver-Lumerman, 2017, Chapter 3). Another example is the proposition that control over the research practices of scientists may be less inhibiting than it might seem when organizations permit "noncompliance to go unrecognized/and or unchallenged," thereby allowing for flexibility in the nature of control over scientific practice (Leahey & Montgomery, 2011, p. 305).

In any event, feasibility of data collection must loom large in any researcher's problem choices. I do not mean to suggest that quantitative research on significant aspects of misconduct is impossible. I do mean to suggest that currently available databases, established for other purposes, are not well-suited to research on misconduct and that a lot of hard work will be required to amplify data in hand, and much more will be needed to collect new and better data for research. Getting from here to there will not be easy.

3. MISCONDUCT: HOW ITS DEFINITION FACILITATES AND COMPLICATES RESEARCH

Histories of the institutionalization of government oversight of research concur that the arriving at an agreement among interested parties about research misconduct did not come easily. It began with Congressional hearings by the Gore Commission in 1981, prompted by the public disclosure of research misconduct cases at four major research centers in 1980. Some 12 cases of research misconduct were disclosed in the United States between 1974 and 1981. Congressional attention to research misconduct was heightened throughout the 1980s by additional allegations of research misconduct and reports that the NIH, universities, and other research institutions were responding inadequately to those allegations[7]. This account only hints at the controversial tone and substance of these discussions. Charges were made about "wasting" public tax monies on research and on the irresponsibility of grant recipients in overseeing how research was done and the funds granted for it were handled. These charges found their way into the press and the press stimulated further controversy.

It ultimately became clear that a consensus had to be reached on how research misconduct should be defined, on what the phenomenon includes and what it leaves out, and it was also necessary to agree on how it should be investigated, who was to investigate it, how it was to be

[7] https://ori.hhs.gov/historical-background.

adjudicated, and what penalties were to be imposed on those judged guilty of having committed it. At the NIH, the prime funder of research in the biomedical sciences in the United States, the Office of Scientific Integrity and its organizational twin, the Office of Scientific Integrity Review, were established in March 1989 to take on these tasks. The former was located in the NIH[8] itself and the latter in the Office of the Assistant Secretary for Health[9]. That two offices were created rather than one was an early indicator of the organizational complexity that would be involved in research oversight. The NSF, also a major funder of research in the United States, would later devise its own rules concerning research misconduct and how it would handle it. It located these activities in the Office of the Inspector General.

It took years of negotiation, for example, for the blanket term *fraud* to be renamed *scientific misconduct* and for scientific misconduct to be renamed, *research misconduct*[10], which was then was subsumed as a problem in the maintenance of "research integrity." Thus, a process of normalizing occurred in which misconduct became what it was not, and along the way the inept term *scientific misconduct* (which, after all, was not scientific at all) was eliminated.

This official definition, limited to fabrication, falsification, and plagiarism, was eventually agreed to by the NIH, the NSF, and the federal government's Department of Health and Human Services and others between 1987 and 1989. It took another 3 years for the National Academy of Sciences to agree to it in 1992. Further reviews took place in 1996 and 1999, with continuing and heated controversy erupting over the terms that the definition of misconduct should contain and how oversight should be exercised and by whom (Wadman, 1996).

Organizational complexity was multiplied by requiring all recipients of funding, including colleges and universities, research laboratories, hospitals, and other organizations receiving federal funds, whether in the United States or elsewhere, to comply with official rules and to play a role in policing research.

The principal reasons given for narrowing of misconduct to just three actions (FFP) were that prior definitions "covered too much ground," were ambiguous, and were not readily applicable in inquiries into misconduct. Those even slightly familiar with behind-the-scenes political wrangling know that official statements do not begin to capture the nuances of the conversations they seek to summarize.

Not long after settling on the FFP criteria, pressure mounted for a uniform definition that could be used across all agencies and across fields of science. (U.S. federal government lists at least 35 separate agencies and additional subagencies tasked with supporting scientific research in one form or another.) Again, multiple exchanges among the relevant actors ended up with a focus on the FFP standard, the most damaging of all to the public record of science.

This left a number of practices that some deem unacceptable (see the earlier discussion of questionable research practices) off the list of actions punishable by the federal government, a decision intended to be administratively practical. Even now, it appears that uniformity in defining research misconduct has not quite been attained on, for example, whether peer reviewers who misappropriate in their own research any ideas, methods, or findings derived in the course of reviewing without recognizing their sources, have violated a major standard of

[8] On the history of National Institutes of Health regulation, see https://ori.hhs.gov/historical-background (Steneck, 1994).

[9] https://ori.hhs.gov/historical-background. See also https://www.nsf.gov/oig/_pdf/presentations/session.pdf.

[10] The term itself evolved from fraud and the misrepresentation of findings, to scientific misconduct, to research misconduct and now, to research integrity, signaling efforts to clarify the phenomenon included and efforts to shape its public image.

conduct or not. The NSF explicitly bars such misappropriation, but not all other agencies do. For now, the FFP standard seems to have garnered some support in the scientific community, but government presence in the oversight process continues to rankle with those who still believe in the long-prized commitment in science to "self-policing."

The narrowing of the definition and controversy surrounding it is germane to quantitative research on misconduct. It has markedly limited the scope of official misconduct and thus has had the effect of making data collected by the variety of oversight groups more uniform than they might otherwise have been; a useful outcome if empirical research were to be done based on official records. At the same time, FFP, self-explanatory as these three forms of misconduct might seem, are individually more complicated than they appear. Consider the current definition of misconduct promulgated by the NSF: Misconduct is "a significant departure form accepted practices of the scientific community for maintaining the integrity of the research record. The misconduct must be committed intentionally, or knowingly, or in reckless disregard of accepted practices." The emphasis on intention and knowing or reckless disregard has the effect of eliminating from consideration of misconduct self-deception as a source of culpability. Self-deception has a long history in science, involving, as it does, scientists' inability to subject their own ideas and research to skepticism as rigorous as they are obliged to subject the work of others. The exclusion of self-deception from considerations of misconduct is not unlike the exclusion of violators of criminal law from responsibility for their acts because they are incapable of judging the seriousness of what they have done, and introduces a measure of mercy into assessment of misconduct. However, intended or otherwise, the damage done to the research record (and its impact on trust among scientists) is the same, whether the miscreant intended it or not. It is not clear to me why so much emphasis was placed on the intention to deceive in judging misconduct. It may be that there are antecedents in penal law or that it permits some measure of leniency to be exercised should investigators think it is needed.

The emphasis on maintaining the integrity of the research record in the official statements equates the record of science[11] with the well-being of science. The maintenance of trust among scientists and the necessity of their being able to rely on one another's work, published or otherwise, is oddly not given its due. Trust among scientists is of major significance for pragmatic rather than expressive reasons.

In matters of quantitative research on misconduct, clarity about the meaning of fabrication, falsification, and plagiarism is an absolute requirement. Questions also arise about how much fabrication or falsification of data has to be involved to qualify as misconduct? Similarly, questions about the extent to which others' work is misappropriated must be addressed. How much work by others must be "borrowed" without proper attribution to qualify as plagiarism? As it happens, plagiarism is the only one of the three acts considered a criminal offense, and a substantial legal record exists on just how much can be misappropriated without proper citation and permission to quote. Similarly, there is a substantial legal record concerning reproduction of images without authorization. The law concerning reproduction of both text and images is intended to protect the intellectual property and thus the financial interests of authors and image makers. In science, however, while the theft of text and images may have financial implications, as important, or more so, is the theft of peer recognition and the reputational rewards that go with it. Hard as it is to assess the losses of authors and artists whose work has been plagiarized and then to compensate them for it, it is also difficult to assess scientists' losses due to the theft of their intellectual property and hard also to compensate the symbolic rewards of recognition they would have received, along with the financial compensation that often goes

[11] https://grants.nih.gov/policy/research_integrity/overview.htm.

with it. In short, research misconduct is a difficult phenomenon to nail down if quantitative research is contemplated.[12]

4. CURRENT AND POTENTIAL FOCI OF RESEARCH ATTENTION: THE REPLICATION CRISIS, RISING JOURNAL RETRACTIONS, AND TECHNOLOGIES FOR DETECTION OF FFP

4.1. The Replication Crisis

Does the replication crisis provide new opportunities for quantitative research on misconduct? In the United States, the press has paid considerable attention to claims that published findings, especially, but not only, in economics, medicine, and social psychology, fail in attempts to replicate them. These failures do not derive from intrinsic difficulties deriving from their being hard to replicate, as for example are field experiments and those focused on phenomena that cannot be repeated at will, such as rare astronomical events. Instead, the replication crisis derives, it is claimed, from scientists employing frequently used but flawed research designs and analytic techniques. The focus here is not on misbehavior by wayward individual investigators but arises from the widespread adoption of research practices that introduce bias into research findings.

Such claims, at their extreme, assert, as one recently did, that "there is increasing concern that in modern research, false findings may be [the] majority or even the vast majority of published research claims" (Ioannidis, 2005). The practices being criticized include the use of very small samples, the investigation of weak effects, the near absence of efforts by investigators to replicate their own findings, and the pervasive use of insufficiently demanding statistical tests of significance, most notably the use of the classic $p \leq .05$ measure (Fanelli, Costas, & Ioannidis, 2017).

The replication crisis has understandably attracted considerable interest among those who do research on research. Indeed, some of the most active investigators on the replication crisis characterize their work as laying the foundations for a new discipline, "the meta-science of science," aimed at improving research methods and practices (Ioannides, Fanelli, et al., 2015).

The replication crisis has also produced proposals for remediation, for example, the suggestion to replace the common statistical standard of significance of $p \leq 5\%$ with one of just $p \leq 1\%$ (Fanelli, 2018). Upping the significance ante in this way would greatly reduce investigator's chances of having no "significant" findings and thus no publishable results.

Other proposed remediations include increasing sample sizes, greater use of large randomized trials, assigning multiple teams to work independently on the same questions, and the movement to institute what is called "preregistration" of research plans. In preregistered protocols, authors seeking to publish are required to submit their research proposals (not their completed papers) to journals agreeing to consider them. The proposals lay out in advance the methods to be used for data collection and analysis, as well as their expected findings. Journals are then tasked with determining whether the plans they receive meet the standards they have set, and if the plans do and they are realized, publication of the results of the preregistered research is guaranteed. Preregistration is intended to preclude authors from selecting and reporting findings that turn up "by chance," despite their satisfying conventional measures of "significance." (The term now used for picking and choosing according to significance is p *hacking*)

Preregistration has both critics and supporters. Some point to the inevitable multiplication of publications containing anticipated and likely humdrum findings. Others propose that preregistration might be useful if it is confined to research that has reached the stage of hypothesis

[12] See Ben-Yehuda and Oliver-Lumerman (2017, pp. 19–62) for the analysis of 748 cases drawn from the NIH database, the NSF database, Danish data, books, and the internet.

testing but is unsuitable for research aimed at hypothesis generation. Still others express concern that preregistration reduces incentives for researchers to take unexpected findings seriously and much less to consider how they might have arisen. The large returns of serendipity in research (when it occurs) should not be forgotten when seeking to apply a one-size-fits-all set of rules (Shiffrin, Börner, & Stigler, 2018). Preregistration might well increase rates of replicability and deter those who are willing to report findings that may have arisen by chance, but it is a costly deterrence to poor research practice that might be discouraged in other ways.

4.2. Retractions as Potential Data for Quantitative Analysis of Misconduct

The absolute number of retracted publications is not only large, as noted earlier, but has been increasing each year. However, when the rising number of published papers is taken into account, the rate of growth in the number of retractions has in fact leveled off since 2012. So much for rising rates of misconduct and other phenomena that retractions might signal. Yet the raw numbers of retractions now published in retraction databases are very large, and for that reason alone they appear to have attracted researchers seeking quantitative information on misconduct. But how useful are they for research on misconduct?

Retractions are intended to earmark published papers containing evidence that is unreliable because it is either erroneous or fraudulent. Errors occur in scientific inquiries, even when agreed-upon standards of practice are met. Scientists know this. They can also result from incompetent or shoddy scientific practice (including, for example, contamination of samples). In that early paper on misconduct (Zuckerman, 1977), I distinguished between "reputable" errors (those arising despite the investigators' best efforts to avoid them) and "disreputable" errors (those arising from disregard of acceptable scientific practice). If intent is taken into account in assessing cases of misconduct, authors making reputable errors are clearly innocent. But those who disregard current standards of practice, intentionally or otherwise, are guilty. Both classes of error undermine the reliability of the scientific record, whatever the intent behind them. And both classes of error waste the time of fellow scientists who take them into account and even more of those who pursue their explanation in research.

Thus, if retractions were to serve as useful evidence for studies of misconduct, they would have to be sufficiently detailed to permit researchers to sort retractions into three groups: those involving misconduct satisfying the FFP standard, those involving reputable errors, and those involving disreputable errors, errors that come close to the NSF definition of culpability due to "reckless disregard" of current standards. The texts of most retractions do not permit such distinctions to be made. Furthermore, some journals publish retractions, while others do not. Fields vary too in whether retractions are published, and editors also differ in their views about the need to publish retractions. As such, in their current state, retraction data are problematic indicators of misconduct.[13]

It has also been argued that the multiplication of retractions is evidence that peer review, said to be a bulwark against research misconduct, fails to prevent it. Does peer review sort out fraudulent research from trustworthy submissions? It very likely does not, as the publication of fraudulent papers suggests. But can peer review serve this purpose? The answer to this question is far from clear. It may serve as a deterrent against submission of fraudulent papers because authors elect not to submit their work to avoid such review. On the other hand, peer reviewers do not generally have access to the raw data on which research is based, and are

[13] See Hesselmann, Graf, et al. (2017, pp. 817–823) for an analysis of inconsistent findings in publications claiming to estimate various kinds of misconduct using retractions data.

thereby unable to check for themselves the data on which research is based and the conclusions drawn from them. It should be possible to ask editors how many papers submitted to their journals were rejected because either they or their reviewers counted them as fraudulent. Whether such evidence exists is not widely known. In general, peer review is intended to sort out those papers that meet standards of scientific practice and present plausible findings from others that fail to meet these criteria. And indeed, this is how peer review in the sciences operates.

A study in which I was involved years ago had two empirical objectives: first, to determine rejection rates of an array of journals in the sciences, social sciences, and humanities, and second, to probe the ways in which reviewers treated submissions from scientists of different standing. We learned that rates of acceptance are far higher in the sciences relative to other disciplines and this had been so for decades (Zuckerman & Merton, 1971). We suggested that the comparatively high rates of acceptance in the sciences resulted in part from the prevailing practice of presubmission review by colleagues and from the prevalence of more uniform standards in the sciences of what constitutes an acceptable paper. Furthermore, of particular importance to this discussion, when asked about the high rates of acceptance, the editors of the journal we studied in detail responded that they believed that reviewers separated reasonable submissions from the rest and that ultimately really good research would prevail and that less convincing research would "sink" in the vast array of published papers that were used rarely, if at all. As a result, the high rate of acceptance they favored allowed the community to judge what it deemed useful and what was not, and that was an effective publication strategy.

For what it is worth, not one of the large number of peer-reviewed papers that formed the basis for the study was identified as fraudulent. That is not to say that judgments were not made about the credibility of submissions and that those that failed to meet that test were rejected. The telling sign was that a certain number of papers carried the notation, "CP," which in the editors' vernacular at the time, stood for "crackpot." These were rejected as such. (Some of these submissions, written in pencil on lined paper, are an indicator that our research was done a long time ago.) No effort was made by editors to determine whether these papers were fraudulent or just plain wrong; they were simply not publishable.

Until more is learned about the "demography" of journals that print retractions, including the disciplines they represent, the reasons cited for the retractions they publish, and who makes the decisions to retract, studies of misconduct based on retraction databases have to be considered as more suggestive than solid.

4.3. Technologies for Detection of FFP

The use of new technologies now available to detect misconduct has not yet (at least to my knowledge) been studied; nor has their effectiveness been evaluated. Among those frequently mentioned are plagiarism searches, statistical techniques for identifying data "too good to be true," and software that identifies photographs and graphic images that have been manipulated or edited. Plagiarism searches are said to be used by college instructors to test the originality of the papers their students submit. They are also used by some journals to identify submissions that contain plagiarized material. Statistical techniques of analysis can be and have been applied to data to identify those that are too good to be true: that are likely (but not proven) to have resulted from selective additions or subtractions. It is now possible to use software to detect whether images or graphs have been "doctored." Information on how to access these programs is available on the NIH website, which details its resources for protecting the

integrity of research. Determining how often these technologies are used, by whom, and in what circumstances, as well as the extent of their effectiveness in identifying research misconduct, would, it seems, provide some evidence about the incidence of misconduct they are intended to detect. But, of course, there is no way of knowing whether their current use has reduced levels of misconduct relative to misconduct occurring in the past, or whether their use has deterred misconduct that in their absence would otherwise have occurred. We do not know now how valid they have proved to be and how often their use has led to rejection of papers for publication, to retraction of flawed papers, or to the identification of otherwise unreliable research. Learning more about the effectiveness of these interventions would be useful.

5. CONSEQUENCES OF RESEARCH MISCONDUCT FOR CONFIDENCE IN SCIENCE

Continuing media attention to research misconduct, to the replication crisis, and now to the frequency of retractions of publications would seem ample reason for the public skepticism about the credibility of science. Media reports have surely led to increased anxiety among scientists about the public's willingness to support scientific and technological research (at least so scientists say). The evidence in hand, based on longitudinal surveys in the United States of public trust in science, however, suggests otherwise. Repetitive surveys of public confidence in science, over three decades or longer, show no major changes in measures of confidence. A substantial share of U.S. citizens say they have moderate or high levels of confidence both in science and in the leaders of the scientific community. Even larger shares say that they believe that scientists "work for the good of humanity" (National Science Board, 2018, Chapter 7). I will not comment on the finding that science in the United States rates second in public confidence only to the military among all occupational groups, and higher than physicians, the clergy, judges, politicians, and lawyers. (The significance of this finding is for the reader to decide.) It would be misleading, however, to fail to note the existence of vigorous antiscience sentiments among sectors of the U.S. population. "Antivaxxers," who believe that vaccinations against diseases are harmful, are numerous, especially among certain religious groups and in certain geographic areas. So are those who reject climate science as having no scientific basis. Those holding unconfirmed, even exotic, beliefs about diet and exercise are also numerous, judging from the retail sales of products they consume, and believers in astrology have long been a staple in the U.S. population. Whatever the sources of these sets of beliefs, research misconduct seems not to have intensified them; nor has it become a widespread *cause célèbre*. Despite the ample attention to misconduct in science in the press, public support for science seems not to have been affected.

6. WHAT LESSONS MIGHT BE LEARNED ABOUT RESEARCH MISCONDUCT AND ABOUT PROSPECTS FOR STUDYING IT QUANTITATIVELY?

1. Thinking about misconduct from an epidemiological perspective, that is, focusing on questions about its incidence and distribution, the effectiveness of means of controlling or remediating it, and careful analysis of the characteristics of "populations at risk" in given instances would help to frame questions worth asking, identify those that can be investigated and determine what needs to be done to make research possible. Adopting this perspective could guide the decisions that investigators make in using available databases or collecting new evidence for research. This was an observation I made back in 1977. It remains more an aspiration than a reality.
2. Using data in hand, as well as collecting new valid and reliable data on scientific "misconduct," is full of pitfalls, much like those that criminologists have encountered

in studying unlawful behavior. The public record on misconduct is biased. No evidence is available on those accused of it but who were exonerated or simply not judged guilty. Self-reports are questionable, as are eyewitness reports, for different reasons. Thus, efforts to estimate the causes of misconduct, its distribution, and its incidence are also very likely to be biased. However, the study of criminal behavior has a much longer history than the study of research misconduct. Careful study of methods already tried and obstacles encountered may benefit the study of research misconduct.

3. The controversy about whether misconduct is better described as the "tip of an iceberg" (and thus is very frequent) or as the result of random "bad apples" sullying the scientific barrel (and thus is very rare) cannot be resolved in light of the available information. The reasons that advocates give for taking one position or the other say less about misconduct and its incidence than about those who support them.
4. Despite major problems with data on misconduct, opportunities for quantitative analysis do exist and should be pursued. Just one example—while official data on the characteristics of scientists judged guilty of misconduct has its drawbacks (they are, after all, only the ones who were caught, out of some unknown larger group of miscreants), it is still possible to identify the characteristics of those who engaged in fabrication, falsification, or plagiarism (one or two or all three), by what means their misconduct was revealed, in what kinds of organizations misconduct occurred, how the investigatory procedures worked, and the gravity of the penalties imposed. Learning more about these cases should reveal how the oversight system works, if not about the epidemiology of misconduct, broadly conceived. Determining the nature of the questions that the available data can answer is part of the challenge.
5. Paying further attention to the consequences that misconduct has for the record of science would be salutary. This opens up a set of questions about the effects of misconduct that have hardly been touched. The line of inquiry would include further study of the rates and persistence of citations to publications judged to violate the FFP standard. What difference, if any, has such research made on to subsequent investigations? Eugene Garfield (1987a, 1987b) began to do this kind of study decades ago and his findings remain illuminating (Zuckerman, 2018). Other studies have begun to examine the consequences for the reputations of first authors and coauthored use of research on retracted papers, whatever the reasons given for retraction. There is some indication that citations accruing to first authors of retracted papers fall after retractions are announced, and there is also evidence that their coauthors suffer a kind of collateral damage (Azoulay, Bonatti, & Krieger, 2017). These findings are highly suggestive, but need far more confirmation. Apart from the effects on citations, serious work is needed on the consequences of misconduct and accusations of it for scientists' careers, over the longer as well as the shorter term, for the institutions in which they worked, for their coworkers and for the fate of the ideas they sought to promote.

Readers may find these comments about research misconduct and the prospects of doing quantitative research on it informative and even, perhaps, thought-provoking.

Quantitative Science Studies will be a great asset to the social study of science and to those of us who think science is an immensely significant part of social life.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

No funding was received for this research.

REFERENCES

Azoulay, P., Bonatti, A., & Krieger, J. L. (2017). The career effects of scandal: Evidence from scientific retractions. *Research Policy*, *46*(9), 1552–1569.

Ben-Yehuda, N., & Oliver-Lumerman, A. (2017). *Fraud and misconduct in research: Detection, investigation, and organizational response*. Ann Arbor, MI: University of Michigan Press.

Brainard, J., & You, J. (2018). What a massive database of retracted papers reveals about science publishing's "death penalty." *Science*, *25*(1), 1–5.

Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? *PNAS*, *115*(11), 2628–2631. www.pnas.org/cgi/doi/10.1073/pnas.1708272114

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714–3719.

Garfield, E. (1987a). What do we know about fraud and other forms of intellectual dishonesty in science? Part 1: The spectrum of deviant behavior in science. *Current Contents. Clinical Medicine*, *15*(14), 3.

Garfield, E. (1987b). What do we know about fraud and other forms of intellectual dishonesty in science? Part 2: Why does fraud happen and what are its effects? *Current Contents. Clinical Medicine*, *15*(15), 3–10.

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology*, *65*(6), 814–845. https://doi.org/10.1177/0011392116663807

Hixon, J. R. (1976). *The patchwork mouse*. Norwall, MA: Anchor Press.

Holton, G. (1994). "We do our damnedest": The evolution of trust in scientific findings. In D. H. Guston and K. Keniston (Eds.) *The fragile contract: University science and the federal government* (pp. 59–81). Cambridge, MA: MIT Press.

Horbach, S. P. J. M., & Halffman, W. (2019). The extent and causes of academic text recycling or "self-plagiarism." *Research Policy*, *48*(2), 492–502. https://doi.org/10.1016/j.respol.2017.09.004

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Ioannidis, J. P., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLOS Biology*, *13*(10).

Leahey, E., & Montgomery, K. (2011). The meaning of regulation in a changing academic profession. In J. C. Hermanowicz (Ed.), *The American academic profession: Changing forms and functions* (pp. 295–311). Baltimore, MD: Johns Hopkins Press.

National Science Board. (2018). *Science and engineering indicators 2018*. Washington, DC: National Science Foundation.

Oransky, I., & Marcus, A. (2010). Why write a blog about retractions. *Retraction Watch*. http://retractionwatch.com/2010/08/03/why-write-a-blog-about-retractions? Retrieved February 27, 2020.

Radick, G. (2015). Beyond the Mendel-Fisher controversy. *Science*, *350*(6257), 159–160. https://doi.org/10.1126/science.aab3846

Shiffrin, R. M., Börner, K., & Stigler, S. M. (2018). Scientific progress despite irreproducibility: A seeming paradox. *Proceedings of the National Academy of Sciences*, *115*(11), 2632–2639.

Steneck, N. H. (1994). Research universities and scientific misconduct: History, policies, and the future. *Journal of Higher Education*, *65*(3), 310–330.

Steneck, N. H., & Mayer, T. (2017). *A global guide to the responsible conduct of research*. London: Imperial College Press.

Wadman, M. (1996). Hostile reception to US misconduct report. *Nature*, *381*, 639.

Zuckerman, H. (1977). Deviant behavior and social control in science. In E. Savarin (Ed.), *Deviance and social change* (pp. 87–138). Beverly Hills, CA: Sage.

Zuckerman, H. (1984). Norms and deviant behavior in science. *Science, Technology, & Human Values*, *9*(1), 7–13.

Zuckerman, H. (2018). The sociology of science and the Garfield effect: Happy accidents, unanticipated developments and unexploited potentials. *Frontiers in Research Metrics and Analytics*, *3*, 20. https://doi.org/10.3389/frma.2018.00020

Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalization, structure and functions of the referee system. *Minerva*, *9*(1), 66–100.