COLLOQUIUM OPINION

Is science really facing a reproducibility crisis, and do we need it to?

Daniele Fanelli[a,1]

Edited by David B. Allison, Indiana University Bloomington, Bloomington, IN, and accepted by Editorial Board Member Susan T. Fiske November 3, 2017 (received for review June 30, 2017)

Efforts to improve the reproducibility and integrity of science are typically justified by a narrative of crisis, according to which most published results are unreliable due to growing problems with research and publication practices. This article provides an overview of recent evidence suggesting that this narrative is mistaken, and argues that a narrative of epochal changes and empowerment of scientists would be more accurate, inspiring, and compelling.

reproducible research | crisis | integrity | bias | misconduct

Is there a reproducibility crisis in science? Many seem to believe so. In a recent survey by the journal *Nature*, for example, around 90% of respondents agreed that there is a "slight" or "significant" crisis, and between 40% and 70% agreed that selective reporting, fraud, and pressures to publish "always" or "often" contribute to irreproducible research (1). Results of this nonrandomized survey may not accurately represent the population of practicing scientists, but they echo beliefs expressed by a rapidly growing scientific literature, which uncritically endorses a new "crisis narrative" about science (an illustrative sample of this literature is shown in Fig. 1 and listed in Dataset S1).

Put simply, this new "science in crisis" narrative postulates that a large and growing proportion of studies published across disciplines are unreliable due to the declining quality and integrity of research and publication practices, largely because of growing pressures to publish and other ills affecting the contemporary scientific profession.

I argue that this crisis narrative is at least partially misguided. Recent evidence from metaresearch studies suggests that issues with research integrity and reproducibility, while certainly important phenomena that need to be addressed, are: (*i*) not distorting the majority of the literature, in science as a whole as well as within any given discipline; (*ii*) heterogeneously distributed across subfields in any given area, which suggests that generalizations are unjustified; and (*iii*) not growing, as the crisis narrative would presuppose. Alternative narratives, therefore, might represent a better fit for empirical data as well as for the reproducibility agenda.

How Common Are Fabricated, False, Biased, and Irreproducible Findings?

Scientific misconduct and questionable research practices (QRP) occur at frequencies that, while nonnegligible, are relatively small and therefore unlikely to have a major impact on the literature. In anonymous surveys, on average 1–2% of scientists admit to having fabricated or falsified data at least once (2). Much higher percentages admit to other QRP, such as dropping data points based on a gut feeling or failing to publish a contradictory result. However, the percentage of scientific literature that is actually affected by these practices is unknown, and evidence suggests that it is likely to be smaller, at least five times smaller according to a survey among psychologists (3). Data that directly estimate the prevalence of misconduct are scarce but appear to corroborate this conclusion. Random laboratory audits in cancer clinical trials, for example, found that only 0.28% contained "scientific improprieties" (4), and those conducted among Food and Drug Administration clinical trials between 1977 and 1988 found problems sufficient to initiate "for cause"

[a]Department of Methodology, London School of Economics and Political Science, London WC2A 2AE, United Kingdom

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: D.F. wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission. D.B.A. is a guest editor invited by the Editorial Board.

[1]Email: email@danielefanelli.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1708272114/-/DCSupplemental.

Published online March 12, 2018.

investigations only in 4% of cases (5). Visual inspections of microbiology papers suggested that between 1% and 2% of papers had been manipulated in ways that suggested intentional fabrication (6, 7).

The occurrence of questionable or flawed research and publication practices may be revealed by a high rate of false-positives and "*P*-hacked" (8) results. However, while these issues do appear to be more common than outright scientific misconduct, their impact on the reliability of the literature appears to be contained. Analyses based on the distribution of *P* values reported in the medical literature, for example, suggested a false-discovery rate of only 14% (9). A similar but broader analysis concluded that *P*-hacking was common in many disciplines and yet had minor effects in distorting conclusions of meta-analyses (10). Moreover, the same analysis found a much stronger "evidential value" in the literature of all disciplines, which suggests that the majority of published studies are measuring true effects, a finding that again contradicts the belief that most published findings are false-positives. Methodological criticisms suggest that these and similar studies may be underestimating the true impact of *P*-hacking (11, 12). However, to the best of my knowledge, there is no alternative analysis that suggests that *P*-hacking is severely distorting the scientific literature.

Low statistical power might increase the risk of false-positives (as well as false-negatives). In several disciplines, the average statistical power of studies was found to be significantly below the recommended 80% level (13–16). However, such studies showed that power varies widely between subfields or methodologies, which should warn against making simplistic generalizations to entire disciplines (13–16). Moreover, the extent to which low power generates false-positives depends on assumptions about the magnitude of the true underlying effect size, level of research bias, and prior probabilities that the tested hypothesis is in fact correct (17). These assumptions, just like statistical power itself, are likely to vary substantially across subfields and are very difficult to measure in practice. For most published research findings to be false in psychology and neuroscience, for example, one must assume that the hypotheses tested in these disciplines are correct much less than 50% of the time (14, 18). This assumption is, in my opinion, unrealistic. It might reflect the condition of early exploratory studies that are conducted in a theoretical and empirical vacuum, but not that of most ordinary research, which builds upon previous theory and evidence and therefore aims at relatively predictable findings.

It may be counter-argued that the background literature that produces theory and evidence on which new studies are based is distorted by publication and other reporting biases. However, the extent to which this is the case is, again, likely to vary by research subfield. Indeed, in a meta-assessment of bias across all disciplines, small-study effects and gray-literature bias (both possible symptoms of reporting biases) were highly heterogeneously distributed (19). This finding was consistent with evidence that studies on publication bias may themselves be subject to a publication bias (20), which entails that fields that do not suffer from bias are underrepresented in the metaresearch literature.

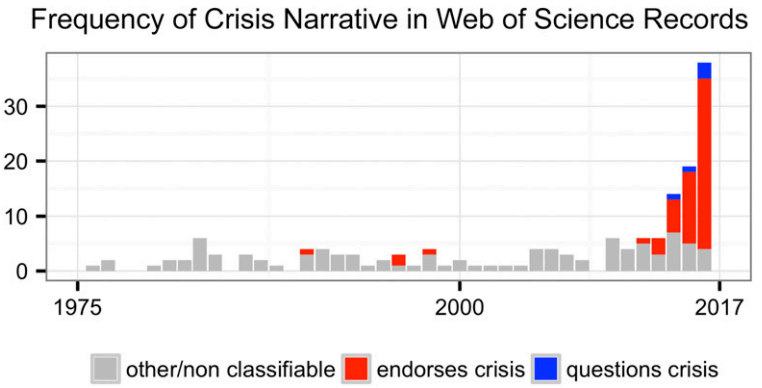

Fig. 1. Number of Web of Science records that in the title, abstract, or keywords contain one of the following phrases: "reproducibility crisis," "scientific crisis," "science in crisis," "crisis in science," "replication crisis," "replicability crisis." Records were classified by the author according to whether, based on title and abstracts, they implicitly or explicitly endorsed the crisis narrative described in the text (red), or alternatively questioned the existence of such a crisis (blue), or discussed "scientific crises" of other kinds or could not be classified due to insufficient information (gray). The complete dataset, which includes all titles and abstracts and dates back to the year 1933, is available in Dataset S1. This sample is merely illustrative, and does not include the numerous recent research articles and opinion articles that discuss the "science is in crisis" narrative without including any of the above sentences in the title, abstract, or keywords.

The case that most publications are nonreproducible would be supported by meta-meta-analyses, if these had shown that on average there is a strong "decline effect," in which initially strong "promising" results are contradicted by later studies. While a decline effect was measurable across many meta-analyses, it is far from ubiquitous (19). This suggests that in many meta-analyses, initial findings are refuted, whereas in others they are confirmed. Isn't this what should happen when science is functional?

Ultimately, the debate over the existence of a reproducibility crisis should have been closed by recent large-scale assessments of reproducibility. Their results, however, are either reassuring or inconclusive. A "Many labs" project reported that 10 of 13 studies taken from the psychological literature had been consistently replicated multiple times across different settings (21), whereas an analysis in experimental economics suggested that, of 18 studies, at least 11 had been successfully replicated (22). The largest reproducibility initiative to date suggested that in psychological science, reproducibility was below 50% (23). This latter estimate, however, is likely to be too pessimistic for at least two reasons. First because, once again, such a low level of reproducibility was not ubiquitous but varied depending on subfield, methodology, and expertise of the authors conducting the replication (23–25). Second, and more importantly, because how reproducibility ought to be measured is the subject of a growing methodological and philosophical debate, and reanalyses of the data suggest that reproducibility in psychological science might

be higher than originally claimed (23, 26, 27). Indeed, the very notion of "reproducible research" can be confusing, because its meaning and implications depend on what aspect of research is being examined: the reproducibility of research methods can in principle be expected to be 100%; but the reproducibility of results and inferences is likely to be lower and to vary across subfields and methodologies, for reasons that have nothing to do with questionable research and publication practices (28).

Are These Problems Getting Worse?

In light of multiple recent studies, there is no evidence that scientific misconduct and QRPs have increased. The number of yearly findings of scientific misconduct by the US Office of Research Integrity (ORI) has not increased, nor has the proportion, of all ORI investigations, that resulted in a finding of misconduct (29). Retractions have risen sharply in absolute terms, but the number of retractions per retracting journals has not, suggesting that the trend is due to the diffusion and improvement of journal retraction policies and practices (29). Errata and corrections have also not increased, nor has the rate of statistical errors made in mainstream psychological journals (29, 30).

The questionable practice known as "salami-slicing," in which results are fractionalized to increase publication output, is widely believed to be on the rise. However, there is no evidence that scientists are publishing more papers today than in the 1950s, once coauthorship is adjusted for (31). Indeed, assessments in various disciplines suggest that, far from becoming increasingly short and trivial, published studies are getting longer, more complex, and richer in data (e.g., refs. 32–34).

Biases in research and reporting were suggested to be on the rise by multiple independent studies, which had found that the relative proportion of "positive" and "statistically significant" results reported in article abstracts has increased over the years (35–37). However, the aforementioned evidence that papers in many (and maybe most) disciplines are becoming longer and more complex suggests that negative results may not be disappearing from the literature, as originally suggested, but perhaps only from abstracts. Negative results, in other words, may be increasingly embedded in longer publications that contain multiple results, and they therefore remain accessible to any researcher interested in finding them.

Finally, pressures to publish have not been convincingly linked to evidence of bias or misconduct. Earlier studies that compared the scientific productivity of countries offered some support for such a link (38, 39). However, later, finer-grained analyses offered contrary evidence, by showing that researchers that publish at higher frequency, in journals with higher impact factor, and in countries where pressures to publish are high, are equally or more likely to correct their work, less likely to publish papers that are retracted, less likely to author papers that contain duplicated images, and less likely to author papers reporting overestimated effects (19, 40, 41). The risk of misconduct and QRPs appears to be higher among researchers in countries that are increasingly represented in the global scientific literature, like China or India (7, 40). Global demographic changes, therefore, might contribute to a rise in the proportion of papers affected by scientific misconduct, but such a trend would have little to do with rising pressures to publish in Western countries.

Do We Need the "Science in Crisis" Narrative to Promote Better Science?

To summarize, an expanding metaresearch literature suggests that science—while undoubtedly facing old and new challenges—cannot be said to be undergoing a "reproducibility crisis," at least not in the sense that it is no longer reliable due to a pervasive and growing problem with findings that are fabricated, falsified, biased, underpowered, selected, and irreproducible. While these problems certainly exist and need to be tackled, evidence does not suggest that they undermine the scientific enterprise as a whole. Science always was and always will be a struggle to produce knowledge for the benefit of all of humanity against the cognitive and moral limitations of individual human beings, including the limitations of scientists themselves.

The new "science is in crisis" narrative is not only empirically unsupported, but also quite obviously counterproductive. Instead of inspiring younger generations to do more and better science, it might foster in them cynicism and indifference. Instead of inviting greater respect for and investment in research, it risks discrediting the value of evidence and feeding antiscientific agendas.

Furthermore, this narrative is not actually new. Complaints about a decline in the quality of research recur throughout the history of science, right from its beginnings (42, 43). Only two elements of novelty characterize the current "crisis." The first is that the validity of these concerns is being assessed scientifically by a global metaresearch program, with results that have been briefly overviewed above (44). The second element of historical novelty is the rising power of information and communication technologies, which are transforming scientific practices in all fields, just as they are transforming all other aspects of human life. These technologies promise to make research more accurate, powerful, open, democratic, transparent, and self-correcting than ever before. At the same time, this technological revolution creates new expectations and new challenges that metaresearchers are striving to address.

Therefore, contemporary science could be more accurately portrayed as facing "new opportunities and challenges" or even a "revolution" (45). Efforts to promote transparency and reproducibility would find complete justification in such a narrative of transformation and empowerment, a narrative that is not only more compelling and inspiring than that of a crisis, but also better supported by evidence.

1 Baker M (2016) Is there a reproducibility crisis? *Nature* 533:452–454.
2 Fanelli D (2009) How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One* 4:e5738.
3 Fiedler K, Schwarz N (2016) Questionable research practices revisited. *Soc Psychol Personal Sci* 7:45–52.
4 Weiss RB, et al. (1993) A successful system of scientific data audits for clinical trials. A report from the Cancer and Leukemia Group B. *JAMA* 270:459–464.
5 Shapiro MF, Charrow RP (1989) The role of data audits in detecting scientific misconduct. Results of the FDA program. *JAMA* 261:2505–2511.
6 Steneck NH (2006) Fostering integrity in research: Definitions, current knowledge, and future directions. *Sci Eng Ethics* 12:53–74.
7 Bik EM, Casadevall A, Fang FC (2016) The prevalence of inappropriate image duplication in biomedical research publications. *MBio* 7:e00809–e00816.
8 Wicherts JM, et al. (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Front Psychol* 7:12.
9 Jager LR, Leek JT (2014) An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15:1–12.
10 Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The extent and consequences of P-hacking in science. *PLoS Biol* 13:e1002106.
11 Bruns SB, Ioannidis JPA (2016) p-curve and p-hacking in observational research. *PLoS One* 11:e0149144.
12 Bishop DVM, Thompson PA (2016) Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ* 4:e1715.
13 Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafo MR (2017) Low statistical power in biomedical science: A review of three human research domains. *R Soc Open Sci* 4:160254.
14 Szucs D, Ioannidis JP (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol* 15:e2000797.
15 Jennions MD, Moller AP (2003) A survey of the statistical power of research in behavioral ecology and animal behavior. *Behav Ecol* 14:438–445.
16 Fraley RC, Vazire S (2014) The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One* 9:e109019.
17 Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2:e124.
18 Button KS, et al. (2013) Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376.
19 Fanelli D, Costas R, Ioannidis JPA (2017) Meta-assessment of bias in science. *Proc Natl Acad Sci USA* 114:3714–3719.
20 Dubben HH, Beck-Bornholdt HP (2005) Systematic review of publication bias in studies on publication bias. *BMJ* 331:433–434.
21 Klein RA, et al. (2014) Investigating variation in replicability. A "Many labs" replication project. *Soc Psychol* 45:142–152.
22 Camerer CF, et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351:1433–1436.
23 Open Science Collaboration (2015) Psychology. Estimating the reproducibility of psychological science. *Science* 349:aac4716.
24 Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA (2016) Contextual sensitivity in scientific reproducibility. *Proc Natl Acad Sci USA* 113:6454–6459.
25 Bench SW, Rivera GN, Schlegel RJ, Hicks JA, Lench HC (2017) Does expertise matter in replication? An examination of the reproducibility project: Psychology. *J Exp Soc Psychol* 68:181–184.
26 Patil P, Peng RD, Leek JT (2016) What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect Psychol Sci* 11:539–544.
27 Etz A, Vandekerckhove J (2016) A Bayesian perspective on the reproducibility project: Psychology. *PLoS One* 11:e0149794.
28 Goodman SN, Fanelli D, Ioannidis JPA (2016) What does research reproducibility mean? *Sci Transl Med* 8:341ps312.
29 Fanelli D (2013) Why growing retractions are (mostly) a good sign. *PLoS Med* 10:e1001563.
30 Nuijten MB, Hartgerink CHJ, van Assen MALM, Epskamp S, Wicherts JM (2016) The prevalence of statistical reporting errors in psychology (1985-2013). *Behav Res Methods* 48:1205–1226.
31 Fanelli D, Larivière V (2016) Researchers' individual publication rate has not increased in a century. *PLoS One* 11:e0149504.
32 Rodriguez-Esteban R, Loging WT (2013) Quantifying the complexity of medical research. *Bioinformatics* 29:2918–2924.
33 Vale RD (2015) Accelerating scientific publication in biology. *Proc Natl Acad Sci USA* 112:13439–13446.
34 Low-Décarie E, Chivers C, Granados M (2014) Rising complexity and falling explanatory power in ecology. *Front Ecol Environ* 12:412–418.
35 Pautasso M (2010) Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics* 85:193–202.
36 Fanelli D (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics* 90:891–904.
37 de Winter JCF, Dodou D (2015) A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 3:e733.
38 Munafò MR, Attwood AS, Flint J (2008) Bias in genetic association studies: Effects of research location and resources. *Psychol Med* 38:1213–1214.
39 Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US States data. *PLoS One* 5:e10271.
40 Fanelli D, Costas R, Fang FC, Casadevall A, Bik EM (2017) Why do scientists fabricate and falsify data? A matched-control analysis of papers containing problematic image duplications. *bioRxiv*, 10.1101/126805.
41 Fanelli D, Costas R, Larivière V (2015) Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS One* 10:e0127556.
42 Mullane K, Williams M (2017) Enhancing reproducibility: Failures from reproducibility initiatives underline core challenges. *Biochem Pharmacol* 138:7–18.
43 Babbace C (1830) Reflections of the decline of science in England, and on some of its causes. Available at https://archive.org/details/reflectionsonde00mollgoog. Accessed November 29, 2017.
44 Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN (2015) Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biol* 13:e1002264.
45 Spellman BA (2015) A short (personal) future history of revolution 2.0. *Perspect Psychol Sci* 10:886–899.