UC Irvine
UC Irvine Previously Published Works

Title
Dialogues about the practice of science

Permalink
https://escholarship.org/uc/item/4wg265qv

Journal
Proceedings of the National Academy of Sciences of the United States of America, 122(5)

ISSN
0027-8424

Authors
Shiffrin, Richard M
Trueblood, Jennifer S
Kellen, David
et al.

Publication Date
2025-02-04

DOI
10.1073/pnas.2423782122

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

PNAS

SPECIAL FEATURE: INTRODUCTION

OPEN ACCESS

Dialogues about the practice of science

Richard M. Shiffrin[a], Jennifer S. Trueblood[a,b], David Kellen[c], and Joachim Vandekerckhove[d,e,f,1]

As times change, institutions adapt. The practice of science is no exception, and these are fast-changing times. Individual scientists tend to approach science in the ways they have been trained and ways that have advanced their careers in the increasingly narrow domains in which they specialize. Many are too busy to step back, look at bigger pictures, and consider the merits of the way science is practiced. In this Special Feature, we ask: Should scientists reform the way they carry out their research, and if so, in what ways? Should we work to reform the institutions that currently serve as scientific gatekeepers? How do recent technological and cultural revolutions affect the practice of science? The current perspectives offer alternative views from a large number of practicing scientists, active in a variety of fields and at diverse stages of their careers.

Consideration of practices that transcend individual disciplines has traditionally been the domain of philosophers of science (e.g., ref. 1), and occasionally the subject of reports by groups of scientists (e.g., ref. 2). The present-day need for revisiting the best ways to practice science arises partly from what has become known as the "reproducibility crisis," partly from rapid advances in intelligent automation (AI and machine learning methods), and partly from the smothering proliferation of scientific reports and scientific outlets.

The perspectives in this Special Feature do not present a case for a particular position or consensus opinion on any given issue. The question of how to practice science calls for a different approach because our usual scientific methods are difficult to apply—science relies on observational data gathered in natural environments, controlled experiments designed to tease out causal factors, and rigorous reflection to connect data to theory. Observational data about the practice of science are available only in a limited sense to each scientist, who typically observes their own practices and those of their colleagues (usually in the same areas of research), leading to limited generalization. Similar limitations hold for observational data gathered by historians, philosophers, or sociologists of science, as the meaning of what they observe is constrained by the selection of historical research they examine and the potential errors within it. If observational analyses are insufficient to determine how scientists should practice, could controlled studies be more effective? It is difficult to imagine proper controls to assess scientific practices. Science is a collaborative effort among scientists worldwide. Even if it were possible to have different subsets of scientists adopt different practices, they would need to address the same issues, meaning their research would not be independent. Even if we ignored this lack of independence, evaluating the success of each group or determining appropriate evaluation criteria would be challenging. The problem of proper controls arises in many scientific domains involving complex systems—such as climate change and cancer research—yet the practice of science remains an even more formidable target of empirical investigation.

Beyond the difficulty of determining best practices empirically, we believe there cannot be a single "best" practice, as different practices suit different goals, values, and desiderata within scientific research. One set of practices might aim to satisfy a "mission-based" goal of immediate application. Such goals are paramount when research is used to justify a societal change—for example, to approve a drug treatment, initiate a vaccination program, make public investments, design ads, and search engines that maximize profits, or regulate media to protect the public from harm. Different practices might be more suitable when the goal is to enhance our understanding of the natural or human world—such as forming causal models of data that explain individual and group behavior, health, or the workings of the universe. Other goals may include developing engineering applications; solving health problems; strengthening the national defense; advancing ethical and diversity goals; improving methodology, experimental design, transparency of reporting, and statistical analysis; and contributing to a fair and democratic society. The variety of these goals is likely one cause of debate about the merits of different practices. As an important aside, we note that a critically important goal involves communicating results to different groups (experts, scientists, students, the lay public, children) and fostering public understanding and acceptance of science. These goals have been addressed in numerous reports by the National Academy of Sciences and other organizations and will not be addressed here.

Given these complexities, the contributing teams were not asked to reach consensus conclusions about the best ways to practice science. Instead, they discussed alternative practices and their likely implications, along with different methods and their pros and cons. A recurring theme is that different practices may be appropriate depending on the specific goals being pursued.

Author affiliations: [a]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405; [b]Cognitive Science Program, Indiana University, Bloomington, IN 47405; [c]Department of Psychology, Syracuse University, Syracuse, NY 13244; [d]Department of Cognitive Sciences, University of California, Irvine, CA 92697; [e]Department of Statistics, University of California, Irvine, CA 92697; and [f]Department of Logic and Philosophy of Science, University of California, Irvine, CA 92697

R.M.S., J.S.T., D.K., and J.V. are organizers of this Special Feature.

Author contributions: R.M.S., J.S.T., D.K., and J.V. wrote the paper.

The authors declare no competing interest.

Copyright © 2025 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: joachim+pnas@uci.edu.

Published January 27, 2025.

The Crisis

The term "reproducibility crisis" emerged from striking demonstrations showing that many published findings—some in clinical research (3) and some in the social sciences (4, 5)—could not be reproduced. At the same time, several high-profile publications highlighted shortcomings in common methodological practices. Ioannidis (6) showed that weak standards of evidence imply that many published findings are unreliable. Simmons et al. (7) pointed out that methodological and statistical flexibility allow any set of results to be presented as significant. A mediagenic publication claiming extrasensory perception (8) became a notorious example of how state-of-the-art practices can nevertheless lead to unreasonable or even absurd conclusions.

These challenges to existing practices combine with a variety of known biases and dysfunctional incentive structures. Particularly prevalent forms of bias include defending one's results and theories as if they were personal possessions (belief bias), seeking out experimental support at the expense of truly critical tests (confirmation bias), and selectively publishing only statistically significant findings (publication bias). To at least some degree these biases are omnipresent (e.g., refs. 9–12); they are often unintentional, unconscious, and therefore difficult to avoid.

Other causes of results and claims that are difficult to replicate, or are often misleading, include simple errors (e.g., in analysis, programming, reporting, or conclusions) and poor scientific practices (e.g., poor design, lack of control, failure to report what was done). Beyond these obvious causes of replication failure are so-called "questionable research practices," which are subject to debate. Both types of practices contributed to the formation of the "Open Science" movement (e.g., the Center for Open Science and the UNESCO Recommendation on Open Science). The term "open science" includes an uncontroversial component of scientific practice: the need to communicate one's designs, procedures, results, models, and simulations accurately and thoroughly. This component has not always been followed, and the open science movement plays an important role in rectifying what is widely agreed to be a necessary component of good practice. Because there is general agreement on this issue, we do not include its discussion here. Less clear is the importance of replication and reproducibility as goals for determining good practice (13).

Reports of failures to replicate published claims are far from new, and go back to the beginnings of science. The numbers of such failures are surely large even if attempts at replication have historically been rare across fields (14). An early example was the report of N-rays (a hypothesized novel type of radiation) in 1903 by Blondlot. N-rays were subsequently supported by some researchers but the findings could not be replicated by others. A definitive explanation—in terms of experimenter bias—waited until Wood (15) showed that "believers" in N-rays continued to find them even after the equipment needed to do so had been made unworkable (see ref. 16). Sometimes nonreplicable findings are harmful, other times harmless, and occasionally useful by leading researchers to improve methodology. For example, in the 1960s, Weber reported the direct experimental detection of gravity waves (17). Although these results could not be replicated, scientists pursuing this finding learned that the equipment used in the original study lacked the necessary precision. Knowledge of the precision needed played a role in the design of the successful LIGO experiments in 2015.

The prevalence of published claims that cannot be reproduced or verified can be interpreted in quite different ways: Some would conclude that reform is needed (e.g., refs. 18 and 19), while others argue that consumers of the scientific literature should always be skeptical concerning reported data and theoretical accomplishments (e.g., refs. 20 and 21). Yet others believe that large swaths of science are in trouble and cannot be trusted (e.g., ref. 22). Are failures of reproducibility now more prevalent or less acceptable than in the past? Should we aim to reduce or eliminate these failures? Is nonreproducibility something to be eliminated or, at least, reduced from present levels? Or is a certain amount of replication failure, perhaps at current levels, necessary to "grease the wheels" of science (20)? These questions do not have clear answers. Movements have begun to change scientific practices, based on the belief that change is needed and that such changes will improve science (e.g., ref. 23). Should we encourage or discourage these movements going forward?

Overview of the Special Feature

The movement to reform scientific practice aimed at improving reproducibility and replicability raises broader questions about scientific practice as a whole. This was the driving force behind the creation of this Special Feature. The organizers were naturally led to pursue these questions because they have been engaged in debates about best practices in science for some time (e.g., refs. 20, 21, and 24–39). The immediate precursor to this feature was a six-day conference in the summer of 2023, where attendees discussed around 50 potential topics before selecting eight for production and submission.

While all eight perspectives relate to the reproducibility crisis in some way, only one addresses it directly. Others address critically important issues that impact the practice of science. Two perspectives address foundational issues in science: measurement and parsimony of explanation. Another three perspectives discuss issues related to scientific gatekeeping by funding agencies, peer review processes, and journals. The remaining perspectives discuss the impact of automation, machine learning, and AI on the practice of science.

Replication and the Reproducibility Crisis. The authors of "How can we make sound replication decisions?" (40) offer a conceptual framework for making sound decisions about whether and how to replicate scientific findings. They explore the role of both epistemic and nonepistemic values, and introduce a discussion of the cognitive attitudes scientists may hold when deciding on replication efforts. These cognitive attitudes lie on a spectrum between a "Book of Truths" (the scientific literature serves mostly to disseminate true claims) and a "Book of Conversations" (the scientific literature serves mostly to exchange ideas between experts). The broader conceptual framework can help guide informed and context-sensitive replication decisions based on alignment between scientific values, cognitive attitudes, and experimental designs.

Foundations for Scientific Progress. Empirical progress relies on the ability to perform accurate measurements. This essential skill underpins scientific reasoning and experimental design—it enables scientists to communicate effectively and identify misleading or unsupported quantitative claims. Problems with measurement can hinder the establishment of scientific discoveries. Once a discovery is made, its verification is crucial for scientific progress. Replication is a widely used method for scientific verification, but it has limitations that can sometimes impede progress. In "Discourse on measurement" (41), the authors emphasize the critical role of measurement in the sciences, providing recent examples of measurement misuse and its detrimental impact on scientific progress. They argue that expertise in measurement literacy is not only fundamental for effective reasoning and making robust scientific claims but also instrumental in the development of successful experimental designs and critical tests of theories. The authors also discuss the problems that can arise when foundational measurement issues are neglected in scientific education.

In "Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony" (42), the authors reevaluate Ockham's razor—the principle of parsimony—in the context of modern scientific modeling. Traditionally, parsimonious models are prized for their greater interpretability, better generalization to new data, usefulness in guiding research, and lower computational requirements. The paper explores how recent advances, such as highly complex models used in protein folding and climate forecasting, challenge the historical preference for simpler models. The authors distinguish between two forms of parsimony: parsimony by components (preferring fewer meaningful elements in models) and parsimony by constraints (preferring less flexibility in models). They argue that parsimony is not always the best guide for scientific progress and that complexity and simplicity can play complementary roles in modeling practice.

Scientific Gatekeeping. The creation and dissemination of scientific knowledge rely on two key elements: financial support for research and effective communication mechanisms for sharing that research. Institutions that facilitate these activities act as gatekeepers of science: Projects without funding may never materialize, and the peer review and publication processes determine which research contributions become part of the scientific record.

The authors of "Alternative models of funding curiosity-driven research" (43) examine the challenges associated with existing funding models, highlighting disparities based on gender and geographic regions, while emphasizing the importance of funding in fostering scientific and technological innovation. The paper critiques current funding schemes and discusses a number of reform proposals—both incremental and radical—including distributed funding, partial lotteries, and the concept of a World Research Council.

In "The present and future of peer review: ideas, interventions, and evidence" (44), the authors examine the integrity and sustainability of the current peer review system. They discuss its current problems, such as biases and inefficiency, and propose possible reforms, such as open and transparent review processes, having reviewers sign their reviews, and publishing reviews alongside the corresponding papers.

In "The misalignment of incentives in academic publishing and implications for journal reform" (45), the authors discuss the problems created by current incentive structures in academic publishing. They note that for many researchers, academic publishing serves the dual purposes of spreading knowledge and building scientific reputations, which are often at odds. Commercial publishers have capitalized on the essential role of publishing in both areas to derive substantial profits from the academic sector. The authors explore alternative models for publication and academic evaluation that aim to realign these incentives.

Impacts of Machine Learning and AI. Rapid advancements in machine learning (ML) and AI are transforming not only society generally but scientists' approach to developing scientific theories and many of the methods used to practice science. ML techniques are challenging traditional criteria for evaluating models, such as the principle of parsimony. Large language models (LLMs) hold the potential to revolutionize the production and consumption of scientific text. More broadly, AI is set to automate various stages of the scientific process, from hypothesis generation to experimentation.

In "How should the advancement of large language models affect the practice of science?" (46), the authors present perspectives from four different groups of scientists on the integration of LLMs into scientific workflows. Opinions vary from viewing LLMs as akin to human collaborators to expressing skepticism about their potential to enhance scientific practices. Schulz et al. argue that using LLMs is akin to working with human collaborators, while Bender et al. emphasize their limitations and risks. Marelli et al. discuss the importance of transparency and responsible use, while Botvinick and Gershman argue that core aspects of scientific decision-making should remain under human control.

Finally, the perspective "Automating the practice of science—opportunities, challenges, and implications" (47) discusses how automation is reshaping science. The authors highlight AI's capacity to overcome human cognitive limitations, thereby accelerating and broadening our discovery capabilities. They provide a thorough overview of specific ways in which automation can transform various aspects of scientific inquiry, from hypothesis generation to experimental design, data collection, and model discovery. The authors discuss both the potential benefits, such as accelerating discovery and enhancing reproducibility, and the limitations and ethical considerations of automation in science.

Conclusion

The practice of science is undergoing significant transformations driven by technological advancements, cultural shifts, and a growing awareness of the limitations and challenges inherent in traditional institutions and methodologies. The eight perspectives in this Special Feature do not and could not cover the vast domain of proper scientific practice, especially since "best" practice—if such a thing exists—varies by discipline and subdiscipline, and between scientists, as different approaches suit different goals, contexts, and values.

The discussions emphasize the importance of flexibility, pluralism, and critical reflection in scientific practice. From the reproducibility crisis to the integration of ML and AI, scientists are called to rethink not only the methods they use but also the underlying assumptions, goals, and values that guide their research. Issues such as measurement literacy, the principle of parsimony, and the role of scientific gatekeeping are central to ensuring that scientific progress remains robust, transparent, and meaningful.

A key takeaway from these perspectives is that science is not a static endeavor but a dynamic process that adapts to new challenges and opportunities. It is our hope that these essays will be of interest to many scientists and spark even more debate and thought about these critical issues than has already appeared in journals, blogs, and public dissemination.

ACKNOWLEDGMENTS. J.S.T. was supported by NSF grants #1846764, #2242962, and #2305559, and the Alfred P. Sloan Foundation. D.K. was supported by NSF grant #2145308. J.V. was supported by NSF grant #2051186.

1. K. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge* (Routledge, 1963).
2. National Academies of Sciences, Engineering, and Medicine *et al*., *Reproducibility and Replicability in Science* (National Academies Press (US), Washington, DC, 2019).
3. T. M. Errington *et al*., Investigating the replicability of preclinical cancer biology. *eLife* **10**, e71601 (2021).
4. H. Pashler, C. R. Harris, Is the replicability crisis overblown? Three arguments examined *Perspect. Psychol. Sci.* **7**, 531–536 (2012).
5. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
6. J. P. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
7. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
8. D. J. Bem, Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* **100**, 407 (2011).
9. T. D. Sterling, Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *J. Am. Stat. Assoc.* **54**, 30–34 (1959).
10. R. Rosenthal, L. Jacobson, Teachers' expectancies: Determinants of pupils' IQ gains. *Psychol. Rep.* **19**, 115–118 (1966).
11. R. Rosenthal, D. B. Rubin, Interpersonal expectancy effects: The first 345 studies. *Behav. Brain Sci.* **1**, 377–386 (1978).
12. R. Rosenthal, "Interpersonal expectations: Effects of the experimenter's hypothesis" in *Artifacts in Behavioral Research*, R. Rosenthal, R. L. Rosnow, Eds. (Cambridge University Press, 2009), pp. 138–210.
13. E. O. Buzbas, B. Devezer, B. Baumgaertner, The logical structure of experiments lays the foundation for a theory of reproducibility. *R. Soc. Open Sci.* **10**, 221042 (2023).
14. M. C. Makel, J. A. Plucker, B. Hegarty, Replications in psychology research: How often do they really occur? *Perspect. Psychol. Sci.* **7**, 537–542 (2012).
15. R. Wood, The n-rays. *Nature* **70**, 530–531 (1904).
16. S. Novella, *The Skeptics' Guide to the Universe* (Hodder & Stoughten, 2018).
17. J. Weber, Evidence for discovery of gravitational radiation. *Phys. Rev. Lett.* **22**, 1320–1324 (1969).
18. B. A. Nosek, C. R. Ebersole, A. C. DeHaven, D. T. Mellor, The preregistration revolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2600–2606 (2018).
19. E. J. Wagenmakers, R. Wetzels, D. Borsboom, H. L. J. van der Maas, R. A. Kievit, An agenda for purely confirmatory research. *Perspect. Psychol. Sci.* **7**, 632–638 (2012).
20. R. M. Shiffrin, K. Borner, S. Stigler, Scientific progress despite irreproducibility: A seeming paradox. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2639–2672 (2018).
21. R. M. Shiffrin, S. Stigler, K. H. Jamieson, "Evaluating evidence requires expert judgment" in *Evidence: The Use and Misuse of Data, Transactions of the American Philosophical Society*, R. M. Hauser, A. Link, Eds. (American Philosophical Society, 2023), vol. 112.
22. T. Yarkoni, The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
23. E. J. Wagenmakers, R. Wetzels, D. Borsboom, H. L. J. van der Maas, Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem (2011). *J. Pers. Soc. Psychol.* **100**, 426–432 (2011).
24. B. Alberts *et al*., Self-correction in science at work: Improve incentives to support research integrity. *Science* **348**, 1420–1422 (2015).
25. D. B. Allison, R. M. Shiffrin, V. Stodden, Reproducibility of research: Issues and proposed remedies. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2561–2562 (2018).
26. B. Devezer, D. J. Navarro, J. Vandekerckhove, E. O. Buzbas, The case for formal methodology in scientific reform. *R. Soc. Open Sci.* **8**, 200805 (2021).
27. A. Etz, J. Vandekerckhove, A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS One* **11**, e0149794 (2016).
28. D. Kellen, G. E. Cox, C. Donkin, J. C. Dunn, R. M. Shiffrin, Against naïve induction from experimental data. *Behav. Brain Sci.* **47**, e51 (2024).
29. M. D. Lee *et al*., Robust modeling in cognitive science. *Comput. Brain & Behav.* **2**, 141–153 (2019).
30. R. M. Shiffrin, Drawing causal inference from big data. *Proc. Natl. Acad. Sci. U.S.A* **113**, 7308–7309 (2016).
31. R. M. Shiffrin, S. H. Chandramouli, "Model selection, data distributions, and reproducibility" in *Reproducibility: Principles, Problems, and Practices*, H. Atmanspacher, S. Maasen, Eds. (John Wiley, New York, 2016), pp. 115–140.
32. R. M. Shiffrin, Commentary on "Robust modeling in cognitive science: Misunderstanding the goal of modeling". *Comput. Brain Behav.* **2**, 176–178 (2019).
33. R. M. Shiffrin, Lord's paradox: A commentary on causal inference. *Ann. Biom. Biostat.* **5**, 1034–1036 (2020).
34. R. M. Shiffrin *et al*., Extraordinary claims, extraordinary evidence? A discussion. *Learn. Behav.* **49**, 265–275 (2021).
35. H. Singmann *et al*., Statistics in the service of science: Don't let the tail wag the dog. *Comput. Brain Behav.* **6**, 64–83 (2023).
36. A. Szollosi *et al*., Is preregistration worthwhile? *Trends Cogn. Sci.* **24**, 94–95 (2020).
37. J. Vandekerckhove, J. Rouder, J. Kruschke, Editorial: Bayesian methods for advancing psychological science. *Psychon. Bull. Rev.* **25**, 1–4 (2018).
38. J. Vandekerckhove *et al*., Robust diversity in cognitive science. *Comput. Brain Behav.* **2**, 271–276 (2019).
39. J. Vandekerckhove, E. J. Wagenmakers, C. S. Peirce on the crisis of confidence and the "No More Bets" heuristic. *Winnower* **3**, e146611.14253 (2016).
40. C. P. Davis-Stober *et al*., How can we make sound replication decisions? *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401236121 (2025).
41. A. P. Pedersen *et al*., Discourse on measurement. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401229121 (2025).
42. M. Dubova *et al*., Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401230121 (2025).
43. G. Gigerenzer *et al*., Alternative models of funding curiosity-driven research. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401237121 (2025).
44. B. Aczel *et al*., The present and future of peer review: Ideas, interventions, and evidence. *Proc. Natl. Acad. Sci. U.S.A* **122**, e2401232121 (2025).
45. J. S. Trueblood *et al*., The misalignment of incentives in academic publishing and implications for journal reform. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401231121 (2025).
46. M. Binz *et al*., How should the advancement of large language models affect the practice of science? *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401227121 (2025).
47. S. Musslick *et al*., Automating the practice of science: Opportunities, challenges, and implications. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2401238121 (2025).