Reproducibility of research: Issues and proposed remedies

David B. Allison[a], Richard M. Shiffrin[b,1], and Victoria Stodden[c]

Reproducibility has been one of the major tools science has used to help establish the validity and importance of scientific findings since *Philosophical Transactions of the Royal Society* was established in 1665 (1). Since that time the process of discovery has evolved to make use of new technologies and methods in a changing regulatory and social environment. The Sackler Colloquium "Reproducibility of Research: Issues and Proposed Remedies," which took place on March 8–10, 2017, convened a wide range of research community stakeholders to address our understanding of transparency and reproducibility in the modern research context with two related questions: what does reproducibility mean in different research contexts, and what remedies increase reproducibility and transparency?

We approach the topic of reproducibility with sensitivity to its complexity, spanning a wide range of issues from data collection and reporting to communication of scientific findings by scientists and nonscientists alike. The Colloquium was organized by David Allison, Richard Shiffrin, Victoria Stodden, and Stephen Feinberg. Before the Colloquium our esteemed and respected friend and colleague, Stephen Feinberg, unfortunately died and could not witness the outcome of his vision. A PNAS retrospective by Larry Wasserman describes Stephen's extraordinary career (2).

The 12 articles in this special issue fall into three categories that shaped the Colloquium. The first article in this special issue is "Issues with data and analyses: Errors, underlying themes, and potential solutions," Andrew W. Brown, Kathryn A. Kaiser, and David B. Allison (3). This article sets the stage for this special issue by providing an intellectual framework for understanding the ubiquity of error in the scientific discovery process and presenting methodological, cultural, and system-level approaches to reducing the frequency of commonly observed errors. The next article is "Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data," by Martijn J. Schuemie, George Hripcsak, Patrick B. Ryan, David Madigan, and Marc A. Suchard (4). This work leverages new large medical claims data to identify sources of error when studies purporting to address the same scientific question actually conflict. They suggest the novel technique of confidence interval calibration to reformulate assessments of uncertainty in discovery and to improve the reconciliation of purportedly conflicted findings. The next article, "Training replicable predictors in multiple studies," by Prasad Patil and Giovanni Parmigiani (5) extends this line of thinking by assessing uncertainty introduced to replications in new studies. They propose ways to capture this uncertainty in inferential studies through the strategic use of cross-study validation and ensemble models.

The next section moves to remedies for reproducibility issues. It starts with a focus on leverage points in the scientific discovery pipeline that provide opportunities for improvement. The first article in this section evaluates the effectiveness of journal policies designed to mitigate irreproducibility by providing access to artifacts of computational research, such as data and code. "An empirical analysis of journal policy effectiveness for computational reproducibility" by Victoria Stodden, Jennifer Seiler, and Zhaokun Ma (6) finds that, while an important step in the right direction, journal policies that require postpublication remission of digital scholarly objects by authors, such as the data and code that support the claims, yields fewer than half such artifacts in practice. These artifacts then enabled the reproduction of about a quarter of the published computational claims in the study. They recommend publishing the digital scholarly objects that support claims in the literature at the same time as the publication of the claims themselves. In the next article, "Standards for design and measurement would make clinical research reproducible and usable," Kay Dickersin and Evan Mayo-Wilson (7) bring attention to the need to unify standards in the reporting of clinical trials results

[a]Department of Epidemiology & Biostatistics, Indiana University Bloomington, Bloomington, IN 47405; [b]Department of Psychological and Brain Sciences, Indiana University Bloomington, Bloomington, IN 47405; and [c]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL 61820

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.B.A., R.M.S., and V.S. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: shiffrin@indiana.edu.

and their implementation. They also assert that enforcement of existing standards is an important gap to fill. Also on reporting standards for clinical trials, Guowei Li et al. (8) present empirical evidence for common reporting problems. In "The preregistration revolution," Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor (9) advocate that researchers routinely register hypotheses and research plans before data collection and analysis to improve the accurate of error assessment associated with empirical inference. The notion of a metastudy is introduced in "Metastudies for robust tests of theory" by Beth Baribault et al. (10) describing the idea of leveraging randomization of confounders to quantify robustness of empirical claims.

The final group of articles places the issues in the larger contexts of science practice and science communication. The first examines reporting standards for scientific findings with the article "Misrepresentation and distortion of research in biomedical literature," by Isabelle Boutron and Philippe Ravaud (11). In this work the authors empirically examine accuracy in reporting by scientists quantifying the prevalence of different types of "spin" that may be applied to the presentation of findings. Kathleen Hall Jamieson, in "Crisis or self-correction: Rethinking media narratives about the well-being of science" (12), suggests ways to improve the public reporting of scientific findings to improve accuracy and transparency. Daniele Fanelli in his article "Is science really facing a reproducibility crisis, and do we need it to?" (13) comments on evidence underlying notions of reproducibility in the scientific context and broader discussion that is underway. The final article addresses "Scientific progress despite irreproducibility: A seeming paradox" by Richard M. Shiffrin, Katy Börner, and Stephen M. Stigler (14). This article takes a historical context, shows that the problems of reproducibility are quite old, and suggests the way that science has evolved to allow continued progress despite the problems and challenges of reproducibility that have been identified.

The discussion across the 3 days of the Sackler Colloquium moved from identification and quantification of issues to remedies and steps forward and placed these important issues in the context of scientific practice and communication. As the reader will see in the articles in this special issue, senses of optimism and seriousness simultaneously pervade the approaches to making progress on transparency and sound practice in scientific research.

1 Steven Shapin and Simon Schaffer (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton Univ Press, Princeton).
2 Wasserman L (2017) Stephen Fienberg: Superman of statistics. *Proc Natl Acad Sci USA* 114:3002–3003.
3 Brown AW, Kaiser KA, Allison DB (2018) Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proc Natl Acad Sci USA* 115:2563–2570.
4 Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA (2018) Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci USA* 115:2571–2577.
5 Patil P, Parmigiani G (2018) Training replicable predictors in multiple studies. *Proc Natl Acad Sci USA* 115:2578–2583.
6 Stodden V, Seiler J, Ma Z (2018) An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc Natl Acad Sci USA* 115:2584–2589.
7 Dickersin K, Mayo-Wilson E (2018) Standards for design and measurement would make clinical research reproducible and usable. *Proc Natl Acad Sci USA* 115:2590–2594.
8 Li G, et al. (2018) Enhancing primary reports of randomized controlled trials: Three most common challenges and suggested solutions. *Proc Natl Acad Sci USA* 115:2595–2599.
9 Nosek BA, Ebersole CR, DeHaven AC, Mellor DT (2018) The preregistration revolution. *Proc Natl Acad Sci USA* 115:2600–2606.
10 Baribault B, et al. (2018) Metastudies for robust tests of theory. *Proc Natl Acad Sci USA* 115:2607–2612.
11 Boutron I, Ravaud P (2018) Misrepresentation and distortion of research in biomedical literature. *Proc Natl Acad Sci USA* 115:2613–2619.
12 Hall Jamieson K (2018) Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proc Natl Acad Sci USA* 115:2620–2627.
13 Fanelli D (2018) Is science really facing a reproducibility crisis, and do we need it to? *Proc Natl Acad Sci USA* 115:2628–2631.
14 Shiffrin RM, Börner K, Stigler SM (2018) Scientific progress despite irreproducibility: A seeming paradox. *Proc Natl Acad Sci USA* 115:2632–2639.