Highlights of the US National Academies Report on “Reproducibility and Replicability in Science”

Harvey Fineberg^{1,2} Victoria Stodden^{3,4,5,6} Xiao-Li Meng^{7,8}

¹Technology Assessment Group, Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, Massachusetts, United States of America,

²Office of the President, Gordon and Betty Moore Foundation, Palo Alto, California, United States of America,

³Department of Statistics, College of Liberal Arts and Sciences, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America,

⁴School of Informative Sciences, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America,

⁵National Center for Supercomputing Applications, Office of the Vice Chancellor for Research and Innovation, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America,

⁶Coordinated Science Lab, Department of Computer Science, Grainger College of Engineering, University of Illinois Urbana-Champaign, Urbana, Illinois, United States of America,

⁷Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts, United States of America,

⁸Harvard Data Science Review, Harvard Data Science Initiative, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Oct 30, 2020

DOI: <https://doi.org/10.1162/99608f92.cb310198>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

In 2019, the National Academies of Sciences, Engineering, and Medicine of the United States released a report on reproducibility and replicability in science. This topic is of keen interest to everyone concerned with the reliability of scientific research and the role of computational and statistical analysis in science. In this interview conducted by Xiao-Li Meng, (*HDSR*’s Editor-in-Chief), report committee chair Harvey Fineberg (President of the Gordon and Betty Moore Foundation) and committee member Victoria Stodden (Associate Professor in the School of Information Sciences at the University of Illinois, Urbana-Champaign) recount the aims and deliberations of the committee, its major recommendations, and calls for concerted efforts from data scientists, research scientists, funding agencies, academic institutions, professional journals, and journalists to ensure scientific rigor and public trust in science.

Keywords: misuse of statistics, National Science Foundation, scientific rigor, scientific reporting, reproducibility, replicability

Xiao-Li Meng (XLM): What was the origin of this study on reproducibility and replicability in science? Why does it matter?

Victoria Stodden (VS): This consensus report originated from an Act of Congress. In the [American Innovation and Competitiveness Act of 2017](#), the Congress compelled the National Science Foundation (NSF) to enter into an agreement with the National Academies of Sciences, Engineering, and Medicine to produce a report that assessed “research and data reproducibility and replicability issues in interdisciplinary research” and made “recommendations for improving rigor and transparency in scientific research.” The National Academies convened a committee of experts to study “[Reproducibility and Replicability in Science](#).” I was invited to be a committee member, and Harvey Fineberg graciously agreed to chair the committee.

In recent years, issues of reproducibility and replicability have engaged researchers in fields as diverse as the life sciences, statistics, geophysics, psychology, computational linguistics, and articles on the topic have even emerged in the lay press. A number of reviews noted shortcomings in the clarity, completeness, and specificity of data analysis methods. Biases affecting reports of statistically significant results raised concerns about replicability of research results. At the same time, voices from a variety of research disciplines have been calling for greater transparency in the computational aspects of published research. Consequently, some journal editors have adopted—and others are considering—policies to make data and code used in published articles available to others. Funding agencies have and are adopting requirements to promote transparency around data management plans and other artifacts, such as research code. This consensus exercise offered our committee a

chance to explore the problem and the current state of reform efforts, and to articulate ways the National Science Foundation and others could potentially improve reproducibility and replicability in research.

As just one example of growing attention to issues of research transparency, in January 2019, President Trump signed the Open, Public, Electronic, and Necessary, (OPEN) Government Data Act into law as part of the [Foundations for Evidence-Based Policymaking Act](#). This Act provides that “government data assets made available by federal agencies ... be published as machine-readable data” in an openly licensed way. There are many steps being taken toward greater transparency by a variety of stakeholders across the research landscape.

XLM: Who served on the committee and how does the committee process work?

Harvey Fineberg (HF): [Committee members](#) were drawn from across the science and research enterprise, ranging from scholars experienced in large scale earthquake simulations to experimental physics to behavioral science. The committee included a number of experts whose research focuses on reproducibility itself, such as Victoria. Our deliberations took place over more than a year, with five in-person public meetings at the National Academies and an additional half-dozen video conferences. Our final meeting was off-site and allowed for internal deliberation and writing. We were ably supported throughout by National Academies staff, foremost by the fantastic study director, Jennifer Heimberg.

The committee’s deliberative process involved invited presentations by members of the interested and relevant communities associated with scientific research. Reproducibility and replication are topics that touch nearly every corner of the academic landscape, and we had a wide variety of presentations, from IEEE and NIST to journal editors and scientific society leadership. At our first committee meeting, Dr. Joan Ferrini-Mundy, then Chief Operating Officer of the NSF, and Dr. Suzi Iacono, Head of the Office of Integrative Activities at NSF, discussed the charge to the committee and provided their perspective on our work as representatives of the main sponsor. The committee and staff reviewed the extensive published literature and commissioned a set of white papers to help illuminate different aspects of this multifaceted topic. At various meetings, we heard from invited experts and gathered information from the community. The committee process involved extensive discussion amongst our members to advance a shared understanding of the issues and to formulate our key findings, conclusions, and recommendations.

XLM: I must confess that I was bit anxious when I was waiting for your committee’s definitions of the terms ‘reproducibility’ and ‘replicability,’ since I was unsure if I had been using them correctly. How did the committee come to a consensus on the terms, given that they had been used either interchangeably or in opposite ways?

VS: Early on, the committee recognized that different individuals and scientific fields have used the terms ‘reproducibility’ and ‘replicability’ in varied and sometimes contradictory ways. The committee distinguished efforts to obtain the same results starting with the same data and computational steps, which we called *reproducibility* or, equivalently, computational reproducibility, from efforts to reach comparable results by obtaining new data in an independent study aimed at the same scientific question, which we called *replication*.

Specifically, we adopted the following definitions:

Reproducibility is obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis. This definition is synonymous with ‘computational reproducibility,’ and the terms are used interchangeably in this report. ... Reproducibility is strongly associated with transparency; a study’s data and code have to be available in order for others to reproduce and confirm results.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

In line with previous work at the National Academies, we also defined generalizability to *refer to the extent that results of a study apply in other contexts or populations that differ from the original one.*

XML: Well, I was certainly relieved that I was using them the same way! But for the benefit of *HDSR*’s broad readership, could you elaborate a bit on what reproducibility and replicability entail?

VS: Of course. Reliance on computation and large data sets pervades research today and can be quite complex. Traditional publication and dissemination mechanisms for new scientific results are finding they need to adapt to enable transparency in the reporting of the computational steps taken in producing the findings. Computational steps could include anything from data generation and collection and cleaning, modeling and parameter fitting, implementation details for algorithms on computational systems, and inference of findings from data. Ensuring reproducibility therefore is a process, which requires good effort and can benefit from innovation.

Replication encompasses a broad range of issues in scientific practices, reporting of empirical and other steps taken in the research, and statistical methods, including calculation of p-values and criteria for statistical significance. Some sources of non-replication are helpful, as they can reveal previously unsuspected influences on the results of a study. Other sources—ranging from simple mistakes to sloppy technique, inappropriate use

of statistics, or even fraud—are unhelpful to the progress of science and should, insofar as possible, be eliminated.

A single scientific study may include aspects of any combination of reproducibility, replicability, and generalizability. Reproducibility and replicability differ in the type of results that should be expected: when a researcher transparently reports a study and makes available the underlying digital artifacts, such as data and code, the results should be computationally reproducible at that time, at least on the same system. In contrast, even when a study was rigorously conducted according to best practices, correctly analyzed, and transparently reported, it may fail to replicate. Thus, in a healthy, well-functioning, scientific research ecosystem, some rate of failure to replicate studies is expected.

XLM: These are all great points. Why then is the topic controversial?

HF: Elements that may contribute to controversy begin with confusion over the use of terms, inadequate tools and training to ensure reproducibility, proprietary data or methods that limit transparency, the difficulty in discerning sources of non-replicability when results fail to be replicated, misunderstanding and misuse of statistics, and concerns about the cost of ensuring greater rigor and transparency in scientific research.

Some see the failure to replicate as a source of potential insight and deeper understanding, while others see it as indicating a lack of rigor in science. Not everyone agrees that the time, effort, and cost required to make every analysis of data reproducible is worthwhile. Some critics fail to differentiate between helpful and unhelpful sources of non-replication. Occasional instances of scientific fraud may be sensational and make an outsized impression on the public mind.

While the committee believes the practice of science is fundamentally sound, it also found gaps and needs to ensure greater computational reproducibility and to eliminate unhelpful sources of non-replication. Different scientific disciplines, which differ in the subjects they study, the questions they ask and the methods they employ, raise issues of reproducibility and replicability in differing ways and to varying degrees.

XLM: I can certainly imagine diverse opinions and perspectives here. What, then, are the key findings and recommendations in the report?

HF: The committee’s distinctive emphasis on computational reproducibility recognizes the growing importance of data and computational resources in research. Our discussion stressed the importance of clarity of exposition, completeness in the description of computational steps, and availability of data. We recommended that NSF consider investing in, or endorsing, repositories that can steward digital artifacts that support findings in the scholarly record, such as data and code. We also suggested programmatic initiatives for

future research such as exploring the limits of computational reproducibility and improving our understanding of when to invest in an independent re-implementation of an experiment. We also suggested that NSF consider investing more broadly in the computational tools and cyberinfrastructure that support science to ensure that the systems we rely on as researchers enable scientific research effectively.

Replication is a nuanced topic, and the dual need is to recognize the potential value when rigorous science fails to replicate and to eliminate the unhelpful sources of non-replication. The report highlights the impact of the prior probability of a scientific result on the likelihood that it may be true based on the findings of a scientific study. The degree to which any scientific system is subject to non-replication depends on the complexity and controllability of the system under study. Uncertainty is an inherent part of science, and it is incumbent on scientists to recognize and characterize the uncertainties in their studies and results, including uncertainties due to statistical and computational methods. Techniques such as systematic reviews and meta-analysis may be more revealing about the central tendency and reliability of scientific results than one-off efforts to replicate a previous study. The report outlines criteria, such as the salience of results for policy or individual decision-making, that would increase the value of a successful replication.

Our findings, conclusions, and recommendations make clear that there is no single stakeholder responsible for the transition to greater transparency and rigor in science. In addition to guidance for best practices by scientists and analysts, the committee identified actions that can be taken by the NSF and other funders. The report calls on scientific societies and journals to adopt and disclose their policies on both computational reproducibility and replication. We also call on institutions to increase and improve training of researchers in computation and statistical methods.

XLM: To whom did you direct your recommendations?

HF: Consistent with its charge, the committee directed many recommendations to the National Science Foundation. We also covered steps that could be taken by stakeholders across the scientific research enterprise. Many influences bear on the way research is conducted and reported, including funding agencies, academic promotion and tenure committees, journal and conference manuscript acceptance policies, the existence and accessibility of infrastructure such as repositories and artifact curation and stewardship resources, and guidance from scientific societies. The committee urged concerted efforts across stakeholders to achieve goals in computational transparency and clarity of research reporting.

XLM: What special opportunities do data scientists have to make a difference?

VS: Researchers increasingly rely on new types and scales of data to accelerate and advance their research. These approaches to computationally-enabled discovery pose a challenge and opportunity for data scientists to attain standards of transparency and reproducibility. Data science methods that expose the underlying data and computation are a fundamental step. It is also important to consider how methods of inference used to generate findings, for example predictive models, are expected to generalize to new samples or even different populations. In short, the need and challenge to data scientists and other researchers is to demonstrate why others in the scientific community should believe inferences from data are correct.

XLM: Being a statistician, I cannot resist the temptation to ask about committee’s discussions and views on how much the misuse of statistics has contributed to non-replicable research findings.

HF: A teacher and titan of statistics in the last century, the late Fred Mosteller, once observed, “It is easy to lie with statistics. It is even easier to lie without them.” The report examines a number of substandard practices in the application of statistics, including failure to report or correct for multiple hypothesis testing and generating hypotheses after the results are known. When the frequency of non-replicability greatly exceeds the stated probability of finding the original results by chance, that suggests a problem in the methods or analysis or both in one or more of the studies. While one can identify many instances of misuse of statistics, the committee did not attempt to estimate the fraction of non-replication due to misuse of statistics.

XLM: Here is another question with strong statistical implications. What’s the committee’s take on the issue of exploratory research versus confirmatory studies? Do they require different guidelines and standards for reproducibility and replicability?

VS: The committee discussed the different aims of studies, either exploratory or confirmatory, for example. Being clear on the purpose of the study from the outset and in reporting results is key and conforms with the first recommendation in the report, to “convey clear, specific, and complete information about any computational methods and data products that support their published results...” We realized we could not give a detailed prescription that would properly apply to every study. Confirmatory studies may seek to obtain findings within a margin of error of the original study in order to be considered a success, or they may be subject to the same reproducibility standards as the original since the approaches in the two studies will almost certainly be compared.

XLM: Yet another data-science specific question: as you know, there has been a movement towards ‘automated inference,’ relying on data and algorithms far more than on developing insight into underlying causal or explanatory mechanisms. Did the committee examine this issue, and, if so, consider this tendency a threat or help to scientific replicability?

VS: The committee did not discuss this, but it is an interesting question, and I can offer a few thoughts regarding its impact on reproducibility, which is easier to answer than replicability. Perhaps counterintuitively, there can be reproducibility benefits to automated discovery or automatic inference. Typically, in such systems, all different combinations of variables are systematically included in models and the most highly-predictive model against some test data is anointed the winner. We have actually been doing some flavor of this for a long time in statistics (forward step-wise, backward elimination, the lasso), but in this setup the tracking of computational operations happens automatically, and so the full set of statistical tests is known and can be factored into the interpretation of the models. So, we have a chance to design inference systems with reproducibility in mind. Traditional approaches to statistical inference may end up outperforming automated approaches over time, since oftentimes reconceptualizing the problem or insightfully including factors beyond those in the available data set deliver the scientific breakthroughs.

XLM: Last but certainly not the least, what were the committee’s discussions on the public perception of non-replicability of scientific research as inflated by media’s tendency only reporting the most attention-grabbing findings, which tend to be extreme and less replicable?

HF: Media play a key role in bringing new scientific discoveries to the public. Public media also tend to overstate dramatic findings and may play a role in amplifying scientific controversies. The committee directed a recommendation to journalists, urging them to “report on scientific results with as much context and nuance as the medium allows.” The recommendation highlights circumstances where particular care in reporting is warranted, including when a result is particularly surprising or at odds with existing bodies of research, when a study system is complex with limited control over confounding influences, when the study deals with an emerging area of science with much disagreement within the scientific community, or when funders or researchers may have conflicts of interest.

XLM: This has been a great discussion. Any final thoughts for our readers?

HF: We hope our report stimulates thought, discussion, and action. Aspects of the report are being presented and discussed in a series of professional meetings and conferences. The Congress has asked the National Science Foundation to report on its planned actions in response to the report. A pair of back-to-back workshops at the National Academies in September 2019 considered, respectively, the roles of different stakeholders in promoting reproducibility and replicability and strategies that can be fostered and adopted by the National Institutes of Health, which has long been concerned with enhancing the rigor and reliability of biomedical

science. Ultimately, the choices and practices of individual scientists, and especially of the data science community, will have the most profound effect on improving reproducibility and replicability of science.

XLM: Thanks so much to both of you for providing very rich food for thought for *HDSR* readers, and most importantly to you and your committee for producing a very timely report, which I am sure will also be time-honored.

Disclosure Statement

Harvey Fineberg, Victoria Stodden, and Xiao-Li Meng have no financial or non-financial disclosures to share for this interview.

©2020 Harvey Fineberg, Victoria Stodden, and Xiao-Li Meng. This interview is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the interview.