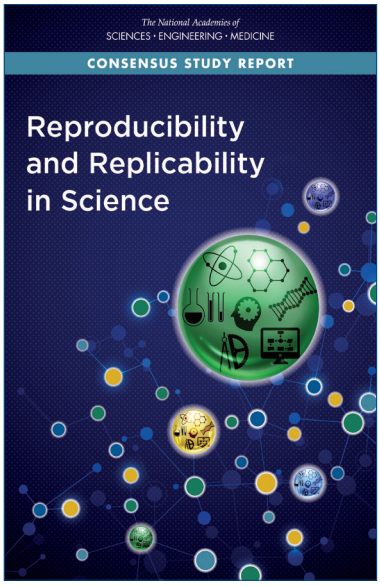
Consensus Study Report
HIGHLIGHTS
May 2019

Reproducibility and Replicability in Science:
Highlights for Social and Behavioral Scientists



Concerns about the reproducibility and replicability of research results have been expressed in both scientific and popular media. As these concerns came to light, Congress requested that the National Academies of Sciences, Engineering, and Medicine assess the extent of issues related to reproducibility and replicability and offer recommendations for improving rigor and transparency in scientific research.

The National Academies' report, *Reproducibility and Replicability in Science* (2019), offers definitions of reproducibility and replicability and examines the factors that may lead to non-reproducibility and non-replicability in research. The report provides recommendations to researchers, academic institutions, journals, professional societies, and funders on steps they can take to improve reproducibility and replicability in science.

This brief offers highlights from the report, focusing on content of interest to researchers in the social and behavioral sciences.

DEFINING REPRODUCIBILITY AND REPLICABILITY

The terms "reproducibility" and "replicability" are often used interchangeably, but the report proposes that each term be used to refer to a separate concept. Reproducibility means computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. Replicability means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. In short, reproducing research involves using the original data and code, while replicating research involves new data collection and similar methods used by previous studies.

These two processes also differ in the expected outcome of a comparison between two results. In general, 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. In contrast, even when a study was rigorously conducted according to best practices, correctly analyzed, and transparently reported, it may fail to be replicated.

REPRODUCIBILITY

The committee's definition of reproducibility focuses on computation because most scientific and engineering research disciplines use computation as a tool, and the abundance of data and widespread use of computation have transformed many disciplines. However, this revolution is not yet uniformly reflected in how scientists use software and how scientific results are published and shared. These shortfalls have implications for reproducibility, because scientists who wish to reproduce research may lack the information or training they need to do so.

To help ensure the reproducibility of computational results, researchers should convey clear, specific, and complete information about any computational methods and data products that support their published results in order to enable other researchers to repeat the analysis.

ASSESSING REPLICABILITY

When researchers investigate the same scientific question using the same methods and similar tools, the results are unlikely to be identical; rather, replicability means obtaining consistent results. Assessing the consistency between two different results or inferences can be approached in a number of ways, and the criteria vary across disciplines.

A number of parametric and nonparametric methods may be suitable for assessing replication across studies. However, an approach that accepts replication only when the results in both studies have attained "statistical significance"—that is, when the *p*-values in both studies have exceeded a selected threshold—is restrictive and unreliable.

The report emphasizes that any determination of replication between two results needs to take account of both proximity (the closeness of one result to the other, such as the closeness of the mean values) and uncertainty (variability in the measures of the results). In addition, to assess replicability, one must first specify exactly what attribute of a previous result is of interest; for example, is only the direction of a possible effect of interest? Is the magnitude of effect of interest? Is surpassing a specified threshold of magnitude of interest? (A full list of principles and characteristics to consider in assessing replication can be found in Chapter 5 of the report.)

SOURCES OF NON-REPLICABILITY

Non-replicability can arise from a number of sources. The report classifies sources of non-replicability into those that are potentially helpful to advancing scientific knowledge, and those that are unhelpful.

Helpful sources of non-replicability. Non-replicability can be caused by inherent but uncharacterized uncertainties in the system being studied, intrinsic variation or complexity in nature, the scope of current scientific knowledge, and the limits of current technologies. When non-replication of results due to these sources is investigated and resolved, it can lead to new insights, better characterization of uncertainties, and increased knowledge about the systems being studied and the methods used to study them.

The susceptibility of any line of scientific inquiry to sources of non-replicability depends on many factors, including those inherent to the system being studied, such as:

- the complexity of the system under study;
- understanding of the number and relations among variables within the system under study;
- the ability to control the variables;
- levels of noise within the system (or signal to noise ratios);
- the mismatch of scale of the phenomena and the scale at which it can be measured;
- stability across time and space of the underlying principles;
- fidelity of the available measures to the underlying system under study (e.g., direct or indirect measurements); and
- prior probability (pre-experimental plausibility) of the scientific hypothesis.

Studies will be more replicable when they are able to better estimate and analyze the uncertainties associated with the variables in the system and control the methods that will be used to conduct the experiment. On the

other end of the spectrum, studies that are more prone to non-replication often involve indirect measurement of very complex systems—for example, human behavior—and require statistical analysis to draw conclusions.

When the sources of non-replicability are knowable, or arise from experimental design choices, researchers need to identify and assess these sources of uncertainty. They should also provide an accurate and appropriate characterization of relevant uncertainties when they report or publish their research.

Unhelpful sources of non-replicability. Non-replicability may be due to shortcomings in the design, conduct, and communication of a study. Whether arising from lack of knowledge, perverse incentives, sloppiness, bias, or fraud, these unhelpful sources of non-replicability reduce the efficiency of scientific progress.

One unhelpful source of non-replicability is inappropriate statistical inference. Misuse of statistical testing often involves post-hoc analysis of data already collected, making it seem as though statistically significant results provide evidence against the null hypothesis, when in fact they have a high probability of being false positives. Other inappropriate statistical practices include *p*-hacking—the practice of collecting, selecting, or analyzing data until a result of statistical significance is found—and cherry picking, in which researchers may unconsciously or deliberately sort through their data and results and selectively report those that satisfy criteria such as meeting a threshold of statistical significance or supporting a positive result, rather than reporting all of the results from their research.

Unhelpful sources of non-replicability can be minimized through initiatives and practices aimed at improving research design and methodology, including training in the proper use of statistical analysis and inference, mentoring, repeating experiments before publication, rigorous peer review, utilizing tools for checking analysis and results, and better transparency in reporting.

ISSUES SPECIFIC TO THE SOCIAL AND BEHAVIORAL SCIENCES

The report explores some topics related to reproducibility and replicability that are currently the subject of discussion by researchers in the behavioral and social sciences.

Replication in psychology. There is a wide range of opinion in psychology about the replicability of research results and about the extent and interpretation of cases in which results cannot be replicated. Some researchers believe that the field faces problems such as frequent use of lax methods that threaten validity, while other researchers disagree with this characterization. Still others have noted that psychology has long been concerned with improving its methodology, and the current discussion of replicability is part of the normal progression of science.

One reason to believe in the fundamental soundness of psychology as a science is that a great deal of useful and reliable knowledge—replicable discoveries about human thought, emotion, and behavior—is being produced. Increasingly, researchers and governments are using such knowledge to meet social needs and solve problems, such as improving educational outcomes and reducing government waste from ineffective programs.

Researchers in psychology have been at the forefront of attempts to study and estimate levels of replicability in their discipline, such as the Open Science Framework. However, there is no definitive estimate of replicability in psychology, in part because no one knows the expected level of non-replicability in a healthy science. Inconsistent results might stem from both unhelpful sources of non-replicability and helpful ones due to innovative research.

Whatever the extent of the problem, unhelpful sources of non-replicability should be sought out and, insofar as possible, eliminated. New practices, such as checks on the efficacy of experimental manipulations, are now accepted in the field. Funding proposals now include power analyses as a matter of course. Longitudinal studies no longer just note attrition but instead routinely estimate its effects. At the same time, not all researchers have adopted best practices, sometimes failing to keep pace with current knowledge. Preregistration as a solution to address unhelpful sources of non-replicability has both advantages and disadvantages. However, the effectiveness of preregistration in improving research reliability practices is unknown.

Social science research using big data. With close to 7 in 10 Americans now using social media as a regular news source, social scientists in communication research, psychology, sociology, and political science routinely analyze information disseminated on these platforms, such as Twitter and Facebook, how that information flows through social networks, and how it influences attitudes and behaviors. These analyses may rely on publicly available data that can be collected by any researcher without input from industry partners, or industry staff may provide access to proprietary data for analysis that may not be available to others.

Both models raise challenges for reproducibility and replicability. In terms of reproducibility, when data are proprietary and undisclosed, the computation by definition is not reproducible by others—which might put the research at odds with publication requirements of journals and other academic outlets. Issues with replicability are raised by the fact that social media platforms regularly modify their application programming interfaces, which influences the ability of researchers to access, document, and archive data consistently. In addition, data are likely confounded by ongoing testing and tweaks to underlying algorithms. In summary, the considerations for social science using big data of this type illustrate a spectrum of challenges and approaches to gaining confidence in scientific studies.

THE VALUE OF META-ANALYSIS

Replicability and reproducibility, useful as they are in confirming scientific knowledge, are not the only ways to gain confidence in scientific results. Multiple channels of evidence from a variety of studies provide a robust means for gaining confidence in scientific knowledge over time. Research synthesis and meta-analysis, for example, are valuable methods for assessing bodies of research.

Using summary statistics or individual-level data, meta-analysis provides estimates of overall central tendencies, effect sizes or association magnitudes, along with estimates of the variance or uncertainty in those estimates. Meta-analyses also test for variation in effect sizes and, as a result, can suggest potential causes of nonreplicability in existing research. Meta-analysis can quantify the extent to which results appear to vary from study to study solely due to random sampling variation, or to variation in a systematic way by subgroups, as well as by characteristics of the individual studies; such analyses must take into account the possibility that published results may be biased by selective reporting and, to the extent possible, estimate its effects.

COMMITTEE ON REPRODUCIBILITY AND REPLICABILITY IN SCIENCE

HARVEY V. FINEBERG (NAM), (*Chair*), Gordon and Betty Moore Foundation; **DAVID B. ALLISON** (NAM), School of Public Health-Bloomington, Indiana University; **LORENA A. BARBA,** School of Engineering and Applied Science, George Washington University; **DIANNE CHONG** (NAE), Boeing Research and Technology (retired); **JULIANA FREIRE,** Tandon School of Engineering, New York University; **GERALD GABRIELSE** (NAS), Department of Physics, Northwestern University; **CONSTANTINE GATSONIS,** Center for Statistical Sciences, Brown University; **EDWARD HALL,** Department of Philosophy, Harvard University; **THOMAS H. JORDAN** (NAS), Department of Earth Sciences, University of Southern California; **DIETRAM A. SCHEUFELE,** Madison and Morgridge Institute for Research, University of Wisconsin–Madison; **VICTORIA STODDEN,** Institute for Data Sciences and Engineering, University of Illinois at Urbana–Champaign; **TIMOTHY D. WILSON,** Department of Psychology, University of Virginia; **WENDY WOOD,** Department of Psychology, University of Southern California and INSEAD-Sorbonne University; **JENNIFER HEIMBERG,** *Study Director;* **THOMAS ARRISON**, *Program Director;* **MICHAEL COHEN**, *Senior Program Officer;* **MICHELLE SCHWALBE**, *Director;* **TINA WINTERS**, *Associate Program Officer;* **THELMA COX,** *Program Coordinator*.

For More Information . . . This Consensus Study Report Highlights for Social and Behavioral Scientists was prepared by the Board on Behavioral, Cognitive, and Sensory Sciences based on the Consensus Study Report *Reproducibility and Replicability in Science* (2019). The study was sponsored by the Alfred P. Sloan Foundation and the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. Copies of the Consensus Study Report are available from the National Academies Press, (800) 624-6242; http://www.nas.edu/ReproducibilityinScience.