INVITED TESTIMONY BEFORE THE U.S. HOUSE OF REPRESENTATIVES' COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY HEARING "STRENGTHENING TRANSPARENCY OR SILENCING SCIENCE? THE FUTURE OF SCIENCE IN EPA RULEMAKING"

Statement of

David B. Allison, Ph.D.

Dean, Distinguished Professor, & Provost Professor

Indiana University School of Public Health-Bloomington

and

Member, Committee on Reproducibility and Replicability in Science

The National Academies of Sciences, Engineering, and Medicine

before the

U.S. House Of Representatives' Committee on Science, Space, and Technology

November 13, 2019

My name is David B. Allison. I currently serve as the dean of the School of Public Health at Indiana University – Bloomington, although on this occasion I am speaking as a member of the Committee on Reproducibility and Replicability in Science and on behalf of the National Academies of Sciences, Engineering, and Medicine (the National Academies) and not Indiana University. I have been asked by The U.S. House of Representatives' Committee on Science, Space, and Technology to testify at their hearing titled, *"Strengthening Transparency or Silencing Science? The Future of Science in EPA Rulemaking"* on November 13, 2019. I understand from the invitation that "The purpose of the hearing is to assess the EPA's proposed rule *Strengthening Transparency in Regulatory Science.*"

In my testimony, I have been asked to address the following topics:

- The definition of reproducibility, as determined by the Committee on Reproducibility and Replicability in Science of the National Academies;
- The potential consequences of EPA's goal to establish a reproducibility standard within its regulatory process by requiring that the underlying data of scientific studies be made available;
- Whether it is appropriate to determine the rigor or regulatory applicability of a study based solely on its reproducibility; and
- Whether a reproducibility requirement could increase the risk that sound science could be excluded from EPA environmental and public health regulations.

These topics will be addressed in addition to several other points. In this testimony, I will provide:

1. A brief background on the Committee on Reproducibility and Replicability in Science and my involvement in it.
2. Some overview remarks about science found in the "Reproducibility and Replicability in Science" report as well as my own personal perspectives that serve as context for this discussion.
3. Responses to the topics posed by the House Science committee.
4. A copy of the Executive Summary of the "Reproducibility and Replicability in Science" report.

My testimony ends with a summary of its main points which are my own personal perspectives.

1. A brief background on the Committee on Reproducibility and Replicability in Science and my involvement in it.

The American Innovation and Competitiveness Act of 2017 directed the National Science Foundation to engage the National Academies in a study to assess reproducibility and replicability in scientific and engineering research and to provide findings and recommendations for improving rigor and transparency in scientific research. The National Academies appointed a committee of experts to carry out this evaluation, representing a wide range of expertise and backgrounds: methodology and statistics, history and philosophy of science, science communication, behavioral and social sciences (including experts in the social and behavioral factors that influence the reproducibility and replicability of research results), earth and life sciences, physical sciences, computational science, engineering, academic leadership, journal editors, and industry expertise in quality control. In addition, individuals with expertise pertaining to reproducibility and replicability of research results across a variety of fields were selected. Dr. Harvey Fineberg, President of the Gordon and Betty Moore Foundation and a past president of the Institute of Medicine—now the National Academy of Medicine—served as the

chair of the Committee. The Committee's report is available for download without charge at: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science.

I was asked to serve as a committee member based on my work as a scientist and my long-term interest in issues related to reproducibility, replicability, and rigor in science such as my involvement in organizing and participation in the 2017 National Academy of Sciences Colloquium which was focused on these issues. My research interests include obesity and nutrition, quantitative genetics, clinical trials, statistical and research methodology, and research rigor and integrity. I have authored more than 600 scientific publications and edited five books. A member of the National Academy of Medicine of the National Academies, I am also an elected fellow of the American Association for the Advancement of Science, the American Statistical Association, the American Psychological Association, the New York Academy of Medicine, the Gerontological Society of America, the Academy of Behavioral Medicine Research, and other academic societies. I have devoted my career to the rigorous pursuit of knowledge through science. It is an honor to represent the Committee on Reproducibility and Replicability in Science and to discuss the content of its report and my perspectives on these topics with the U.S. House Committee on Science.

2. **Science as a shared communal process for objectively determining the truth of propositions about the world.**

Science is a method by which society tries to discover and share knowledge about the state of the world. It is fundamentally a communal process in which communicating the research process and findings, helping others to understand the knowledge obtained, and subjecting conclusions and the bases for them to public questioning and scrutiny are all essential components. What makes science special both in its claims to have access to objective knowledge about the world as well as in its communal process involves the methods by which scientific knowledge is generated. In particular, "in science, three things matter: the data, the methods used to collect the data (which give them their probative value), and the logic connecting the data and methods to conclusions."[1] These are the substrates of science.

Because of the critical role of methods in this process, it is an essential tenet of science that the methods used to collect or produce data and to analyze them be as thoroughly and transparently described as possible so that others may understand what was done and thereby judge the probative value of the data. Thus, transparency is critical to one of the three fundamental elements of science as I have described. As the Committee states in its report (p. 32), "*When research is communicated with clear, specific, and complete accounting of the materials and methods used, the results found, and the uncertainty associated with the results, other scientists can know how to interpret the results. The communal enterprise of science allows scientists to build on others' work, develop the necessary skills to conduct high quality studies, and check results and confirm, dispute, or refine them.*" In short, observability of

[1] https://www.pnas.org/content/115/11/2563

methods and observability of data, which might both be considered under the rubric of "transparency", support the objectivity and communal validation process of science.

That is, as scientists or individuals consuming and judging the validity, value, and utility of science, we need to know more than one's answer, we need to know how one got that answer. The phrase so many of us heard from our middle school math teachers "show your work" is an apt description. Only by seeing the process of the work done to produce an answer in science can we judge that answer. This observability requires transparency. This observability and transparency in turn makes reproducibility possible.

Reproducibility is a word that is used in multiple different ways in the scientific and general communities. Most recently, as I will state in Section 4, the term reproducibility was defined in the Reproducibility and Replicability in Science report as follows (p. 46) "*reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with 'computational reproducibility'*." Notably, reproducibility is neither a necessary nor sufficient condition for a particular scientific project to be judged as valid for supporting any conclusions drawn from it. It is a valuable aspect of science, but only one aspect of science that is valuable, and it is not clear that reproducibility should merit a privileged position as the sole arbiter of whether a particular study or data set should be admitted into a discussion of evidence.

It is worth noting that the Committee on Reproducibility and Replicability in Science did not consider the EPA proposed rule in its tasking and, since the proposed rule was released during Committee deliberations, the Committee's report was not publicly available while EPA's proposed rule was underdevelopment.

The proposed EPA rule does not necessarily state that reproducibility *per se* or even that transparency will be the sole arbiter of the admission of evidence into the policy making process, but it might be construed as implying this. Part of the challenge with the proposed rule is the substantial number of terms including *reproducibility*, *transparency*, *independent validation*, and others which are not all explicitly defined. This leads to ambiguity in how the rule may be interpreted and utilized. Rulemaking is arguably not served by ambiguity nor is science itself. Though some ambiguity is inherent in all language, we should strive to be precise in terms. Therefore, if some variant of the proposed EPA rule were to go forward, the public interest would likely be served by defining all terms as precisely as possible, by including factors other than reproducibility (at least as the Committee's report has defined it) as key factors in determining how to evaluate evidence, as well as potentially making other modifications.

From my perspective, it is important to consider what the ultimate goals of science and policy making are in considering what those other modifications might be. The ultimate goal of science is to uncover and communicate truths about the state of the world. The ultimate goal of policy making is to serve the interests of the public. Science is a valuable input to policy making decisions but can never be fully dispositive of policy-making decisions which also must take into account moral, social, economic, political, and other factors. But the evaluation of the science *per se* should be based only on the science and not on these other factors.

Science can inform us about the plausible truth of propositions. These propositions can relate to things such as how much of a substance is in the environment, whether the amount of a substance in the environment has increased or decreased, what may have caused exposure to various substances, what the effects of exposure to various substances at various times in various doses are in humans, etc. Reproducibility is of interest because it potentially helps us to evaluate the extent to which a study supports the truth of some proposition and, in the long run, buttresses the entire enterprise of science and thereby ensures that we are better able to pursue truth through science. As the Committee report states (p. 33): "*Science is engaged in a continuous process of refinement to uncover ever closer approximations to the truth*." In the report, Conclusion 2-1 states (p.33):

> "*CONCLUSION 2-1: The scientific enterprise depends on the ability of the scientific community to scrutinize scientific claims and to gain confidence over time in results and inferences that have stood up to repeated testing. Reporting of uncertainties in scientific results is a central tenet of the scientific process. It is incumbent on scientists to convey the appropriate degree of uncertainty in reporting their claims.*"

The degree of certainty about the truth of any proposition in science comes from many sources including but not limited to reproducibility. The overall rigor of the science such as the quality of the measurement instruments used, the extent to which the findings have been replicated (as opposed to simply reproduced), the degree of transparency and reporting of the science, the extent to which it has been thoroughly peer reviewed, the extent to which results fit with a larger body of data available to the scientific community, are all factors that can come into play in judging the extent to which we have a scientific basis for believing that any particular proposition is true. Collectively, all of these things might be called "rigor." My colleagues and I on the Committee wrote (p.52):

> "*Rigor is defined as 'the strict application of the scientific method to ensure robust and unbiased experimental design' (National Institutes of Health, 2018e). Rigor does not guarantee that a study will be replicated, but conducting a study with rigor—with a well-thought-out plan and strict adherence to methodological best practices—makes it more likely. One of the assumptions of the scientific process is that rigorously conducted studies 'and accurate reporting of the results will enable the soundest decisions' and that a series of rigorous studies aimed at the same research question 'will offer successively ever-better approximations to the truth' (Wood et al., 2019, p. 311).*"[2]

From my personal perspective, it may not be apt for a governmental rule to define the admissibility of evidence into a discussion on consideration of a policy that can and should be informed by science solely on the basis of reproducibility. I have stated that one reason for this

[2] See National Institutes of Health. (2018e). Rigor and Reproducibility in NIH Applications: Resource Chart. Available: https://grants.nih.gov/grants/RigorandReproducibilityChart508.pdf and Wood, A.C., Wren, J.D., and Allison, D.B. (2019). The Need for Greater Rigor in Childhood Nutrition and Obesity Research. JAMA Pediatrics, 173(4), 311-312. doi:10.1001/jamapediatrics.2019.0015.

is that reproducibility is neither a necessary nor a sufficient condition for a scientific study or data set to be valid or useful. This is so for many reasons.

First, a study can be reproducible and transparent and yet completely invalid. If a second analyst repeats the entire process of the first analyst applied to the same data, including the first analyst's mistakes or to a data set that is fundamentally flawed and inappropriate, an answer may be reproduced, the process may be transparent, and yet the answer may be worthless and invalid.

Additionally, a scientific project may not be reproducible because the available information is insufficient to allow someone to reproduce it. The original raw data may not be available or for many reasons may not be able to be made public. The original investigators may not have sufficiently documented their steps to allow a full evaluation of exactly what was done permitting an exact reproduction. These are certainly limitations and should be noted. And yet, limitations are not necessarily the same as invalidating factors that should exclude information from further inquiry. A general tenet of scientific evaluation is that one should consider all of the available evidence. One may weigh the individual elements of evidence differentially, but it is uncommon to exclude particular evidence from consideration because it contains some limitations. Virtually all empirical evidence is imperfect and has some limitations. It is vital that in the scientific process those limitations are noted and some of those limitations may preclude firm conclusion-making. Yet the evidence should still be weighed and considered.

In considering the rationale for this approach, the fundamental distinction between the idea of conclusion-making and decision-making is called for. Scientific conclusion-making may depend on certain key types of data. Scientific conclusion-making may depend upon scientific evidence which supports a sufficient degree of certainty that rules out alternative explanations that would compete with a proposition being accepted as true to some reasonable degree of certainty. For example, in biomedical research, and many other domains, scientists will often not be prepared to state unequivocally that it has been demonstrated by scientific methods that '*x causes y*' unless there has been a randomized controlled experiment in which experimental units (e.g., people in medical trial) have been randomly assigned to different levels of *x* (e.g., to take a drug vs. a placebo or to eat diet A vs. diet B). Yet, in medicine, nutrition, public health, and other applied domains we are often called upon to make recommendations to individual patients, citizens, or society at large and often must do so in the absence of data that would be sufficient to allow us to draw a firm scientific conclusion that *x causes y*. We may have to make our recommendation simply by saying that *it seems likely that x causes y* even though it has not been *demonstrated that x causes y*. When we make a recommendation that somebody should *act as though x causes y* even though we have not demonstrated scientifically that x causes y, we are involved in decision-making not conclusion-making. The scientific conclusion can remain unclear while we still proceed with a recommendation. In all cases that recommendation should be made with honesty, letting those to whom we communicate it know that we have not yet demonstrated that x causes y only that it seems a reasonable and plausible proposition given the available information.

This distinction was put eloquently by Sir Austin Braford Hill in 1965 who considered issues such as whether smoking caused lung cancer. He recognized that there were not randomized

controlled trials demonstrating unequivocally that smoking causes lung cancer but that the evidence for an association between smoking and lung cancer was extremely strong and, combined with much other information in the scientific domain, has led virtually all scientists to accept the proposition that smoking causes lung cancer as true beyond any reasonable doubt.[3] In discussing the thought process involved in this, Sir Austin Bradford Hill stated "*in passing from association to causation I believe in 'real life', we shall have to consider what flows from that decision. On scientific grounds, we should do no such thing. The evidence is there to be judged on its merits and the judgment (in that sense) should be utterly independent of what hangs upon it – or who hangs because of it.*"[4]

Similarly, in a recent New York Times' article considering the controversy around the health effects of red meat,[5] I was quoted as describing the distinction between evidence for conclusion-making versus evidence for decision-making, stating "*The standards of evidence for the former are scientific matters and should not depend on extra scientific considerations. The standards of evidence for the latter are matters of personal judgment or in some cases legislation. People should be aware of the uncertainty and make their decisions based on that awareness.*"

This recognition of the difference between decision-making for applied purposes, the fundamental aspect of policy making, and conclusion-making for scientific purposes underlies the very credible approaches taken by multiple other government organizations with respect to their consideration of evidence around key questions. For example, in their discussion of what constitutes adequate evidence for making their decisions about such things as drug approvals, the U.S. Food and Drug Administration has stated (p. 5):[6]

> "*The need for independent substantiation has often been referred to as the need for replication of the finding. Replication may not be the best term, however, as it may imply that precise repetition of the same experiment in other patients by other investigators is the only means to substantiate a conclusion. Precise replication of a trial is only one of a number of possible means of obtaining independent substantiation of a clinical finding and, at times, can be less than optimal as it could leave the conclusions vulnerable to any systematic biases inherent to the particular study design. Results that are obtained from studies that are of different design and independent in execution, perhaps evaluating different populations, endpoints, or dosage forms, may provide support for a conclusion of effectiveness that is as convincing as, or more convincing than, a repetition of the same study.*"
>
> . . . (p.17) "*However, situations often arise in which studies that evaluate the efficacy of a drug product lack the full documentation described above (for example, full patient records may not be available) or in which the study was conducted with less monitoring than is ordinarily seen in commercially*

[3] https://www.americanscientist.org/article/reasonable-versus-unreasonable-doubt
[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1898525/
[5] https://www.nytimes.com/2019/09/30/health/red-meat-questions-answers.html
[6] https://www.fda.gov/media/71655/download

sponsored trials. Such situations are more common for supplemental indications because postapproval studies are more likely to be conducted by parties other than the drug sponsor and those parties may employ less extensive monitoring and data-gathering procedures than a sponsor. Under certain circumstances, it is possible for sponsors to rely on such studies to support effectiveness claims, despite less than usual documentation or monitoring."

Similarly, the "Reference Manual on Scientific Evidence" produced by the National Research Council of the National Academies and the Federal Judicial Center states (p.330): [7]

> *"A party that offers data to be used in statistical work, including multiple regression analysis, should be encouraged to provide the following to the other parties: (a) a hard copy of the data when available and manageable in size, along with the underlying sources; (b) computer disks or tapes on which the data are recorded; (c) complete documentation of the disks or tapes; (d) computer programs that were used to generate the data (in hard copy if necessary, but preferably on a computer disk or tape, or both); and (e) documentation of such computer programs. The documentation should be sufficiently complete and clear so that the opposing expert can reproduce all of the statistical work."*

Yet, also states (Preface, p. xiv):

> *"In the final analysis, a judge does not have the option of suspending judgment until more information is available, but must decide after considering the best available science."*

In the academic community, we have a system called GRADE.

> "*GRADE (Grading of Recommendations, Assessment, Development and Evaluations) is a transparent framework for developing and presenting summaries of evidence and provides a systematic approach for making clinical practice recommendations.[1-3] It is the most widely adopted tool for grading the quality of evidence and for making recommendations with over 100 organisations worldwide officially endorsing GRADE.*"[8]

In using systems like GRADE, while limitations of individual studies are noted, "*…the credibility and trustworthiness of* ***the totality of evidence*** [emphasis added] *across studies in relation to a specific research question*"[9] is key. This reliance on the totality of evidence via GRADE is also a hallmark of the process for generating dietary recommendations used by Federal Agencies.[10] Thus, GRADE is used to help evaluate evidence that can potentially support decisions about public health recommendations. Importantly, GRADE defines principles for standards of evidence and helps evaluate individual pieces of evidence so that they may be properly weighed in an analysis. In contrast, GRADE does not specifically state that certain types of evidence will

[7] https://www.fjc.gov/sites/default/files/2015/SciMan3D01.pdf
[8] https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001464/
[10] https://www.ncbi.nlm.nih.gov/books/NBK465019/

simply be excluded from discussion, but rather outlines which types of evidence should be given greater value and lead to more confident conclusions versus less confident conclusions.

All of this leads one to ask whether the public interest would be well-served by modifying the current proposed EPA rule to increase clarity around definitions and procedures for its implementation. Or would the public interest be better-served by a more thorough and expansive statement of principles as to what constitutes good scientific evidence, about ideals of scientific evidence which include, but are not limited to, reproducibility and transparency, and suggestions for how to weigh and evaluate evidence both for drawing scientific conclusions and for making prudent decisions. A statement of such broad principles may serve the interests of the public and of science by promoting openness in science, good quality science, rational policy making, and transparency in both science and government, more so than does a rule which serves to exclude certain information from consideration.

3. Executive Summary of the "Reproducibility and Replicability in Science" of the National Academies.

The executive summary of the "Reproducibility and Replicability in Science" report of the National Academies appears as Appendix A to this document.

4. Responses to Specific Questions.

a) The definition of reproducibility, as determined by the Committee on Reproducibility and Replicability in Science of the National Academies.

The term reproducibility is defined in Conclusion 3-1 in the Committee's report, "Reproducibility and Replicability in Science" (p. 46):

> *Reproducibility* is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "computational reproducibility". . .

The Committee's definition of replicability is also important. The same section of the report defines:

> *"Replicability* to mean obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data."

b) The potential consequences of EPA's goal to establish a reproducibility standard within its regulatory process by requiring that the underlying data of scientific studies be made available.

In my opinion, the answer to this depends upon exactly how the rule is implemented and modified. If reproducibility were to become the sole arbiter of whether information, a study, or a data set were included in policy making considerations, I believe the effects would be deleterious for the reasons I have stated above. Some high-quality information that, for any number of reasons, cannot be made fully reproducible and transparent would be excluded. Moreover, the rule might lead to the mistaken conclusion that information that was judged to

be admissible because it met a transparency or reproducibility standard was valid information, and as I have indicated above there can be much that is reproducible and transparent but is nonetheless invalid or otherwise flawed.

The likely quality of the outcomes as a result of the proposed rule would also depend upon the extent to which the request that underlying data be transparent and the studies be reproducible be implemented flexibly and in an unbiased manner or inflexibly or in a biased manner. Were a rule to be implemented that provided strong encouragement and incentives for making science reproducible and transparent, that would be good. In contrast, if such a rule became *dicto simpliciter* and a sole arbiter of whether information could be included, that would be bad. Certainly, the current EPA rule contains many situations in which exceptions can be made. That is wise. Yet what is unclear to me is whether the rule is necessary at all and, if it is valuable, how these exceptions will be adjudicated and whether the process of making them will lead to excessive use of time, excessive exclusion of studies, and potential bias in terms of which studies and datasets ultimately are allowed to be included.

c) Whether it is appropriate to determine the rigor or regulatory applicability of a study based solely on its reproducibility.

No, from my perspective, it would not be appropriate to determine the rigor or the regulatory applicability of a study based solely on its reproducibility as reproducibility is defined in the National Academies' report for the reasons I have stated above. In short, reproducibility is neither a necessary nor a sufficient condition to determine the validity of a study for in turn determining the truth of a proposition.

d) Whether a reproducibility requirement could increase the risk that sound science could be excluded from EPA environmental and public health regulations.

It is not clear to me that the currently proposed rule definitively proposes a reproducibility requirement as the sole arbiter or a *sine qua non* for which studies and datasets can enter into policy making because the proposed rule only addresses certain aspects of policy making and it allows for exceptions. Yet, for the reasons I have described above, I do think there is some danger that if reproducibility is poorly defined and more importantly if it becomes the sole and essential criterion for inclusions of data, then yes, such a requirement could risk that sound science would be excluded from EPA environmental and public health regulations.

5. Summation.

In summation, the National Academies Committee and I as both a member and an individual scientist are a strong proponents of reproducibility and replicability, of transparency in science, and more importantly and more broadly of the utmost rigor in the execution of science and in the unvarnished truthful communication of scientific information among scientists and to society at large. I personally believe that any effort that serves to promote the goals of reproducibility, transparency, scientific rigor, and truthful communication in and about science should be supported. To the extent that EPA can enact guidance, statements, policies, and procedures that promote these practices, that is all to the good. Yet there must be flexibility

such that we may consider and speak openly about data even when those data have limitations including, but not limited to, incomplete transparency or reproducibility of some datasets and studies. Just as other scientific communities and other government regulatory bodies relying on scientific information must do, in this realm, I advocate that we consider *all* the information, while providing the most weight to the best information.

Appendix A

The Executive Summary from "Reproducibility and Replicability in Science" is copied below. The full report may be downloaded without charge at: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

EXECUTIVE SUMMARY

When scientists cannot confirm the results from a published study, to some it is an indication of a problem, and to others, it is a natural part of the scientific process that can lead to new discoveries. As directed by Congress, the National Science Foundation (NSF) tasked this committee to define what it means to reproduce or replicate a study, to explore issues related to reproducibility and replicability across science and engineering, and to assess any impact of these issues on the public's trust in science.

Various scientific disciplines define and use the terms "reproducibility" and "replicability" in different and sometimes contradictory ways. After considering the state of current usage, the committee adopted definitions that are intended to apply across all fields of science and help untangle the complex issues associated with reproducibility and replicability. Thinking about these topics across fields of science is uneven and evolving rapidly, and the report's proposed steps for improvement are intended to serve as a roadmap for the continuing journey toward scientific progress.

We define *reproducibility* to mean computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, and code, and conditions of analysis; and *replicability* to mean obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data. In short, reproducibility involves the original data and code; replicability involves new data collection and similar methods used by previous studies. A third concept, *generalizability,* refers to the extent that results of a study apply in other contexts or populations that differ from the original one.[11] A single scientific study may entail one or more of these concepts.

Our definition of reproducibility focuses on computation because of its large and increasing role in scientific research. Science is now conducted using computers and shared databases in ways that were unthinkable even at the turn of the 21st century. Fields of science focused solely on computation have emerged or expanded. However, the training of scientists in best computational research practices has not kept pace, which likely contributes to a surprisingly low rate of computational reproducibility across studies. 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. Proprietary and non-public data and code add challenges to meeting transparency goals. In addition, many decisions related to data selection or parameter setting for code are made throughout a study and can affect the results. Although newly developed tools can be used to capture these decisions and include them as part of the digital record, these tools are not used by the majority of scientists. Archives to store digital artifacts linked to published results are inconsistently maintained across journals, academic and federal institutions, and

[11] The definition of generalizability used by the NSF (Bollen, et al., 2015).

disciplines, making it difficult for scientists to identify archives that can curate, store, and make available their digital artifacts for other researchers.

To help remedy these problems, the NSF should, in harmony with other funders, endorse or create code and data repositories for long-term preservation of digital artifacts. In line with its expressed goal of "harnessing the data revolution," NSF should consider funding tools, training, and activities to promote computational reproducibility. Journal editors should consider ways to ensure reproducibility for publications that make claims based on computations, to the extent ethically and legally possible.

While one expects in many cases near bitwise agreement in reproducibility, the replicability of study results is more nuanced. Non-replicability occurs for a number of reasons that do not necessarily reflect that something is wrong. Some occurrences of non-replicability may be helpful to science—discovering previously unknown effects or sources of variability—while others, ranging from simple mistakes to methodological errors to bias and fraud, are not helpful. It is easy to say that potentially helpful sources should be capitalized on, while unhelpful sources must be minimized. But when a result is not replicated, further investigation is required to determine whether the sources of that non-replicability are of the helpful or unhelpful variety or some of both. This requires time and resources and is often not a trivial undertaking.

A variety of standards are used in assessing replicability, and the choice of standards can affect the assessment outcome. We identified a set of assessment criteria that apply across sciences highlighting the need to adequately report uncertainties in results. Importantly, the assessment of replicability may not result in a binary pass/fail answer; rather, the answer may best be expressed as the degree to which one result replicates another.

One type of scientific research tool, statistical inference, has had an outsized role in replicability discussions due to the frequent misuse of statistics such as the *p*-value and threshold for determining "statistical significance." Inappropriate reliance on statistical significance can lead to biases in research reporting and publication; although publication and research bias are not restricted to studies involving statistical inference. A variety of ongoing efforts is aimed at minimizing these biases and other unhelpful sources of non-replicability.

Researchers should take care to estimate and explain the uncertainty inherent in their results, to make proper use of statistical methods, and to describe their methods and data in a clear, accurate, and complete way. Academic institutions, journals, scientific and professional associations, conference organizers and funders can take a range of steps to improve replicability of research. We propose a set of criteria to help determine when testing replicability may be warranted. It is important for everyone involved in science to endeavor to maintain public trust in science based on a proper understanding of the contributions and limitations of scientific results.

A predominant focus on the replicability of individual studies is an inefficient way to assure the reliability of scientific knowledge. Rather, reviews of cumulative evidence on a subject, to assess both the overall effect size and generalizability, is often a more useful way to gain confidence in the state of scientific knowledge.

==========