Harvard Data Science Review • Issue 2.4, Fall 2020

Self-Correction by Design

Marcia McNutt¹

¹National Academy of Sciences, District of Columbia, United States of America

Published on: Jan 11, 2021

DOI: <https://doi.org/10.1162/99608f92.32432837>

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

ABSTRACT

Science has always been self-correcting. Through traditional approaches such as establishing benchmarks and standards, and reproducing, replicating, or generalizing prior research results, work that is at the same time important and suspect will be subjected to further testing. With science now becoming simultaneously critically important to society and an expensive enterprise, I argue that we need to adopt an enterprise-wide approach to self-correction that is built into the design of science. Principles of such an approach include educating students to perform and document reproducible research, sharing information openly, validating work in new ways that matter, creating tools to make self-correction easy and natural, and fundamentally shifting the culture of science to honor rigor.

Keywords: replication, reproducibility, trust, geosciences

Marcia McNutt - Video Abstract

1. Introduction

All citizens intrinsically are interested in scientific results, as they matter to our health, security, safety, economic prosperity, and quality of life. Scientists are therefore obligated to use all means at our disposal to ensure that science advances from intriguing hypothesis to validated consensus as quickly and efficiently as possible. In the current age of megascience, the need to build quality assurance into the scientific method from its inception is even more critical.

In this perspective, I recall from my background in geophysics the various quality control methods that have served science well in past decades, and how they now must be institutionalized and formalized into the culture of all disciplines of science. Science has always been self-correcting, but in the past the self-correction was left

to chance. In the future, we need a scientific enterprise with self-correction more intentionally designed into the system.

2. How Trust Is Built in the Geosciences

From my vantage in the geosciences, many aspects of the replicability issue have seemed somewhat remote. My own subfield, marine geophysics, has traditionally had a better policy on sharing data than has been the case for most other scientific disciplines (astronomy being a notable exception) on account of the shared nature of oceanographic ship time. In particular, data collected while the vessel was underway, such as bathymetry, gravity, and magnetic data was expected to be deposited in the National Geophysical Data Center (NGDC) after a period for quality control. Policies at funding agencies, such as the National Science Foundation (NSF), have long stipulated that all data collected with government funds be placed in public repositories within 2 years, although such policies were problematic because enforcement was absent unless the investigator was flagged as being out of compliance with the data policy while applying for a follow-on grant. Equally troublesome, public repositories did not cover all data types, and were best suited for those that could be reduced to a series of numbers. There were and still are notable exceptions to these policies, such as proprietary data acquired by industry, but in general marine geophysicists would have to make a case why data sharing was not the default option.

Thus, *reproducibility* (the act of checking someone else's work using their data to ensure that the result is robust) was not a challenge, although sharing of code was far less common. It was practically guaranteed that if another researcher questioned your data or your interpretation of the data, he or she would be able to check your conclusions against other possibilities. For example, decades ago when I first proposed that a massive area of abnormally shallow seafloor in the central Pacific Ocean related to excess heat deep within Earth (McNutt et al., 1987), several other laboratories were quick to check my work using the same archived data from the NGDC. Could my result be explained by improper corrections for sediment? Did my data processing adequately remove any bias from the many volcanoes on the average depth? Did the result depend upon what baseline model I used for the deepening of the seafloor with age? After nearly a decade, the matter was settled that the feature did exist. Along the way, we developed more robust methods of data processing that addressed some earlier concerns of potential sources of bias (McNutt et al., 1996).

On the other hand, *replication*—the re-creation of the exact same investigation from start to finish using new data collection and expecting the same result—was less common. Many phenomena in the geosciences are ephemeral and cannot be repeated: the Loma Prieta earthquake, the Tohoku tsunami, Hurricane Harvey. Even many standard Earth measurements cannot be expected to be exactly repeated because of secular variation. Mountains erode, climate changes, and biogeochemical cycles are evolving. The one exception is laboratory-based geoscience, such as the experiments performed on rock mechanics or geochemistry of Earth materials, which have always had strong traditions of intercalibration against known samples and strict quality control. Nevertheless, even field geoscientists develop confidence in their physical models proposed to explain field

measurements through *generalization*—the confirmation of the applicability of a physical model by testing it against new, independent data sets collected from different events or field areas. Geoscientists also obtain confidence in conclusions through a form of replication that avoids the nonstationarity of natural phenomenon, and that is by studying the same event using two independent methods, ideally with different sources of bias and error. An example would be measuring the degree to which the oceans have warmed from climate change by observing changes in the speed of propagation of acoustic waves in the SOFAR (Sound Fixing and Ranging) channel and, alternatively, by averaging the in-situ temperature data from National Oceanic and Atmospheric Administration’s global array of profiling drifters.

2.1. The Importance of Standards

One of my first encounters with the importance of maintaining standards in the geosciences happened at a meeting of the American Geophysical Union. The research group of A. B. Watts from Lamont-Doherty Earth Observatory presented a poster paper with an analysis of marine geophysical data quality from the archives of the NGDC. The team had computed the cross-over error in shipboard bathymetric soundings and marine gravity measurements when oceanographic research vessels from two different institutions passed over the same patch of seafloor. If the data quality control for the vessels of both institutions was high, the data should agree to within a reasonably small error, mostly explained by navigation uncertainty. What the analysis showed was that the cross-over errors were typically small, except for the comparisons with one particular oceanographic institution, which systematically failed to agree with the data collected by all of the others, indicating sloppy quality control. Government agencies took note. At the next round to consider funding of the ships for the oceanographic fleet, the NSF did not renew that institution’s bid to continue as a ship operator.

2.2. The Value of Benchmarking

Computer modeling is one area of replicability where the issues are common across most fields of science, including the geosciences, and this was my first encounter with what can go wrong when a community fails to establish benchmarks. It was the 1980s in the heyday of using computer codes to model convection in Earth’s mantle. Constraints on the physical parameters for the viscous flow equations inside a rotating, spherical Earth came from experiments on the properties of likely mantle mineralogy at ambient pressures and temperatures (performed in rock mechanics laboratories), and the predictions from the codes were compared with surface observations such as variations in the height of Earth’s equipotential surface (the geoid) to ascertain more information on the flow characteristics. The problem was that various expert groups at different institutions were deriving inconsistent results for the details of the convection amplitude, pattern, and thermal structure, and thus the predicted observables, using different methods (spectral, finite element, finite difference, finite volume), even using the exact same input parameters. This was a classic case of irreproducibility that was in danger of stifling progress in the field, as researchers did not know whose output to believe. The researchers had been so eager to explore novel scenarios with their codes that they had not paused to benchmark their performance against known standards. Because computing resources for large-scale simulations were a

precious resource at the time, much effort had been put into numerical optimization of code performance for speed. This left many wondering whether some ‘features’ in the output were merely numerical artifacts from code optimization.

To address the problem, researchers from all the major groups (Harvard, Caltech, Johns Hopkins, UCLA, MIT, etc.) running codes converged for a series of summer workshops at Los Alamos to look over each other’s shoulders as they intercompared their codes, step by step. They concluded that they had to start by benchmarking the easiest problems first, get the same answer, and then bootstrap their way up to gradually, piece by piece, add complexity until the real source of the discrepancy was found (Travis et al., 1990). Much of the difference was reduced to how well the advection term in the flow equations—velocity $\times$ grad(temperature)—was parameterized. One of the models did not program compressibility quite right. Greater understanding was accomplished in those few days in Los Alamos than ever could have been achieved by passing around preprints and manuscript reviews. The field moved forward.

Today the current best practice to encourage reproducibility is to share code. Designing code that can be easily shared is certainly a positive move for reproducibility, as it allows others to determine if they can obtain the same result that the original author did using the author’s data and code. However, it is not clear that sharing code would have solved the problem in replicability that I describe with the various teams modeling convection in Earth’s mantle. Another investigator could have run the same code with the same input conditions and verified the original author’s output, but that would not have provided any additional guarantee that the result reflected some valid feature about convection inside the planet. Nor would it have resolved discrepancies between different codes. The lesson here is that all disciplines should be alert for opportunities to set benchmarks and standards that allow researchers to test code or laboratory methods against known or validated results.

2.3. Replicability Versus Independent Validation

I experienced one of the more dramatic demonstrations of how replication might not be enough to gain trust in the geosciences during the *Deepwater Horizon* oil spill in 2010, the largest marine oil spill in U.S. history. At the time I was the director of the U.S. Geological Survey. I was dispatched by the secretary of the Interior to Houston to oversee a team of government scientists and engineers working with BP professionals to try to contain the oil and control the well. One of my assignments was also to estimate the flow rate for the failed well, although at the time there was no proven method for assessing the discharge rate of a well a mile deep beneath the ocean surface.

Different teams I convened deployed a variety of methods to estimate the flow rate of the well. Some methods relied on measuring the growth in area of the oil slick at the ocean surface over time. Another popular approach analyzed video camera footage to estimate the velocity of particles entrained in the fluid and the volume of the escaping hydrocarbons. Yet another relied on theoretical calculations using geophysical measurements from the

oil reservoir. While each method produced very consistent results by independent researchers (good replicability), none of the independent methods of measuring the flow rate agreed on the discharge rate of the well (questionable accuracy of any one method). It was not until a team from Woods Hole Oceanographic Institute deployed a current meter and imaging sonar from a remotely operating vehicle that we finally obtained flow rate estimates that converged with the higher rates from other methods, and we felt confident that we had a reliable estimate (McNutt et al., 2011). This experience illustrated the importance of being cautious about trusting results that have not been replicated by an independent method. The flow-rate estimates from analysis of the video data were both reproducible and replicable by the same method, but they consistently underestimated the flow rate of the well.

3. Furthering Reproducibility Through Journal Policies

By the time I took on the position of editor-in-chief of the *Science* family of journals, a bit less than 10 years ago, the lack of replicability had become a serious issue in biomedical sciences and psychology, and thus I made addressing replication an important part of my agenda as editor-in-chief. My concern at the time was that if the scientific community itself did not confront this problem head-on and in a forthright manner, it had the potential to seriously damage the reputation of science and scientists, a group that perennially is near the top of the list of the most trusted voices in America. I further reasoned that if we did not find solutions to whatever factors might be leading to lack of replicability, then solutions could be forced onto the community from the outside that might be less compatible with the culture of researchers and their institutions. Thus began a fruitful partnership with other top journals, such as *Nature*, and organizations, such as the Center for Open Science and the National Institutes of Health. With funding from the Laura and John Arnold Foundation, *Science* organized a series of workshops on the replicability issue, the most consequential of which resulted in the TOP (Transparency and Openness Promotion) guidelines (Nosek et al., 2015) for promoting reproducibility and replicability. TOP's eight modular standards allow journals to opt in at three different levels of rigor in their requirements for author transparency in the availability of data, materials, code, and methods, in the preregistration of the analysis plan or the study, and in publication of replications. To date these standards have been implemented by more than 1,100 scientific journals and are expanding to universities.

Having a consistent set of clear standards that applied across all fields actually made my job easier as editor-in-chief. I was able to hold authors accountable for making data, materials, and other essential information available for reproduction and replication, except for the few papers that received waivers on account of concern for personal privacy or another similar overriding factor. If an author had failed to follow the TOP protocol, simply the threat of attaching an "Editorial Expression of Concern" to the publication on account of its lack of transparency would be enough to motivate compliance.

While I was editor-in-chief at *Science*, we didn't just publish the TOP standards but just one month later also published the first meta studies exploring the limits in replicability in 100 important papers in psychology (Open Science Collaboration, 2015). This experience was my first encounter with the challenge of how to

decide when replication achieved a result close enough to that of the original that one can decide that the replication is a success. While nearly all of the original studies obtained significant results, just 36% of the higher powered replications produced significant results and 39% of effects were subjectively judged to have supported the conclusions of the original study. The strongest predictor of which studies replicated was the strength of the evidence in the original study. I also gained an appreciation for how interested the research community is in the trustworthiness of published studies: this landmark paper has been cited more than 4,600 times to date.

Despite the progress made on promoting transparency and open science practices that aid in reproduction and replication, journals would never have the resources to ensure compliance with journal policies. I recall one editor of a rather small journal that published just a few hundred articles per year commenting that it was a full-time job for one person to assess compliance with their open data policy. Do the links in the paper to the underlying data work? Are they directed to a valid repository (as opposed to the author's webpage)? Is the archived data indeed the data upon which the paper is based? Is all of the metadata available as well? Open science practices needed to become the default behavior of the research community rather than one more onerous requirement imposed by journals.

My experience on the publishing side also exposed the bias in what gets published. Null results are less likely to be published in top journals, although with the expansion of open access journals most find an outlet somewhere. Splashy findings that might not stand the test of time are more likely to catch both the editor's and the reader's eye than a very rigorous, careful study that is not paradigm-shifting. However, the latter study might be exactly that solid foundation that an entire new scientific pillar can be built upon. I understand calls for not wanting to unduly punish authors whose work fails to reproduce. After all, if that were to happen, authors would stop making data, materials, code, and methods available for reproduction out of fear of negative consequences. However, every community knows those in their field whose work is iron-clad, well-done, and totally trustworthy. We need to honor those scientists more and encourage students to emulate them.

4. The National Academies Report

All of these early efforts culminated when I was elected president of the National Academy of Sciences in 2016, and soon after had the opportunity to launch a formal consensus study at the direction of Congress on this topic, funded by the NSF and the Alfred P. Sloan Foundation. The all-star committee, ably chaired by Harvey Feinberg, president of the Gordon and Betty Moore Foundation and former president of the Institute of Medicine, delivered its report several years later titled *Reproducibility and Replicability in Science* (National Academies of Sciences, Engineering, and Medicine [NASEM], 2019). While I deserve no credit for the substance of this report, I take some satisfaction in the observation that many of the report's findings and recommendations mirrored my experiences with the subject and provided a reasoned path forward for improving the conduct of science.

For example, the report makes the distinction between helpful and not helpful sources of irreproducibility and failures to replicate. Numerous researchers over the years relayed to me situations in which a failure to replicate revealed a factor that everyone had thought was inconsequential to the outcome of an experiment was in fact an independent variable—for example, type of bedding in mouse cages impacting pharmacological studies, gender of the experimenter impacting the tolerance for pain in laboratory animals (or optimization of code impacting output!).

The report also places an emphasis on the community tightening up its standards to reduce the nonhelpful sources of irreproducibility, caused by poor study design, sloppy execution, and incomplete communication of all the information necessary to repeat the study. Some laboratories have a tradition of asking incoming students to replicate the results of graduating students. This practice encourages high quality in experimental results and life-long good habits in experiment execution.

Next, the report's focus on computational reproducibility is surely well placed. One of the most thorny workshops in the series on transparency that *Science* hosted during my tenure there was on this topic (Stodden et al., 2016). The issue had only become more complex since my first encounter with it during the days of benchmarking mantle convection codes. Scientists now predict complicated phenomena using an array of home-grown, commercially available, open-source, and proprietary codes. It may not even be legal to move the proprietary software to another machine or make it publicly available. Furthermore, some code is optimized for use on specialized processors that are not available for general use, and the code cannot be guaranteed to produce consistent results if ported to another computing platform. Not all scientists are even informed users of the commercial packages they use (e.g., statistical analysis software). The report recommends everything from simply more attention to version control to depositing of archival versions of source codes on common platforms such as GitHub.

The recommendations contained in the NASEM report serve to make laboratory, field, and computational aspects of science more trustworthy and more robust. Ideally, I would like to see aspects of this report incorporated into all methods courses taught at universities in order to train students in good habits for obtaining reliable results that stand the test of time.

Perhaps one of the most enduring contributions of the report is the recommendation for investment in open-source tools and infrastructure to create transparency and enable reproducibility as a natural part of sound science. This recommendation mirrors my own experience at *Science* that the effort to publish research that is easily verified is best achieved when we make transparency and open science practices easy to do and an integral part of the culture of science.

5. Is There a Replication Crisis?

There is an obvious connection between the open science movement and the use of replication as a tool to assess the validity of scientific findings. Some researchers have noted that the lack of replicability of research

did not emerge as an issue until the late 1990s or perhaps the 21st century, and have attempted to place the blame on the growth of the research enterprise and attendant pressures on researchers to cut corners and push for sensational results. Although there have been some well-publicized cases of research failing to replicate because of outright fraud on the part of the researcher, such cases are still rare.

I would argue instead that it was far more difficult to ascertain prior to the open science movement whether studies replicated or not. Decades ago research may also have been difficult to replicate (although surely experiments have become more complex), but because data, materials, and methods were less easily shared, it was impossible to show that a certain study had been exactly replicated but with a different result. Thus, the rise of the movement toward open science has enabled replication, and revealed what actually may have been a long-standing situation. And if the situation has indeed been long standing, it is quite difficult to argue that we suddenly have a crisis.

Long ago, as now, the research community has built up a healthy degree of skepticism concerning the results of any single study unless verified by another research group, preferentially using alternative investigation methods and generalized to other systems. A single study is merely a suggestion of a promising direction that is worth verifying, using best practices such as preregistration of the study design and the analysis plan to help control for bias. Funder mandates and journal policies to promote openness and transparency are certainly an improvement in creating an environment that could allow the self-correction tradition of science to proceed in a more efficient and orderly way. In the current age of the multimillion-dollar science experiment, I would even argue that we, as scientists, have an obligation to use all of the transparency and openness tools at our disposal to make the self-correcting process of science proceed as quickly and efficiently as possible so that we are excellent stewards of precious resources.

One corollary of this hypothesis is that as the research enterprise becomes ever more open, even more replication failures will be uncovered, and even more harm will be done to the trustworthiness of science unless efforts to bolster the rigor of experimental protocols and eliminate bias from interpretations are successful. A second corollary is that we must develop the open-source culture, tools, and infrastructure called for in the NASEM report if we want to build trust in our research enterprise.

6. Conclusion: Establishing a Strategic Council

The report from the National Academies has laid out a roadmap for a holistic approach to achieve that “self-correction by design” that I called for in the introduction to this perspective. We must go beyond haphazard use of standards, benchmarks, and calls for openness instead transparency. While in the past the admirable efforts of individual research disciplines, funders, journals, and research institutions have had an impact, at this time they all must pull together. To help achieve this integrated action by all stakeholders I have asked the National Academies to set up a Strategic Council to help guide this effort. The goal is to provide a seat at the table for those entrusted with the quality and integrity of the research enterprise from its inception at the funding stage

to its eventual dissemination and potential incorporation into actions and public policy. The Strategic Council can provide a venue for coordinating requirements across the different stages of the research process and sharing best practices. It can serve as a sounding board for new ideas and a support group for implementing the best of them. It will break barriers, catalyze progress, and anticipate problems to safeguarding the effectiveness of the research enterprise. The Academies look forward to working with the research community on this critical effort.

Acknowledgments

Thank you to Scott King and Brad Hager for jogging my memory on important details concerning the mantle convection workshops held so many years ago. I also thank Brooks Hanson and an anonymous reviewer for comments on an earlier draft of this manuscript.

Disclosure Statement

Marcia McNutt has no financial or non-financial disclosures to share for this article.

References

- McNutt, M. K., Camilli, R., Crone, T. J., Guthrie, G., Hsieh, P. A., Ryerson, T., Savas, O., & Shaffer, F. (2012). Review of flow rate estimates of the *Deepwater Horizon* oil spill. *PNAS*, 109(50), 20260–20267. <https://doi.org/10.1073/pnas.1112139108>
- McNutt, M. K., & Fisher, K. M. (1987). The South Pacific superswell. In B. Keating, P. Fryer, R. Batiza, & G. W. Boehlert (Eds.), *Seamounts, islands, and atolls* (pp. 25–34). Geophysical Monograph #43. American Geophysical Union. <https://doi.org/10.1029/GM043p0025>
- McNutt, M. K., Sichoix, L., & Bonneville, A. (1996). Modal depths from shipboard bathymetry: There IS a South Pacific superswell. *Geophysical Research Letters*, 23(23), 3397–3400. <https://doi.org/10.1029/96gl03053>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. <https://doi.org/10.17226/25303>
- Nosek, B. A. et al. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., & Tauber, M. (2016). Enhancing reproducibility for computational methods. *Science*, 354(6317), 1240–1241.

<https://doi.org/10.1126/science.aah6168>

Travis, B. J., Anderson, C., Baumgardner, J., Gable, C. W., Hager, B. H., O'Connell, R. J., Olson, P., Raefsky, A., & Schubert, G. (1990). A benchmark comparison of numerical methods for infinite Prandtl number thermal convection in two-dimensional Cartesian geometry. *Geophysical & Astrophysical Fluid Dynamics*, 55(3–4), 137–160. <https://doi.org/10.1080/03091929008204111>

Disclosure Statement

Marcia McNutt has no financial or non-financial disclosures to share for this article.

©2020 Marcia McNutt. This article 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 article.