Toward Reproducible and Extensible Research: From Values to Action

Aleksandrina Goeva¹ Sara Stoudt^{2,3} Ana Trisovic⁴

¹Macosko Lab, Stanley Center for Psychiatric Research, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, United States of America,

²Berkeley Institute for Data Science, University of California Berkeley, Berkeley, California, United States of America,

³Statistical & Data Sciences Program, Smith College, Northampton, Massachusetts, United States of America,

⁴Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Dec 16, 2020

DOI: <https://doi.org/10.1162/99608f92.1cc3d72a>

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

ABSTRACT

The National Academies report [Reproducibility and Replicability in Science](#) outlines values and broad goals for reproducible and replicable scientific research, emphasizing the importance of its availability and transparency. Confronted with these goals, we examine what it means for research material to be transparent, available, and reproducible. As early-career researchers, we advocate for the focus to be placed on reusable and extensible research, which is naturally achieved when research material is reproducible, transparent, and open. When we understand the inner workings of a study, have access to its inputs and outputs, and are able to achieve the same results ourselves, we can more easily reuse, extend, and generalize prior efforts. To achieve the outlined reproducibility goals, one needs actionable steps that facilitate implementation and help mark progress in practice. We discuss concrete guidelines and checklists as tools for actionable reproducibility. We emphasize how a set of hierarchical recommendations that give an order of importance to the various aspects of reproducibility, transparency, and openness would help researchers facing limited resources prioritize accordingly. Finally, we discuss how training, new roles focusing on reproducibility, and supportive work environments would be beneficial.

Keywords: reproducibility, replicability, extensible research, transparency, openness

1. Introduction

In recent years, the terms "open data," "reproducibility," and "replicability" have become, to a certain extent, buzzwords. In particular, the term "reproducibility crisis" has been used to raise concern about research practices and motivate new development of tools that work to abate the crisis. Markedly, Baker's survey titled "1500 Scientists Lift the Lid on Reproducibility" (2016) has been cited more than 1500 times since it appeared in 2016 (*1500 Scientists Lift the Lid on Reproducibility* - Google Scholar, n.d.), emphasizing a growing interest in reproducibility and replicability in science.

The timely release of the National Academies report, "Reproducibility and Replicability in Science," (National Academies of Sciences, Engineering, and Medicine (U.S.) et al., 2019) in 2019 clarifies expectations and realities around the scientific method. Authored by a committee of experts, it introduces definitions, explores issues across science and engineering, and assesses their impact on public trust. It provides recommendations for improving rigor and transparency in research that would be valuable to researchers, funders, academic institutions, and journals alike.

Figure 1. The report Reproducibility and Replicability in Science (National Academies of Sciences, Engineering, and Medicine (U.S.) et al., 2019) as a word cloud (Trisovic, 2020).

As early-career researchers, we often contribute heavily to the technical and computational aspects of research projects, and while we understand the challenges of computational reproducibility, we look forward to guidance on how to address them. The Academies' report outlines values and broad goals for reproducible and replicable scientific research (see word cloud visualization of the report in Figure 1). As we try to translate these values into our own practice, we contribute some additional nuances on the outlined terms and share our ideas on the next steps that we would find valuable as early-career researchers.

2. Sharing Research Effectively Takes More than Reproducibility

Acknowledging that science has become increasingly computational, the report defines reproducibility as "obtaining consistent results using the same input data, computational steps, methods, code, and conditions of analysis." The committee also states that "a study's data and code have to be available in order for others to reproduce and confirm results." Making research data and code available to the community for examination and critique is the core principle of openness (Peters, 2010) and it is often linked to the open science movement. The committee also views "transparency as a prerequisite for reproducibility," though

“transparency” in research is generally applied to different stages of the process, such as reporting (Weissgerber et al., 2016), artifact sharing (Nosek et al., 2015), and peer-review (Wolfram et al., 2020).

Despite all these terms being tied to research sharing, we want to emphasize that reproducibility, openness, and transparency do not perfectly overlap when applied to research materials (see Figure 2). In particular, we see research material as transparent only when it is well-documented and understandable to the point that it could be reused for education or new research. Ever-increasing data volumes and intricate code can be captured as an executable “black box” in an automatized workflow, virtual container, or a package that reproduces results with a single command. Despite these tools facilitating reproducibility, they may decrease researchers’ access to the underlying processes and hence their transparency. In a similar vein, research data and code may be unintelligible or shared without comments or documentation, however, as long as they produce reported results, they, by definition, enable reproducibility. Alternatively, research materials can be reproducible and transparent but not open, when they are shared ‘by request’ or when published results are based on sensitive data. These instances, in principle, could be verified reproducible by third-party services behind ‘closed doors.’ Finally, results can be open and transparent but not reproducible due to, for example, missing dependencies or unreported random seeds.

Research material that is the trifecta of reproducible, transparent, and open should be considered a gold standard for dissemination. Such material is not only reproducible but also the easiest to learn from and build upon. We believe that the focus should be placed on reproducible as well as reusable and extensible research. When we understand the inner workings of a study, have access to its inputs and outputs, and are able to achieve the same results ourselves, we can more easily think about reusing and extending prior work and its data and code. We use “extensibility” to refer to the ability to extend software code (and likewise other research materials) by adding new functionality or modifying existing functionality and “generalizability” to refer to results of a study applying in other contexts or populations that differ from the original one.

Figure 2. Nuances in reproducible, open, and transparent research material.

The Venn diagram illustrates three different scenarios: (1) Research data and code can be freely available online, with thorough documentation, but if code breaks due to, for example, missing dependencies, or an unreported random seed, it will not be reproducible. (2) Research data and code can be transparent and reproducible but only shared “by request” or with a fee. (3) Research data and code are available online and reproducible, but they could be unintelligible due to a lack of documentation.

3. Synthesizing Valuable Insight from Both Discoveries and Null-Studies

Although computational reproducibility, with focus on code and data sharing, is often the star of the show, decisions made throughout the investigative and analysis process are also valuable in effectively documenting the outcomes. In particular, they are key in enabling replicability, which the report defines as “obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.” The report acknowledges that “many decisions related to data selection or parameter setting for code are made throughout a study and can affect the results,” and recommends researchers to “capture these decisions and include them as part of the digital record.” These recommendations echo the hallmarks of proper inference outlined as a principle in the ASA Statement on Statistical Significance and P-Values (Wasserstein & Lazar, 2016).

However, even if these choices are recorded, there is no guarantee that they were not cherry-picked after testing multiple values (or even multiple modeling approaches) and selected based on which one yielded the

most favorable results. For example, if multiple analyses are performed on the data, but only the ones with a p -value below a certain threshold are reported, both validity and replicability of the finding might be compromised, contributing to a spurious excess of statistically significant results in the published literature (Wasserstein & Lazar, 2016). In the current publication culture, researchers are incentivized to engage in such activities because studies with statistically significant results are more likely to get published and more likely to be cited than studies with null findings (Jannot et al., 2013).

We believe that scientific knowledge is enriched by both positive and null-result studies, and in particular, that all well-executed studies are worth sharing. Publishing a null-result study helps reduce repetition of the study by another group and may help us refine our research questions. Experimental high-energy physics is an example of a community that commonly publishes null-result studies, referring to them as “searches” (Junk & Lyons, 2020). These values could be adopted more widely, if for example, publishing bodies include pre-registering studies in their submission and review process and publish articles based on the premise and regardless of whether the findings are significant or not, given that the study is carried out rigorously (e.g. *Journal of Development Economics [JDE Pre Results Review Papers]*, n.d.).

The report finds that “rare, unexpected, or novel events and topics are much more likely to be covered by news media,” which also influences researchers’ incentives and warps expectations of science. As the report suggests, a solution may be in facilitating and publishing meta-analyses that study a multitude of scientific investigations on the same topic in prestigious journals that journalists often mine for story ideas. Researchers can then present a united front, working as a team with journalists to report on the nuances of these synthesis studies rather than competing to get the latest one-off study into the public eye. Similarly, public policy officials looking to shore up their decisions will be nudged in the direction of more stable findings. Such meta-analyses highlight that science is iterative, requiring multiple studies on a similar topic before a consensus is reached. In addition to emphasizing and promoting the importance of replicable results, they also provide a higher-level picture of scientific developments and make them understandable to a wider audience. For example, if the same result is reached via different approaches, that makes the finding even more robust.

Finally, communicating some less formal aspects of the scientific process could sometimes enhance the replicability and extensibility of a study’s findings. Questions like what motivated the authors to pursue their research; how they got inspired to create their models; and what unexpected, strange or unsuccessful approaches they tested and why they failed, are sometimes left behind the scenes. When shared, these insights could be useful to students and early-career researchers hoping to replicate, reuse, or extend the study. Therefore, it would be valuable to facilitate the sharing of the less formal aspects of the research process, for example, through blog posts, interviews, or podcast conversations.

4. Need for Actionable Reproducibility

The report urges researchers to convey "clear, specific, and complete description of how the reported result was reached" to enable reproducibility and replicability. We stand behind these values, and as early-career researchers, we seek concrete action-items to help us apply them in practice. For example, computer scientist Joelle Pineau has led an effort to create a machine learning reproducibility checklist with a set of criteria for reporting work, focusing on models and algorithms, theory, datasets, code, and experimental results (Pineau, 2019). The Neural Information Processing Systems (NeurIPS) conference introduced a reproducibility program in 2019, which used the checklist as part of the paper submission process. The reproducibility program also included a code submission policy and a community-wide reproducibility challenge. The program had several positive outcomes, such as an increase in code submissions, and it found that the reproducibility checklist was useful to both authors and reviewers (Pineau et al., 2020).

Checklists are simple yet powerful tools that are already proven to be valuable in high-risk environments such as aviation and surgery. Therefore, a worthwhile extension of the report could be translating its recommendations into a general-purpose reproducibility checklist of action items that could then be customized for specific fields. Furthermore, the checklist could feature a set of hierarchical recommendations that give an order of importance to the various aspects of reproducibility, transparency, and openness. Such ordering would help researchers facing limited resources prioritize accordingly. An action-oriented approach with achievable steps would both motivate us and measure progress towards the gold standard.

In a similar step towards action, it would be useful to identify and recommend an optimal time frame for research dissemination and archiving. For instance, researchers may be hesitant to share their data due to the fear of being 'scooped,' or their institution may be unable to fund a long-term infrastructure for open data. A general recommendation that would state that for highest impact, research materials should be made available, for example, up to one year after publication, and should be available for a minimum of ten years, could provide achievable goals to both the researchers and their institution. The data embargo period of one year would enable a researcher to fully explore and share their data without the fear of being scooped. In comparison, a ten-year availability period would help an institutional repository advocate for the necessary funding. Therefore, such a concrete guidance for longer-term access would help facilitate research openness, and likely its reproducibility and extensibility.

While a checklist and an envisioned time frame would provide researchers with specific tasks for meeting reproducibility goals, comprehensive training would help researchers acquire an understanding, skillset, and confidence in navigating the systems for sharing data and code. The report states that "researchers need to understand the importance of reproducibility, replicability, and transparency, to be trained in best practices, and to know about the tools that are available," and that "adequate education and training in computation and statistical analysis is an inherent part of learning how to be a scientist today." We commend and reiterate this

statement, and also advocate for a communication training that would help the researchers share their findings with broader audiences in an accessible way.

Developing content for such specific training, while also creating reproducibility checklists, tools, and infrastructure could be a full-time job itself. There is important work to be done investigating the limitations of reproducibility and consolidating the existing developments with new technical requirements. Therefore, academic and other research institutions could consider creating designated roles revolving around reproducibility, replicability, and generalizability of research (Meng, 2020). Above all, it is crucial to foster supportive work environments, and by introducing achievable well-defined tasks, specific training, and new roles for reproducibility experts, workplace environments may become more constructive, collaborative, and would sustainably support progress toward the gold standard across the scientific fields.

Figure 3. A reproducibility strategy. Illus. Aleksandrina Goeva.

5. Conclusions

The "Reproducibility and Replicability in Science" report has initiated a timely conversation on scientific methodology and it introduces values to carry forward in our research. As early-career researchers, we often start our projects by reproducing or replicating previous results and building upon prior studies. Therefore, in addition to reproducibility, we believe that there should be an increased focus on its intersection with transparency and openness that would allow prior research to be easily extended. Additionally, we see value in communicating non-significant results, attempted but failed approaches, and the less computational and more human aspects of research, beyond the bare minimum that ensures reproducibility.

Generating reproducible and extensible research is often the responsibility of the authors of the study alone, and it is likely a time-consuming process. Creating actionable goals and new roles for experts to advise on the matter, would both motivate and help researchers invest their limited time and resources into strengthening the reproducibility, transparency and openness of their work. Finally, high priority should be placed on developing

supportive work environments that provide the needed training, constructive feedback, and credit for all contributions toward more reproducible and extensible research.

Disclosure Statement

Aleksandrina Goeva, Sara Stoudt, and Ana Trisovic have no financial or non-financial disclosures to share for this article.

References

- 1500 scientists lift the lid on reproducibility. (n.d.). Google Scholar. Retrieved November 24, 2020, from https://scholar.google.com/scholar?hl=en&as_sdt=0%2C22&q=1500+Scientists+Lift+the+Lid+on+Reproducibility&btnG=
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454. <https://doi.org/10.1038/533452a>
- Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology*, 66(3), 296–301. <https://doi.org/10.1016/j.jclinepi.2012.09.015>
- JDE Pre Results Review Papers. (n.d.). Retrieved November 24, 2020, from <http://jde-preresultsreview.org/>
- Junk, T. R., & Lyons, L. (2020). Reproducibility and replication of experimental particle physics results. *Harvard Data Science Review*, 2(4). <https://doi.org/10.1162/99608f92.250f995b>
- Meng, X.-L. (2020). Reproducibility, replicability, and reliability. *Harvard Data Science Review*, 2(4). <https://doi.org/10.1162/99608f92.dbfce7f9>
- National Academies of Sciences, Engineering, and Medicine (U.S.) (Eds.). (2019). *Reproducibility and replicability in science*. The National Academies Press. <https://doi.org/10.17226/25303>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., & others. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Peters, M. A. (2010). Openness, Web 2.0 technology, and open science. *Policy Futures in Education*, 8(5), 567–574. <https://doi.org/10.2304/pfie.2010.8.5.567>
- Pineau, J. (2019). *Machine learning reproducibility checklist*. <https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf>

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2020). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 Reproducibility Program). *arXiv*. <https://doi.org/10.48550/arXiv.2003.12206>

Trisovic, A. (2020). Word cloud of reproducibility and replicability in science [Data set]. *Harvard Dataverse*. <https://doi.org/10.7910/DVN/HOLVXA>

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>

Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., & Prager, E. M. (2016). Transparent reporting for reproducible science. *Journal of Neuroscience Research*, 94(10), 859–864. <https://doi.org/10.1002/jnr.23785>

Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: Promoting transparency in open science. *Scientometrics*, 125(2), 1033–1051. <https://doi.org/10.1007/s11192-020-03488-4>

©2020 Aleksandrina Goeva, Sara Stoudt, and Ana Trisovic. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.