

# Demonstrations of reproducibility are not replicability



**R**eplicability and reproducibility are core tenets of science as they inform the reliability, robustness and generality of results. Given their importance, journals have set forth measures that aim to promote them. We argue that many recent measures have been adopted rapidly with little justification, fail to achieve these aims, have costs for scientists and possibly undermine efforts towards replicability and reproducibility.

All forms of reproducibility or replicability aim to determine whether we obtain the same answer when we ask the same scientific question. We define computational reproducibility as the ability to reproduce a set of quantitative results using the data and code from the original publication; analytical reproducibility as obtaining the same results using the original data and following the same analytical methods with an independently produced code; and conclusive reproducibility as when the same conclusions are reached in an alternative valid analysis of the same data (Fig. 1). The term replicability is used when a study repeats the collection of new data. Although we use this broad definition of replicability, it is often divided into direct replicability, which follows the same methods and species, or conceptual replicability, which uses different methods and/or species (Fig. 1).

Over a decade ago, mandatory data archiving began to be implemented in several journals in ecology and evolution<sup>1</sup>. Mandatory data archiving provides clear benefits, including facilitating reproducibility and promoting assessment of replicability via meta-analysis, making raw data available for other uses, and allowing detection of rare instances of fraud. The past several years have seen many journals take additional measures in the name of reproducibility and replicability. These measures, in their extreme, have aimed to achieve the mandatory archiving of code that runs on any computer<sup>2</sup>, with detailed metadata files and other requirements that have progressed to a state in which computational reproducibility should be verifiable by a non-expert with the click of a mouse, and without having to even read the original paper (Supplementary Table 1).

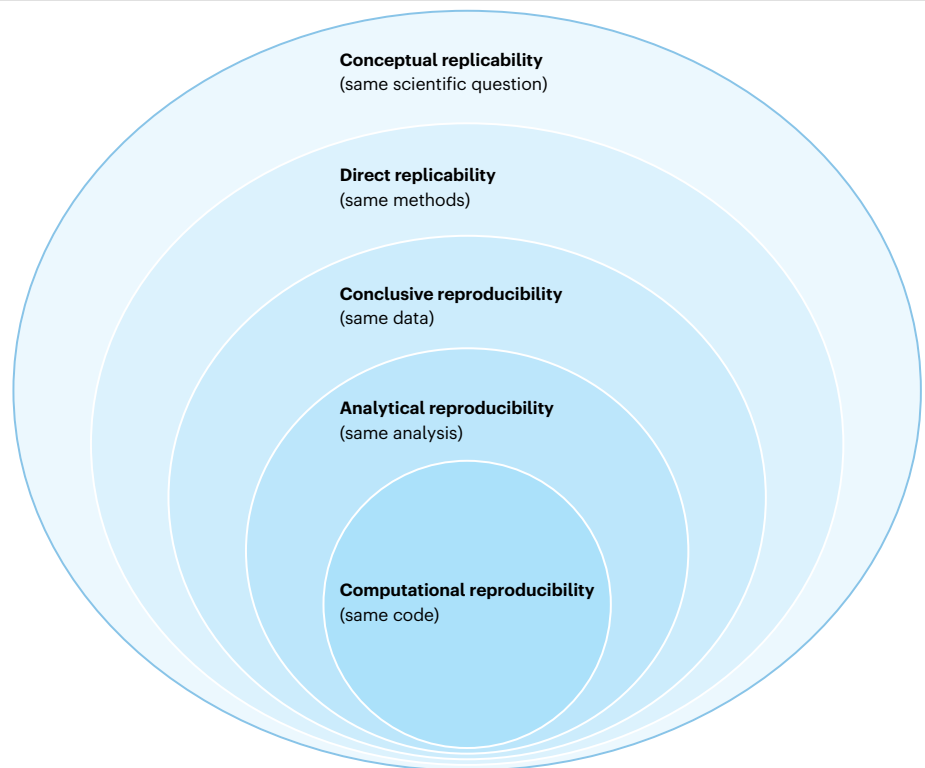


**Fig. 1 | Levels of replicability and reproducibility.** Fundamentally, all levels of replicability and reproducibility aim to assess whether an answer to a scientific question remains unchanged when we repeat or replicate a study or some aspect of it. At the lowest level, computational reproducibility aims to assess whether the exact quantitative results are reproduced by rerunning the computer script. At the highest level, conceptual replicability aims to assess whether we obtain the same general answer to the scientific question when we obtain new data using different methods (such as a different species or experimental design). Although there is a natural nesting to these levels of replicability and reproducibility, they are not necessarily dependent: for example, a quantitative result could be non-reproducible computationally because of non-seeded stochastic sampling reflecting an approximation of *P* values, yet the result could be reproducible and replicable at all of the other levels.

We think code archiving is good for science and that authors should be encouraged to make computer scripts available. Moreover, computational reproducibility is fundamentally important, and something that every scientist should (and probably does) verify while performing an analysis. However, when considering the push for elaborate demonstrations of computational reproducibility, it is worthwhile to consider that this is the most trivial level of reproducibility to enforce

demonstrations of, in the broader context of scientific progress. There are three reasons for this: computational reproducibility does not ensure an analysis was performed correctly or is appropriate to the question or data; there is no guarantee of forward and backward compatibility of the software or packages used; and it sometimes places emphasis on statistically nonsensical aims, such as exact replication of a stochastic Markov chain or bootstrap, and gives an air of exactness to values that are

approximations subject to sampling error. In fact, most editorial efforts focus only on computational reproducibility or have no foreseeable direct effect on reproducibility nor replicability (Supplementary Table 1). Although we agree that researchers should take great care to ensure computational reproducibility in their work, we argue it is worth considering that mandating demonstrations of reproducibility might better focus on replicability, conclusive reproducibility and analytical reproducibility, which test the robustness and generality of results.

In parallel, such editorial measures can have tangible costs. On top of the time investment in producing performative documents, scientists collectively spend 100 million hours in peer review<sup>3</sup>, and we hazard that code itself is unlikely to be carefully reviewed in many cases. Alternatively, many journals now have ‘code editors’, who typically ensure that code is annotated and runs. We suggest that resources directed to achieve public demonstrations of computational reproducibility could instead be focused on achieving other forms of reproducibility or replicability. For example, an alternative to a code editor would be a statistical editor who attempts analytical reproducibility, at least for some manuscripts (such as those receiving mixed reviews or a random selection). Alternatively (or in addition), authors could be required to present an alternative valid analysis of particularly

complex datasets to ensure conclusive reproducibility. Making such a requirement would do far more to ensure robustness of published results than, for instance, enforcing meta-data file formats. Journals can also actively solicit and publish studies that aim to attempt replication<sup>4</sup>, and some are now doing so (for example, *Ecology and Evolution*).

Mandating and micromanaging the process of code sharing might also drive costs that limit the progress of science and, ironically, limit reproducibility. Code and analytical errors have often been revealed when an independent analysis (analytical reproducibility) fails to yield the same conclusions as the original study (for example, see ref. 5). Moreover, much of these efforts at micromanaging the computational practices of the field might lead to a false sense of accomplishment and to the belief that we have done much to tackle a ‘replicability crisis’ (for which arguably little evidence exists (Supplementary Information)). Finally, this focus on making computational reproducibility effortless for others to assess might lead to the assumption that this is the most important level of reproducibility or replicability to be assessing.

The focus of most current efforts does little to advance the main reasons why we need replicability in science in the first place. We believe future efforts should take care when considering explicitly how they may promote, assess or demonstrate meaningful

reproducibility and replicability that advances progress of ideas in ecology and evolution.

**Miguel Gómez-Llano**  & **Stephen P. De Lisle**

Department of Environmental and Life Sciences, Karlstad University, Karlstad, Sweden.

✉ e-mail: [miguel.gomez@kau.se](mailto:miguel.gomez@kau.se)

Published online: 13 February 2026

## References

1. Moore, A. J., McPeck, M. A., Rausher, M. D., Rieseberg, L. & Whitlock, M. C. *J. Evol. Biol.* **23**, 659–660 (2010).
2. Bolnick, D. et al. Guidelines for archiving code with data. *amnat.org* <http://comments.amnat.org/2021/12/guidelines-for-archiving-code-with-data.html> (2021).
3. Aczel, B., Szasz, B. & Holcombe, A. O. *Res. Integr. Peer Rev.* **6**, 14 (2021).
4. Palmer, A. R. *Annu. Rev. Ecol. Evol. Syst.* **31**, 441–480 (2000).
5. Munday, P. L. *PLoS Biol.* **20**, e3001809 (2022).

## Acknowledgements

We thank L. Rowe, J. Stinchcombe, I. Moodie, D. Wilkinson, A. Sánchez-Tójar and S. Georganos for comments on the manuscript. S.P.D.L. is funded by the Swedish Research Council (VR 2024-04599), Formas (2021-01096) and KK-stiftelsen (SFSG-2024).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41559-026-03000-w>.