Is reproducibility good enough?

Ivan Stordal

Tuesday 23 Sep, 2025

Abstract

The credibility of science depends on whether research findings are robust, reliable, and reproducible. Over the past two decades, several fields have faced what is now called the "replication crisis" where publication bias, small samples, and flexible analysis practices have weakened confidence in results. This paper builds on the lecture Robust and Reliable Science and reviews literature on reproducibility and replicability, highlighting how large replication studies and methodological critiques have revealed shortcomings in current practices. The paper argues that computational reproducibility should be treated as a basic requirement for credibility, while full replication is more demanding and often limited by resources. Tools such as Quarto and GitHub, combined with journal policies, preregistration, and better training in open science practices, provide promising ways forward. Although progress has been made, reproducibility is still not sufficient. Stronger standards and incentives are needed to move science toward greater robustness and reliability.

1 Introduction

The trustworthiness of research relies on whether results can be reproduced and verified. In the last two decades, many disciplines have faced growing concern over the "replication crisis", raising doubts about the reliability of published findings (Ioannidis, 2005). Reproducibility refers to obtaining the same results using the original data and methods, while replicability means validating those results with new data (Goodman et al., 2016). Both are essential for cumulative science, yet evidence shows that current practices often fall short.

This paper asks whether reproducibility today is "good enough." Drawing on the lecture Robust and Reliable Science and key sources, it reviews problems such as publication bias, Type I errors, and replication failures in psychology and economics. It then discusses potential solutions including preregistration, data and code sharing, and technical tools such as Quarto. The goal is to evaluate the current state of reproducibility and outline what is still needed to make science more reliable.

2 Literature review

The distinction between reproducibility and replicability has become central to debates about research quality. Reproducibility means running the same

analysis on the same data to obtain the same result, while replicability requires re-testing findings on new material (Goodman et al., 2016; Peng, 2011). Scholars often describe reproducibility as the baseline for credible work, with replicability as the gold standard (Munafò et al., 2017).

One of the main barriers to reliability is publication bias, sometimes called the "file drawer problem" (Rosenthal, 1979). Positive and surprising results are more likely to be published, which creates a distorted scientific record. Simmons et al. (2011) point out that this issue is worsened by the risk of Type I errors, especially when researchers have in how they collect or analyze data. Studies show that these practices inflate reported effect sizes, reducing the trustworthiness of the literature (Button et al., 2013).

Large replication projects illustrate the scale of the problem. The Open Science Collaboration (2015) tried to replicate 100 studies in psychology and succeeded in reproducing only about one third. In economics, Dewald et al. (1986) showed that many empirical results could not be reproduced because data or code were missing. They used data from the 1982 *JMCB Data Storage and Evaluation Project*.

Proposed solutions include stronger journal policies requiring authors to provide code and data, sometimes linked to permanent repositories (e.g., Barnes, 2010; Bechhofer et al., 2013). Complementary reforms such as preregistration and registered reports aim to reduce questionable research practices (Nosek et al., 2015). Technical frameworks also play a role: literate programming (Gentleman, 2005; Xie, 2015), practical guidelines for computational reproducibility (Sandve et al., 2013), and newer publishing systems like Quarto (Allaire et al., 2020) make it easier to integrate code, data, and text into workflows.

3 Discussion of the reseach question

Should replicability be the norm?

Most researchers agree that computational reproducibility should be the minimum requirement for published work (Peng, 2011). Replication, however, is more expensive and not always feasible, especially in large or sensitive studies. A pragmatic approach is to enforce reproducibility across the board, while prioritizing replication in high-stakes areas such as clinical trials or economic policy (Munafò et al., 2017).

Can Quarto documents help?

Quarto provides a promising way to make research workflows more transparent. By combining code, narrative, and results in one document, it reduces the risk of errors from copy-pasting (Wickham & Grolemund, 2016; Xie, 2020). When linked with GitHub, it also makes collaboration and peer review easier. Experience from teaching shows that adopting such tools can lower the barrier to working openly and reproducibly, especially for students and early-career researchers (Broman & Woo, 2018).

What problems remain?

Despite technical progress, cultural and institutional barriers remain. Incentives

still reward novelty over careful replication, making it less attractive to reproduce others' work (Begley & Ellis, 2012). Small samples and flexible analytic choices continue to generate unstable results (Button et al., 2013). Even when journals require data and code, enforcement is inconsistent (Miguel et al., 2014).

Several concrete steps could help:

Policy and enforcement stricter journal checks for code and data availability.

Preregistration and registered reports limiting bias by committing to methods before results are known (Nosek et al., 2015).

Training and infrastructure promoting reproducible pipelines and collaborative platforms (Sandve et al., 2013).

Environment capture using tools such as renv in R to freeze dependencies and ensure analyses remain reproducible over time (Starr et al., 2015).

4 Conclusion

The reproducibility debate shows both progress and persistent challenges. Psychology, economics, and biomedicine all provide evidence that many results cannot be reproduced (Begley & Ellis, 2012; Collaboration, 2015; Dewald et al., 1986). At the same time, reforms such as preregistration, journal data policies, and computational tools are helping to improve research practices (Munafò et al., 2017; Sandve et al., 2013).

The overall conclusion is that reproducibility is not yet good enough. Technical obstacles have been reduced, but incentives and enforcement still lag behind. A realistic path forward is to require computational reproducibility for all published work, while encouraging replication where the stakes are high. With wider use of Quarto, GitHub, and data archives, science can continue moving toward being more robust and reliable.

5 Software used

5.1 R version

R version 4.5.1 (2025-06-13)

5.2 Attached packages:

attached base packages: [1] stats, graphics, grDevices, utils, datasets, methods, base,

other attached packages: [1] lubridate_1.9.4, forcats_1.0.0, stringr_1.5.1, dplyr_1.1.4, purrr_1.1.0,
[6] readr_2.1.5, tidyr_1.3.1, tibble_3.3.0, ggplot2_3.5.2, —

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). *Rmarkdown: Dynamic documents for r*.

Barnes, N. (2010). Publish your computer code: It is good enough. *Nature*, *467*(7317), 753–753.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., & Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems*, *29*(2), 599–611.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Broman, K. W., & Woo, K. H. (2018). *Data organization in spreadsheets* (No. e3183v2). PeerJ Inc.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, *76*(4), 587–603.

Gentleman, R. (2005). Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L. D., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., & Simonsohn, U. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Sert, N. P. du, Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

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., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227.

Rosenthal, R. (1979). *The file drawer problem and tolerance for null results.*

Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple

rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., Haak, L. L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J. E., Lin, J., Nielsen, L. H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., … Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, *1*, e1.

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data* (pp. XXV, 492). O'Reilly.

Xie, Y. (2015). *Dynamic documents with R and knitr* (Second). Chapman and Hall/CRC.

Xie, Y. (2020). *Knitr: A general-purpose package for dynamic report generation in r* [Manual].