Is reproducibility good enough?

Karin Liang

Tuesday 16 Sep, 2025

Abstract

This paper asks, "Is reproducibility good enough?" It reviews key literature, discusses whether reproducibility should be treated as the norm, and considers tools such as Quarto. The conclusion is that reproducibility is essential but not sufficient; replicability and broader reforms are also required for credible science.

2 Introduction

This paper is based on the lecture "Reproducibility" in the MSB105 Data Science at the Western Norway University of Applied Sciences, which addresses the central question: "Is Reproducibility Good Enough?" We review the literature on reproducibility and replicability, and notes that many studies have reported difficulties in repeating or independently verifying published results. Such evidence suggests that reproducibility alone is insufficient for ensuring reliable knowledge. The paper also considers whether replicability should be treated as a scientific norm, examines the role of tools such as Quarto in supporting reproducible workflows, and briefly points to technical and cultural challenges. Finally, it outlines potential solutions, including dynamic documents, data citation, and institutional reforms to strengthen scientific credibility.

Reliable knowledge is the cornerstone of scientific progress. When findings cannot be verified, the path of research becomes uncertain and future decisions risk being misguided. Both reproducibility and replicability are needed to keep science trustworthy.

3 Literature review

Reproducibility has been described as a core requirement of trustworthy science (McNutt, 2014). Still, as Peng (2011) emphasizes, it is a necessary but not sufficient condition for replicability. The National Science Foundation distinguishes reproducibility, replicability, and generalizability as complementary dimensions of robust science (Bollen et al., 2015; *Dear Colleague Letter*, n.d.), and Goodman et al. (2016) emphasizes methods reproducibility, results reproducibility, and robustness as distinct but interrelated concepts.

Serious obstacles remain. Publication bias leads to the overrepresentation of studies with Type I errors (Rosenthal, 1979). The consequences of Type I errors are

severe, including persistence in the literature, wasted resources, and damage to scientific credibility (Simmons et al., 2011). Meta-analysis, defined as a statistical method for synthesizing results across studies (wikipedia, 2020), is also distorted when publication bias is present (Iyengar & Greenhouse, 1988). In addition, Ioannidis argued that most published research findings may actually be false (Ioannidis, 2005), and in medicine, irreproducible preclinical studies have been shown to waste resources and delay progress (Begley & Ellis, 2012).

Large replication efforts confirm these concerns. In psychology, replication rates are much lower than in the original studies (Klein et al., 2018; Nosek et al., 2015), while in economics, attempts often fail due to missing data, poor documentation, or flawed code (McCullough et al., 2008), and even when data and code are provided, successful replication is rare. Surveys across disciplines further show that most researchers have failed to reproduce another scientist's experiments (Baker, 2016).

Technical solutions have been proposed. Research objects and data citation systems integrate and track research components (Bechhofer et al., 2013; Brase, 2009). Literate programming (Knuth, 1992) and its derivatives such as noweb (Ramsey, n.d.), Sweave (Leisch, 2002), knitr (Xie, 2014, 2015, 2020), and R Markdown (Allaire et al., 2020; Xie et al., 2018) paved the way for Quarto, which supports multi-language workflows and reproducible documents. Yet studies show that even with code and data available, replication can fail (Pimentel et al., 2019; Rule et al., 2018).

Cultural reforms are equally crucial. Guidelines for authors and editors have been proposed to reduce false positives (Simmons et al., 2011), while broader initiatives call for preregistration, data sharing, and openness (Munafò et al., 2017). As noted by Hardwicke et al. (2020), lasting improvements in efficiency and credibility will depend not just on technical solutions but also on broader institutional and cultural change.

4 Discussion of the reseach question

4.1 Should replicability be the norm or is this to much to ask for now?

Replicability should be regarded as a scientific norm rather than an optional goal. While reproducibility is widely recognized as a baseline (Peng, 2011), it cannot ensure reliable findings on its own. Evidence from psychology, medicine, and economics shows that many published results fail independent verification (Ioannidis, 2005; McCullough et al., 2008; Nosek et al., 2015), and replication success remains limited even when code and data are provided (Pimentel et al., 2019).

The implications are serious: false positives and selective reporting waste resources and undermine confidence in research (Simmons et al., 2011). As noted by *Dear Colleague Letter* (n.d.), robust science requires more than isolated reproducibility, and Bollen et al. (2015) stresses that credibility depends on reproducibility, replicability, and generalizability together.

Possible remedies already exist. Proposals include author and editor guidelines to reduce false positives (Simmons et al., 2011), research objects and data citation systems (Bechhofer et al., 2013; Brase, 2009), literate programming traditions (Knuth, 1992), and broader reforms promoted by meta-research and open science initiatives

(Hardwicke et al., 2020; Munafò et al., 2017). These efforts suggest that normalizing replicability is not excessive but a necessary step toward credible science.

4.2 Can Quarto documents help with reproducibility?

Quarto offers a powerful framework to improve reproducibility by integrating code, data, and analysis into one document, reducing problems caused by fragmented research outputs (Bechhofer et al., 2013). It continues the literate programming tradition (Knuth, 1992), evolving from earlier tools like knitr (Xie, 2014, 2015) and R Markdown (Xie et al., 2018; Xie, 2020), which first enabled the seamless combination of code and narrative.

Still, reproducibility cannot be guaranteed simply by sharing code and data. Studies of computational notebooks reveal that many analyses fail to rerun due to missing dependencies or poor documentation (Pimentel et al., 2019). Similar shortcomings have been observed in economics, where archives rarely supported full replication (McCullough et al., 2008). Thus, while Quarto reduces technical barriers, it cannot by itself resolve issues such as publication bias (Rosenthal, 1979) or false positives (Simmons et al., 2011).

Where Quarto shows its greatest strength is in aligning with open science practices. It supports version control, citation management, and multiple output formats (Allaire et al., 2020; Xie et al., 2018), enabling consistent communication of results. More importantly, it complements wider reforms promoted by meta-research (Hardwicke et al., 2020) and open science initiatives advocating transparency, preregistration, and data sharing (Munafò et al., 2017). In this way, Quarto functions not as a complete solution, but as a key step toward more trustworthy science.

4.3 What problems remains and how can these be solved?

Even with progress in reproducibility, significant challenges remain. Structural incentives still prioritize novel, positive results, encouraging publication bias and sustaining false positives, which undermines efficiency and damages trust in science (Rosenthal, 1979; Simmons et al., 2011). Technical hurdles also persist: differences in software versions, dependencies, and computing environments often prevent replication even when data and code are shared (McCullough et al., 2008; Nosek et al., 2015; Pimentel et al., 2019). Moreover, replication work often lacks cultural and institutional support, while limited transparency, preregistration, and open data sharing further discourage such efforts (Baker, 2016; Hardwicke et al., 2020; Munafò et al., 2017).

Solutions require both technical and cultural strategies. On the technical side, tools like Quarto, along with predecessors such as R Markdown and knitr (Allaire et al., 2020; Xie, 2014, 2015, 2020; Xie et al., 2018), help integrate code, results, and text, supporting reproducibility across platforms. Culturally, change must come through author and editor guidelines to limit false positives (Munafò et al., 2017; Simmons et al., 2011), systemic reforms promoting openness (Hardwicke et al., 2020), and adoption of practices such as data citation (Brase, 2009). Meta-research has outlined concrete steps in this direction, including stronger peer review, preregistration, and broader sharing of code and data (Hardwicke et al., 2020).

5 Conclusion

This paper reviewed the distinction between reproducibility and replicability, the contribution of Quarto, and the remaining barriers to reliable science. The evidence shows that reproducibility is necessary but insufficient, as many published research results fail to be independently verified (Ioannidis, 2005; McCullough et al., 2008; Nosek et al., 2015). Replicability and generalizability must also be part of the standard for credible research (Bollen et al., 2015).

Quarto and earlier tools such as knitr and R Markdown provide valuable technical support by combining code, data, and text into dynamic documents (Allaire et al., 2020; Xie, 2014; Xie et al., 2018). These tools reduce fragmentation but cannot alone solve deeper problems like publication bias and false positives (Rosenthal, 1979; Simmons et al., 2011).

Lasting progress depends on both technical and cultural reforms.Dynamic documents, environment management, and data citation systems improve reproducibility, while preregistration, open sharing, and systemic incentives strengthen replicability (Hardwicke et al., 2020; Munafò et al., 2017).

In sum, reproducibility is not enough. Science must normalize replicability and embed openness into its practices to ensure long-term credibility.

5.1 Software and Packages

- R version 4.5.1 (2025-06-13)
- Base packages: stats, graphics, grDevices, utils, datasets, methods, base

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*.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://doi.org/10.1038/533452a

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. https://doi.org/10.1016/j.future.2011.08.004

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

Bollen, K., Cacioppo, J. T., Krosnick, J. A., Olds, J. L., & Kaplan, R. M. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science* (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences). NSF.

Brase, J. (2009). DataCite - A Global Registration Agency for Research Data. *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, 257–261. https://doi.org/10.1109/COINFO.2009.

66

Dear Colleague Letter: Robust and Reliable Research in the Social, Behavioral, and Economic Sciences (Nsf16137) | NSF - National Science Foundation. (n.d.). https://www.nsf.gov/pubs/2016/nsf16137/nsf16137.jsp.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S. N., & Ioannidis, J. P. (2020). Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application*, *7*(1), 11–37.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, *3*(1), 109–117.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Knuth, D. E. (1992). *Literate Programming*. Cambridge University Press.

Leisch, F. (2002). Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis. In W. Härdle & B. Rönz (Eds.), *Compstat* (pp. 575–580). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-57489-4_89

McCullough, B. D., McGeary, K. A., & Harrison, T. D. (2008). Do economics journal archives promote replicable research? *Canadian Journal of Economics/Revue Canadienne d'économique*, *41*(4), 1406–1420. https://doi.org/10.1111/j.1540-5982.2008.00509.x

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229–229. https://doi.org/10.1126/science.1250475

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(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. https://doi.org/10.1126/science.aab2374

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Pimentel, J. F., Murta, L., Braganholo, V., & Freire, J. (2019). A large-scale study about quality and reproducibility of jupyter notebooks. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, 507–517.

Ramsey, N. (n.d.). *Noweb home page*. https://www.cs.tufts.edu/~nr/noweb/.

Rosenthal, R. (1979). *The file drawer problem and tolerance for null results. 86*, 638–641.

Rule, A., Tabard, A., & Hollan, J. D. (2018). Exploration and explanation in computational notebooks. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

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. https://doi.org/10.1177/0956797611417632

wikipedia. (2020). Meta-analysis. *Wikipedia*.

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman and Hall/CRC.

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].

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman and Hall/CRC.