Is reproducibility good enough?

Harald Bjarne Vika

18,09,2025

Abstract

This text explores the ongoing challenges and solutions related to reproducibility in economic research. While early efforts, such as those by Econometrica, emphasized data transparency, increasing model complexity over time made replication difficult. Modern tools such as RStudio, knitr, and Quarto now enable the integration of code, data, and analysis into fully reproducible documents. However, problems remain, including poor compliance with data-sharing policies, lack of incentives, and technical barriers. Addressing these issues requires stronger enforcement, cultural change, and broader access to training and user-friendly tools

1 Introduction

The advancement of science relies heavily on trust in the discovery of new data. Reproducibility plays a key role in building this trust, as it allows scientist to verify and gain confidence in the conclusions drawn form research. However, in recent times, the scientific community has raised concerns about the growing number of peer-reviewed preclinical studies that cannot be reproduced(McNutt, 2014).

In this paper, we will explore the concept of reproducibility in scientific research, why it is important for building trust and whether current practices are sufficient to ensure its reliability for future scientists. We begin by reviewing the literature to clarify the definitions and distinctions between "reproducibility" and " replicability" and their roles in the scientific process. Next, we will discuss whether reproduccibility is essential or if replication alone can be considered adequate. We will also examine hoew tools such as R and Quarto documents can enhance research reproducibility, identify common challenges in this process, and consider possible solutions. Finally, we will conclude by reflection on these issues and sharing our perspective on the state and future of reproducibility in scientific data.

2 Literature review

3 Theory on reproducibility

THe cover story of The Economist, titled "How Science Goes Wrong", adopts the terminology for reproducibility from (Barba, n.d.), defining it as the ability

to regenerate results using the original researcher's data and code. In contrast, the same source describes replicability as either the process of collecting new data to arrive at the same scientific findings as a previous study, or when an independent research team reproduces the results using the original author's materials. These definitions highlight that while reproducibility and replicability are closely related and sometimes overlapping, they refer to distinct aspects of scientific validation.

Another important concept is generalizability, which refers to the extent to which the results of a study can be applied to different contexts or populations beyond the original study setting. Together, reproducibility, replicability and generalizability contribute to what is often described as robust and reliable science, a foundation essential for scientific progress and the ability to build confidently on prior research.

In this paper, we will focus primarily on the role of reproducibility in scientific data and its importance in maintaining the integrity and trustworthiness of research.

Reproducibility is a necessary, but not sufficient, condition for replicability. As noted by Peng (2011), reproducibility should be considered a minimum requirement for scientific publications. As previously discussed, reproducibility helps scientists build confidence and trust in their findings—but what exactly does "research reproducibility" mean?

Goodman et al. (2016) outlines three distinct dimensions of research reproducibility:

- **Methods reproducibility** – This refers to providing sufficient detail about the study's procedures and data so that the same research can be repeated, both theoretically and in practice.
- **Results reproducibility** (also referred to as *replicability*) – This involves obtaining the same results as a previous study when conducting an independent study using the same methodology.
- **Robustness and generalizability** – While sometimes used interchangeably with reproducibility, these terms highlight additional aspects of scientific reliability. *Robustness* refers to the stability of experimental conclusions under variations in baseline assumptions or procedures. *Generalizability*, on the other hand, refers to the extent to which a study's findings remain valid in different settings or populations outside the original experimental framework.

Together, these concepts form a broader understanding of what it means for scientific research to be reliable and trustworthy.

4 Publication bias

When discussing reproducibility in scientific research, one major challenge is *publication bias*, which can significantly affect the ability to publish findings. This bias can lead to a disproportionate representation of positive results in the literature, while studies reporting null or negative results are often left unpublished.

One serious consequence of publication bias is the increased risk of Type I errors, false positives. A *Type I error* occurs when the null hypothesis (H) is incorrectly rejected, meaning we conclude that there is an effect when, in fact, none exists. The significance level (commonly denoted as) represents the probability of making this kind of error and reflects the level of risk the researcher is willing to accept before conducting the hypothesis test (*Type i and Type II Errors*, n.d.).

If scientific journals systematically favor studies that reject the null hypothesis, the literature may become saturated with false positives. This situation undermines the foundation of future research, especially when later studies attempt (and fail) to replicate the original findings. One well-known example of this issue is the *"File Drawer Problem"* described by Rosenthal (1979), where studies that fail to reject the null hypothesis are effectively "filed away" and never published, while only statistically significant results make it to print. This selective publication practice skews the overall body of evidence, making false positives more likely.

The implications are far-reaching. If a prestigious journal publishes a study with a false positive, it may discourage replication efforts due to the perceived credibility of the source. Additionally, investing resources into research programs based on such flawed findings can be costly and may lead to ineffective or even harmful policy decisions.

Fields that regularly publish false positives risk losing credibility and the trust of the scientific community (Simmons et al., 2011). As Young et al. (2008) notes, *"More alarming is the general paucity in the literature of negative data. In some fields, almost all published studies show formally significant results, so that statistical significance no longer appears discriminating."* This lack of negative results contributes further to the distortion of the scientific record.

Publication bias also compromises meta-analyses, which aggregate data from multiple studies to draw broader conclusions (Russo, 2007). If the individual studies included are themselves biased toward positive findings, the meta-analytic results will also be skewed, compounding the problem.

In short, addressing publication bias is essential to preserving the reliability, replicability, and credibility of scientific research.

When it comes to economics and the issue of research reproducibility, historical efforts can be traced back to journals like *Econometrica*. In an editorial note, Frisch (1933) emphasized the importance of transparency, stating: *"Statistical and other numerical work presented in Econometrica, the original raw data will, as a rule, be published, unless their volume is excessive"*. This reflects an early recognition that making data publicly available is essential for reproducibility.

However, over time, as economic research grew in complexity and scale, models became larger and more computationally intensive. As a result, researchers began publishing only final results, while the raw data and underlying code were often omitted. This shift made it increasingly difficult if not impossible for others to replicate or reproduce published findings.

To address this problem, various solutions were proposed. One such initiative was the creation of data archives, such as the American Economic Association

(AEA) data repository (*American Economic Association*, n.d.), aimed at preserving and sharing datasets used in published research. However, these archives often fell short of their goal. As McCullough et al. (2008) points out: *"All the long-standing archives at economics journals do not facilitate the reproduction of published results. The data-only archives at the Journal of Business and Economic Statistics and the Economic Journal fail in part because most authors do not contribute data."* In other words, the mere existence of archives was insufficient—author compliance was inconsistent, and critical components such as code and documentation were often missing.

A more robust solution emerged in the form of computable documents research articles that integrate text, data, and executable code in a single, reproducible format. This approach allows other researchers to fully reproduce the original study, including data preparation, model estimation, and results generation.

One example of this is WaveLab, a software package designed to reproduce figures from published wavelet research. Developed with reproducibility in mind, WaveLab includes all the source code required to replicate results [*WaveLab and Reproducible Research | SpringerLink* (n.d.);]. The authors encouraged readers to inspect and use the code, promoting what they called "really reproducible" research.

Gentleman (2007) reinforces this idea by emphasizing the role of software in enabling reproducibility: *"The step from disseminating analyses via a compendium to reproducible research is a small one. By reproducible research, we mean research papers with accompanying software tools that allow the reader to directly reproduce the results and employ the methods that are presented in the research paper"* (Gentleman & Temple Lang, 2007).

In this model, the article itself includes:

- The research text
- Code to import and clean the data,
- Code for model estimation and testing,
- And code to generate the final results.

By submitting such materials alongside the manuscript, the research becomes fully transparent and reproducible. This not only strengthens the credibility of the work but also enables future researchers to build on it more effectively.

To implement effective data archiving and reproducible research, tools such as Sweave were originally developed to integrate code and documentation. Sweave, the predecessor of knitr, allowed users to embed R code within LaTeX documents to dynamically generate reports (*Sweave*, n.d.).

Today, knitr is the more widely used and advanced tool, fully supported within RStudio. RStudio itself is an evolution of earlier tools like the R Notebook and offers an integrated development environment (IDE) that allows users to run code in sequentialchunks, while also enabling individual execution for experimentation and debugging (Lander, 2017).

RStudio supports the full data science workflow: it helps structure and clean data, perform transformations, visualize results, and build statistical or machine

learning models. One powerful feature is the use of Quarto documents, which allow users to combine narrative text with code in a single file. These documents can be rendered into multiple output formats—including HTML, PDF, Word, and presentation slides—making it easy to share fully reproducible research reports.

By using these tools, researchers can ensure that their analyses are transparent, replicable, and accessible to others.

5 Discussion of the research question

Replicability *should* be the norm in scientific research, and while it presents challenges, expecting it is not too much to ask—it is, in fact, essential for credible and trustworthy science.

Historically, even early journals like *Econometrica* emphasized the importance of transparency. As Frisch (1933) noted, raw data should be published alongside statistical work unless prohibitively large (Frisch, 1933). This shows that the value of replicability was acknowledged from the beginning. However, as research became more complex and models more data-intensive, researchers increasingly published only results, making it nearly impossible for others to replicate their work.

Solutions like the American Economic Association's data archive were a step forward, but they often failed in practice because many authors did not submit their data (McCullough et al., 2008). This highlights that infrastructure alone isn't enough stronger cultural and institutional expectations are needed.

More promising are tools that support computable documents, such as knitr and Quarto in RStudio, which allow researchers to embed code, data, and analysis directly within documents. These tools make replicability not only possible but practical. They allow for seamless integration of text, code, and output, which can be rendered into shareable formats like HTML, PDF, or slides. With these modern tools, the barriers to replicable research are significantly lower than they once were.

In short, while replicability may have seemed burdensome in the past, today's tools have made it far more achievable. As the cost of irreproducible research grows—damaging trust, wasting resources, and leading to false conclusions—it becomes increasingly clear that replicability is not an unreasonable demand. Instead, it should be a standard practice that protects the integrity and progress of science.

Yes, Quarto documents can significantly help with reproducibility by allowing researchers to combine code, data, analysis, and narrative in a single, integrated document.

As discussed, tools like Sweave and later knitr paved the way for embedding code into research documents, enabling dynamic report generation (*Sweave*, n.d.). Today, RStudio, building on these foundations, supports Quarto, a modern tool designed specifically for reproducible research workflows. Quarto allows researchers to write documents that include both the text of their analysis and

the underlying code that generates the results.

What makes Quarto especially powerful is its flexibility: researchers can use it to generate outputs in various formats—including HTML, PDF, Word, and slides all from the same source file. This means a single Quarto document can serve both as a scientific report and as a reproducibility package. Code chunks can be run individually for testing or sequentially for full analysis, and because the document includes both the data-processing steps and the models, readers can see exactly how the results were generated.

By using Quarto, researchers reduce ambiguity, eliminate manual errors, and make it easier for others to validate, understand, and build upon their work. It supports the broader goal of "really reproducible research," as described in efforts like WaveLab and in the principles set out by Gentleman & Temple Lang (2007), who emphasized that reproducible research means papers accompanied by the actual tools needed to regenerate results.

In short, Quarto is a practical and powerful solution for making reproducibility not just possible, but efficient and well-integrated into the research process.

Despite major advancements in tools and workflows that support reproducibility, such as Quarto, knitr, and data archives, several key problems still remain.

One major issue is incomplete data and code sharing. While journals and organizations like the American Economic Association have created data archives, these often fall short because many authors do not contribute their data or fail to provide sufficient documentation or code (McCullough et al., 2008). Without access to the exact datasets and the code used in analysis, reproducibility becomes impossible, regardless of available tools.

Another persistent problem is cultural and institutional inertia. Although tools like RStudio and Quarto make reproducibility easier than ever, many researchers still do not adopt these practices. There may be a lack of incentives, a fear of being scooped, or simply insufficient training. As a result, many studies continue to be published without the necessary materials for replication, and efforts like computable documents remain underutilized.

Additionally, technical barriers still exist for some users. While tools like Quarto are powerful, they require familiarity with R, LaTeX, markdown, and version control systems, all of which can be intimidating to researchers without a computational background.

6 Conclusion

So how can we look at these problems?

- Enforce data and code availability policies: Journals and funding bodies should require that authors submit both datasets and code used for analysis as a condition for publication. This should be monitored and enforced, not just encouraged.
- Promote and normalize the use of reproducible tools: More training and education should be offered to help researchers adopt tools like RStudio

and Quarto. Universities, research institutions, and conferences should prioritize workshops and resources on reproducible workflows.
- Standardize reproducible workflows: Encouraging the use of "computable documents" as a standard format—where all text, code, and data are embedded—can simplify expectations and promote consistency across disciplines.
- Recognize and reward reproducible research: Researchers who go the extra mile to ensure their work is fully transparent and replicable should be recognized through citations, awards, or funding opportunities.
- Lower technical barriers: Continued development of user-friendly tools, interfaces, and templates can make reproducible research more accessible to non-technical users.

In conclusion, while powerful solutions now exist to make reproducibility achievable, real change requires a cultural shift one that embraces transparency, provides proper incentives, and equips researchers with both the tools and the training to adopt reproducible practices across the board

7 References

R version 4.5.1 (2025-06-13 ucrt) Platform: x86_64-w64-mingw32/x64 Running under: Windows 11 x64 (build 26100) Matrix products: default LAPACK version 3.12.1 locale: [1] C time zone: Europe/Oslo tzcode source: internal attached base packages: [1] stats graphics grDevices utils datasets methods base loaded via a namespace (and not attached): [1] compiler_4.5.1 cli_3.6.5 rsconnect_1.5.1 tools_4.5.1 rstudioapi_0.17.1 lifecycle_1.0.4 [7] rlang_1.1.6

American Economic Association. (n.d.). https://www.aeaweb.org/journals/data.

Barba, L. A. (n.d.). *Terminologies for reproducible research*. https://doi.org/10.48550/arXiv.1802.03311

Frisch, R. (1933). Editor's note. *Econometrica*, *1*(1), 1–4. https://www.jstor.org/stable/1912224

Gentleman, R., & Temple Lang, D. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, *16*(1), 1–23.

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

Lander, J. P. (2017). *R for everyone: Advanced analytics and graphics* (Second edition). Addison-Wesley, Pearson.

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.

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229–229.

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.*

Russo, M. W. (2007). How to review a meta-analysis. *Gastroenterology & Hepatology*, *3*(8), 637–642. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3099299/

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.

Sweave: Dynamic generation of statistical reports using literate data analysis | SpringerLink. (n.d.). https://link.springer.com/chapter/10.1007/978-3-642-57489-4_89

Type i and type II errors: What are they and why do they matter? - c j smith, 2012. (n.d.). https://journals.sagepub.com/doi/10.1258/phleb.2012.012j04

WaveLab and reproducible research | SpringerLink. (n.d.). https://link.springer.com/chapter/10.1007/978-1-4612-2544-7_5

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why Current Publication Practices May Distort Science. *PLoS Medicine*, *5*(10), e201.