SUERF
The European Money and Finance Forum

SUERF Policy Brief
No 288, March 2022

Economists: not enough transparency? Evidence from a reproducibility exercise of a top economic journal




By Sylvérie Herbert, Hautahi Kingi, Flavio Stanchi and Lars Vilhuber*

Keywords: Replication, reproducibility, transparency, economics research, journal policies.

Policy-making relies increasingly on empirical economics research. Thus, for these policies to be credible, the research underpinning them should be transparent, reproducible, and replicable. We test whether peer-reviewed journals, often times used to measure the quality of economists' research, can be guarantors of reproducibility, through their requirements of providing codes and data. Focusing on a high profile journal with a data availability policy, American Economic Journal: Applied Economics, we find that only between 25% and 43% of articles are reproducible. Fully reproducible studies are not cited more often. Our study suggests that systematic journal-based reproducibility checks may bring the economist profession to a better equilibrium of transparent and reproducible research.

* **Sylvérie Herbert**, Banque de France; **Hautahi Kingi**, Google; **Flavio Stanchi**, Airbnb; **Lars Vilhuber**, Cornell University.

As in many other scientific disciplines, transparency and openness are essential to the credibility of economics research. Transparency is especially critical when that economic research informs economic policy decisions. Indeed, in recent years, governments and policy institutions, and in particular central banks, have pushed for "evidence-based" policy-making. Central banks' (or any other institutions) forecasts, indicators or models for policy evaluation are all based, to some extent, on research. This research and expertise helps better design policies, but this means that economic research needs to trustworthy for these policies to be credible. To be trustworthy, research needs to be transparent, and a key way to ensure that the research is truly fully transparent is to test for *reproducibility*.[1]

The peer review process in economics journals ensures the originality and quality of the research article, and provides a stamp of approval. However, editors and referees do not have the obligation to check that the same data and code yields similar results as the authors. Hence, the referee process does not provide a check on the reproducibility. In an effort to foster greater reproducibility, journals put in place data availability policies (DAP henceforth) as early as 1933, with *Econometrica*, asking authors to provide data and codes. Other journals followed quite early as well, such as the *Journal of Money, Credit and Banking* in the late 1990's. First amongst the discipline's "top 5" journals, the *American Economic Review* introduced a DAP in 2005, with the *Quarterly Journal of Economics* following in 2016. However, as of 2017, only 54% of 343 economic journals of the Thomson Reuters Social Science citation index had a DAP (Höffler 2017).

In Herbert et al. (2021), we test whether journals' data availability policies with light enforcement, such as the one enforced by the *AEA* in 2005, yields reproducible research in the first place. While there is no systematic check of the codes with such policies, making codes and data publicly available should in theory enhance transparency. Providing undergraduate students only with the code, data, and information provided by the authors, our study tests if they are able to successfully reproduce the results in the published articles. Such a reproducibility check verifies whether the provided materials are accurate and complete, thus verifying that they are fully transparent.

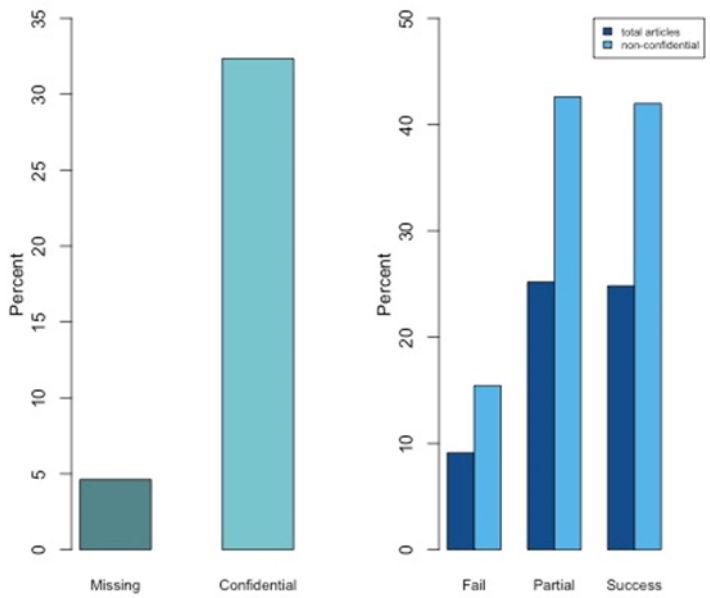
Reproducing the AEJ:Applied. Our analysis is based on articles published in the *AEJ:Applied* between 2009 and 2018, which we gathered along with their codes and data from the journal's website. Each replicator (or assessor) first evaluated the reproduction task or "expected" level of reproducibility by filling out a questionnaire, gathering information on the availability of the required components (data, code, documentation). The questionnaire recorded a selection of article characteristics, including the type of data used (e.g., restricted, confidential, public...), the presence of an online appendix, the programs used by the author, the documentation on how to run them (readme files). They also gathered information on the completeness and clarity of the documentation of the programs and data. Most articles had a supplementary dataset, but not all articles were accompanied by the full data set necessary for replication. The assessor recorded whether the articles were accompanied by the necessary data, and any apparent reason if not. Having gathered all this preliminary information, the assessor gave an estimate of the "expected" difficulty of reproduction. They then attempted to reproduce the analysis – trying to reproduce the empirical analysis from the provided code and data, with minimal changes and no interaction with the original authors. Throughout the exercise, the assessor was asked to keep track of any changes made to the code to run the analysis. Whenever substantial changes were needed for

[1] We follow definitions of "replicability" and "reproducibility" by the US National Academy of Sciences. They define "reproducibility" as "obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis" (National Academies of Sciences, Engineering, and Medicine, 2019, pg. 36). Replicability is achieved for instance by collecting new data, implementing different methods or code, and then "obtaining consistent results across studies aimed at answering the same scientific question".

the code to run, they were told to limit the time allowed for this article's reproduction. Once the reproduction completed, assessors answered an exit-questionnaire to capture information about the success or failure of replication, and the reasons behind the failure.

A moderate success, mainly due to non-provided data. We find only moderate success in reproducing the articles in spite of a data availability policy. When considering the sample of assessed articles, only 25% were successfully reproduced in full. Conditional on selecting papers using non-confidential data, we find a higher reproducibility rate of 43%. Another 43% of these articles were partially replicated, meaning that the most important tables were similar in the paper and the replication. Such low reproducibility is thus driven by confidential or missing data (see figure 1), which constitute more than a third of our sample. However, even when restricting on non-confidential data, our analysis shows that a substantial numbers of articles had different results than those obtained from their replication packages (14%), which is non negligible.

Figure 1: Proportion of articles with confidential and missing data (left) and reproducibility ratio (right)



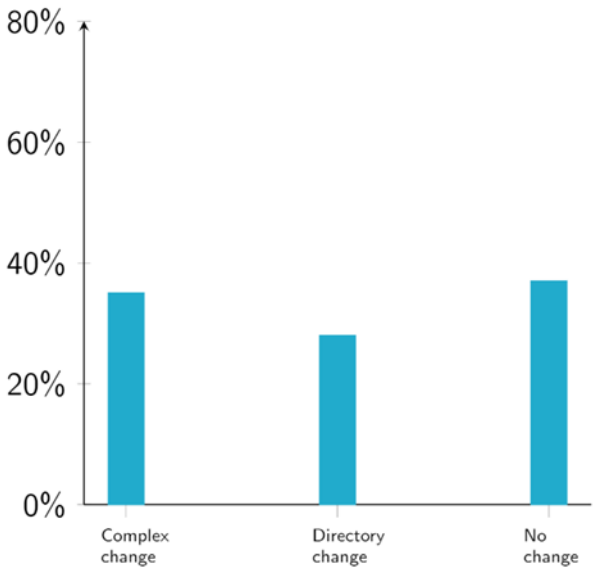
Note: Ratio of articles with missing and confidential data to the left, over the sample of our articles. Failure, partial and full reproducibility ratio according to different definitions of the sample (to the right).

A small fraction of failures stemmed from corrupted data, code errors, and unavailable software (24% total with equal proportion), while the largest fraction of failures stemmed from inconsistent values (64%). The computer programs successfully ran, but the numerical values were inconsistent with those reported in the articles, and the replicators were unable to find a convincing reason.

Reproducible research is not necessarily easily operated. This non-negligible fraction of failures highlights the need for a systematic reproducibility check so as not to threaten the credibility of the expertise developed in research articles. Lightly monitored data availability policies by themselves thus seem necessary but not sufficient to ensure full reproducibility and easily accessible research. Moreover, we found that even

reproducible articles required complex code changes to reach similar results as in the papers (figure 2). A third of articles required involved substantial changes, another third minimal changes, and only a minority required no change. This means this research cannot easily be taken off the shelf to inform policy discussions.

Figure 2: Proportion of replicated articles requiring no change, small or complex changes



No reproducibility bonuses. Most importantly, we further show that reproduced papers have not earned a citation bonus. Increased citations may thus not be enough of an incentive device to generate transparent documentation and coding practices. Our analysis therefore calls for a systematic check of articles during the referee process, as a way to reach a "good reproducibility" equilibrium.

References

Herbert, S., Kingi, H., Stanchi, F. and Vilhuber, L. 2021. "The Reproducibility of Economics Research: a Case Study", Working papers 853, Banque de France.

Höffler, Jan H. 2017. "Replication and Economics Journal Policies." *American Economic Review*, 107 (5): 52-55.

About the authors

Sylvérie Herbert *is a research economist in the Monetary Policy Division at the Banque de France. Prior to joining the Banque de France, and throughout her PhD, she gained experience at the Federal Reserve Bank of St Louis and the Federal Reserve Bank of Richmond. Previously, she interned in the Monetary Policy Research Division of the ECB's Directorate General Research and the Monetary Policy Strategy Division of the Directorate General Economics. Her research covers central bank communication, monetary policy, expectations formation, and more broadly information economics (dispersed information). She holds a PhD from Cornell University.*

Hautahi Kingi *is an economist and Data Scientist at Google. Prior to joining Google, he worked as a Data Scientist at Facebook, and as a Senior Research Economist at IMPAQ International, in Washington DC. He worked for the International Monetary Fund as a Research Associate. He holds a PhD from Cornell University.*

Flavio Stanchi *is a Data Scientist at Airbnb. He holds a PhD in economics from Cornell, where he researched the impact of innovation and technology diffusion on markets.*

Lars Vilhuber *is the Executive Director of the Labor Dynamics Institute at Cornell University, and a Senior Research Associate in the Economics Department of Cornell University. He works with the Research and Methodology Directorate at the U.S. Census Bureau on a variety of projects. Since 2018, he is the American Economic Association's Data Editor, and Co-Chair of the Innovations in Data and Experiments for Action (IDEA) Initiative at J-PAL. He is the Managing executive Editor of the Journal of Privacy and Confidentiality.*

SUERF Publications

Find more **SUERF Policy Briefs** and **Policy Notes** at www.suerf.org/policynotes

SUERF
The European Money and Finance Forum

SUERF is a network association of central bankers and regulators, academics, and practitioners in the financial sector. The focus of the association is on the analysis, discussion and understanding of financial markets and institutions, the monetary economy, the conduct of regulation, supervision and monetary policy.

SUERF's events and publications provide a unique European network for the analysis and discussion of these and related issues.

SUERF Policy Briefs (SPBs) serve to promote SUERF Members' economic views and research findings as well as economic policy-oriented analyses. They address topical issues and propose solutions to current economic and financial challenges. SPBs serve to increase the international visibility of SUERF Members' analyses and research.

The views expressed are those of the author(s) and not necessarily those of the institution(s) the author(s) is/are affiliated with.

SUERF Secretariat
c/o OeNB
Otto-Wagner-Platz 3
A-1090 Vienna, Austria
Phone: +43-1-40420-7206
www.suerf.org • suerf@oenb.at