

# Comments on Reproducibility in Finance and Economics

Toni M. Whited

November 2022

## 1. Introduction

Reproducibility is defined as obtaining consistent results using the same data and code as the original study. Most of the discussion of reproducibility has centered around the many obvious benefits. Reproducible research advances knowledge for several reasons. It reduces the risk of errors. It also makes the processes that generate results more transparent. This second advantage has an important educational component, as it helps disseminate not just results but processes. However, reproducibility is not without costs. Good research procedures consume resources both in terms of a researcher's own efforts and in terms of the involvement of arms-length parties in actually reproducing the research. This second cost is not just a time cost; it is pecuniary as well.

Thus, reproducibility is a good that is costly to produce and that has many positive externalities. Researchers internalize many of the benefits of reproducibility, especially in terms of research extendability and personal reputation. However, they do not internalize any of the benefits to the research community at large. Because reproducibility is costly, it is unlikely to be produced at a socially optimal rate by any individual researchers. Thus, the questions are the extent to which reproducibility should be subsidized and who should subsidize it. Should all research be reproduced by arms-length

parties, and what are the least costly policies that facilitate reproducible research? The rest of this note is organized around policies regarding actual reproduction and proprietary data.

## **2. Code, Data, and Arms-Length Reproduction**

One low-cost and easily implementable set of policies that enhances the reproducibility of research is journals' data and code disclosure policies. In the age of inexpensive data storage and an abundance of public repositories, the costs of these policies are small, and the policies should be implemented. They impose some costs on researchers in terms of organizing data and code, but well-organized data and code are already an essential part of the research process, so these costs should be small.

While simple to implement, this low-cost policy is not without non-pecuniary drawbacks for journals. The code and data can be incomplete, poorly documented, or unusable. Moreover, journal editors have to retract articles that, after publication, cannot be reproduced. In economics, these concerns have prompted journals to start arms-length reproduction of results. The benefit of this policy is primarily that authors and journals can be confident that the code submitted with an article actually works to reproduce the results.

However, the pecuniary costs of this policy can be substantial. It is expensive for journals to hire data editors and well-trained research assistants, and many academic journals run on tight budgets. It is often time-consuming for authors to comply with reproducibility requirements. This last issue is particularly burdensome for authors who cannot afford research assistance.

While the above issues involve costs, the following are more fundamental. Reproducibility policies give researchers incentives to do research that is easier to reproduce, thus restraining research innovation that requires either large data or intense comput-

ing. Most importantly, code that can run on data and reproduce results can still contain errors.

These arguments imply that while individual researchers are likely to underproduce reproducibility, it is also unlikely optimal for the progress of science that all research be reproduced before publication. Some papers, even those in the very best journals, rarely get read or cited, and the benefits of reproducing these papers are small.

However, ex-ante, it is hard to know which papers will attract attention and which will not. One solution that lies between data and code disclosure and arms-length reproduction is verification. It is much less expensive to verify the contents of a replication package than to do an actual reproduction. Verification might consist of checking for the existence of replication instructions, an execution script, or either data or pseudo-data. This type of service could be provided by journals or other third parties, much as copy editors fix syntax and grammar errors before articles are submitted. At that point, reproducibility would be left up to the academic community, with the more important pieces of research being subject to greater scrutiny.

A final issue with reproducibility is education. In economics and finance, students are not taught how to create reproducible research. An improvement that would go a long way toward improving the culture surrounding reproducibility would be to teach PhD students how to organize research projects and to write code in such a way that others can reproduce results easily. This type of education would lower the costs to individual researchers of making their own research reproducible.

### **3. Proprietary Data**

A possibly larger challenge for reproducibility than verification or arms-length execution of code is proprietary data. A clarification is necessary because not all types of data with restricted access are completely secret, that is, available only to the data provider

and a researcher. For example, commercial data sets are not secret, just costly to obtain. Similarly, administrative datasets are not secret. They just require special permission. In contrast, proprietary data cannot be offered to the research community at large for the purposes of reproducing the results. So the question is whether journals should discourage the use of this type of data or require that verifiers have access to the data. Given the large number of studies using proprietary data, this issue is possibly more important than the issue of running code.

#### **4. Conclusion**

In conclusion, the reproducibility of research is essential for the advancement of science. However, it is not without costs, so blanket statements that all research should be reproducible are not feasible. Instead, feasible policies include those that lower the costs for others to replicate research. Data and code disclosure is a low-cost policy that should be implemented widely. Verification of code and data packages is a slightly more costly option. Arms-length reproduction is a much more costly alternative. Finally, perhaps the most important issue that impedes reproducibility in finance and economics is the use of proprietary data.