Transparency, Clarity, Replicability, Reproducibility

Christian Schluter

Replication and Reproducibility

The ability to **replicate** a research finding is central to the scientific paradigm, and helps convince us that a particular result is "real."

In the physical sciences, other labs should be able to exactly reproduce what the original authors did and generate the same results

There is a widespread perception that much work in economics is not reproducible, which has contributed to what has become known as a **credibility crisis**.

Practices are (slowly) changing to help put economic research on a more scientific basis.

Examples of evidence on (un)reliability.

One of the most important recent studies is Camerer et al. (2016), which repeated 18 behavioral economics lab experiments originally published between 2011 and 2014 in the AER and the QJE to assess their replicability.

The estimated effects were statistically significant with the same sign in 11 of the 18 replication studies (61.1%).

Chang and Li (2015) systematically tested the reproducibility of 67 macroeconomics papers

Thirty five articles are published in journals with data and code sharing requirements, but Chang and Li could obtain data for only 28 of these (80%) from the journal archives

Of the 26 papers in journals without data sharing requirements, Chang and Li were unable to obtain 15 datasets (58%).

The overall replication success rate is 29 of 67 (43%) overall, or 29 of 61 (48%) among those using non-proprietary datasets, so roughly half.

Pure and statistical replication

But what exactly is replication?

Perhaps surprisingly, multiple definitions are used by different scholars, across fields and over time

Hamermesh (2007) proposes a distinction between:

- **Pure replication**: an exercise to verify that the same results are obtained if one uses the same data and same methods can discover errors in the original analysis.
- **Statistical replication**: an exercise using alternative methods and/or data to test the same hypothesis. Perhaps a better label is **reproducibility** or re-analysis.

Example of high profile pure replication controversy:

Leimer and Lesnoy (1982) found a coding error in the famous Feldstein (1974) Journal of Political Economy paper claiming that Social Security expansion had reduced private savings by 50% (!), with potentially large adverse consequences for U.S. economic growth

In a pure replication exercise that corrected the error, the original result was over-turned.

Example of a high profile reproducibility controversy:

Albouy (2012) disputes the construction of historical data used in the famous Acemoglu, Johnson and Robinson (2001, AER) using historical settler mortality as an instrumental variable (IV) for rule of law, which concludes that institutions were the key determinant of comparative economic growth outcomes over hundreds of years.

Using Albouys modified data (which he claims corrects multiple errors), the IV first stage is weaker, implying that the method is not longer appropriate.

These two original papers were extremely influential, with potentially important implications for social science and public policy.

Another example is Easterly, Levine, and Roodman (2004), commenting on a high profile paper by Burnside and Dollar (2000, AER).

The comment extended the earlier data set to some additional countries and a few additional years, so that it falls somewhere between pure replication and scientific replication.

Its main result was to demonstrate that the original finding that the amount of foreign aid a developing nation receives interacts with good macroeconomic policy to induce growth but does nothing absent such policy did not seem to be robust to the addition of relatively few data points.

But replication work remains uncommon in the social sciences: the data from the median empirical paper published in a field journal is not shared at all (within a few years of publication), and even for articles in leading journals, the data is rarely accessed even 6-7 years later (and sometimes even then probably used for other purposes, such as graduate teaching).

Hamermesh (2007) proposes a number of explanations (e.g. incentives, institutionalising replications), as well as possible remedies going forward (change in social norms and practices).

Current practices have improved considerably, but there are still problems.

There is a widespread view that the quality of posted materials on the AER/AEJ sites is often quite low: the materials are never carefully checked by journal staff, and are often unusable (i.e., variable labels not included). Hence compliance is not what it ought to be.

Many other leading journals in Economics, and even more so in other social science fields, do not have a similar data and code posting requirement.

Hammermesh also links reproducibility to a perspective on the limitations of empirical research (which is often forgotten by authors seeking to advertise the perceived importance of their results)"

"By far the most important justification for scientific replication in non- experimental studies is that one cannot expect econometric results produced for one time period or for one economy to carry over to another. Temporal change in econometric structure may alter the size and even the sign of the effects being estimated, so that the hypotheses we are testing might fail to be refuted with data from another time. This alteration might occur because institutions change, because incentives that are not accounted for in the model change and are not separable from the behaviour on which the model focuses, or, crucially, that even without these changes the behaviour is dependent on random shocks specific to the period over which an economy is observed."

Claerbouts Principle

Koenker and Zeileis (2009): "Claerbouts Principle:

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

We view this as a desirable objective for econometric research."

Literate programming

The technology to (purely) replicate fully and easily now exists in the form of notebooks (eg Jupyter for R or Stata, Rstudio for R).

The process of generating the output reported in the paper can now be fully automated.

All this helps to make your research more **credible**.

Hammermesh (2007): "our ideas are unlikely to be taken seriously if our empirical research is not credible, so that the **likelihood of positive payoffs to our research is enhanced** if we maintain our data and records and ensure the possibility of replication."