

The robustness of doubt

Why replication games are a public good

As a graduate student of economics, one of the best ways to learn your trade is to download the replication package (files containing the information needed to allow others to reproduce the work) of a paper you admire and try to make the numbers come out. You see how a research question became a research design, and how a causal claim became a defensible comparison. You begin to understand the judgment calls that every paper contains.

This replication work matters beyond the seminar room – economic research is used to set interest rates and to allocate public funds. Knowing the strength of the research foundations of such decisions is a public good, not a private reckoning. The spread of mandatory replication packages across journals has now made it possible to check published work at scale – and researchers have started doing it together rather than alone.

Reproducible, robust and replicable

There are three terms worth keeping straight. A result is computationally reproducible if the original code on the original data returns the published numbers; robust if it survives reasonable alternative choices on that data; and replicable if it holds when new data are used. Most of what we call “the replication crisis” is about robustness, and it is the hardest to see from outside, because the sturdy result and the fragile one look identical on the page.

In 2023, I joined a replication games event in Norwich, run by the Institute for Replication, to pressure-test a published paper's data and code. I was paired with two colleagues I had not collaborated with before. The organisers gave us a list of articles from the top economic journals, and the three of us settled on one whose central finding we all found striking: that in response to power shortages an energy-saving campaign, by discouraging use of air conditioning, had raised mortality on the hottest days. It matched our interests and we felt equipped to interrogate it.

The code ran and the numbers matched: the result was computationally reproducible. But robustness checks are redesigns. We asked whether the finding described the average person or the average region, and how much rode on that distinction. We tested alternative assumptions to see which comparison was doing the causal work. And where a key variable had been constructed from search data, we rebuilt it from the ground up. What we found was a distribution: reassuringly, the claim held under most plausible designs.

Throughout, this was a collaborative process: our team worked together to draft the report

and, in the end, heard back from the original authors, who confirmed the results and thanked us for the care we had taken. None of us had a stake in whether the paper survived, and that professional disinterest is exactly what an author cannot supply about their own work. Checking a result is a contribution to the profession, not a case against its author.

Our exercise was one of 110 reproductions across 12 leading journals that contributed to a study published in *Nature* (Brodeur et al, 2026). The results were reassuring but showed why the profession cannot be complacent: more than 85% of claims were computationally reproducible; and 72% of significant results stayed significant under alternative specifications.

In contrast, another study, which sampled 600 social and behavioural science papers at random, found that only a quarter had shared data (Miske et al, 2026). In this case where data and code were open, the findings mostly held (74% were approximately reproducible). The larger danger lies in just how much research remains closed to scrutiny.

Building trust

Earlier this year, I went back for a stricter version: a King's College London event where reproduction itself had turned from a craft into an object of study, with artificial intelligence as the treatment. Does a language model help you to reproduce a paper faster, to catch more of its errors and to judge its robustness as well as a human would? Those are exactly the questions the field should be asking about its own tools before leaning on them.

Verification has long been thankless and lonely, seldom published and occasionally career-poisoning. The replication games turn it into something else: a professional service, closer in spirit to being asked to review a paper. On the day, you look around the room and see people at every career stage, a first-year PhD student next to a full professor, all doing the same unglamorous but essential work.

The credibility of published research is not something that we inherit; it is something that we build, measure and repair. For me, replication began as the small pleasure of poking around in someone else's data. The replication games left me with a clearer sense of what it takes to trust a number, and to deserve the trust that the public places in economics.

Further reading

Brodeur, A. et al. (2026) “Reproducibility and robustness of economics and political science research,” *Nature* 652, 151–156.

Miske, O. et al. (2026), “Investigating the reproducibility of the social and behavioural sciences,” *Nature* 652, 126–134.

