ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Brodeur, Abel; Cook, Nikolai; Mikola, Derek; Totarelli, Maddalena

Research Report

Evidence under scrutiny: How reliable is the research backing public policy?

EconPol Policy Brief, No. 85

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Brodeur, Abel; Cook, Nikolai; Mikola, Derek; Totarelli, Maddalena (2026) : Evidence under scrutiny: How reliable is the research backing public policy?, EconPol Policy Brief, No. 85, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/343151

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

ECONPOL
Policy Brief

06 | 26
85
Volume 10

Evidence Under Scrutiny: How Reliable Is the Research Backing Public Policy?

Abel Brodeur, Nikolai Cook, Derek Mikola, Maddalena Totarelli, et al.

- 85 percent of published economics and political science findings can be independently verified
- One in four significant results disappear under different reasonable choices, and one in four studies contain undetected coding errors
- Published results cluster just above significance thresholds, which is evidence of publication bias
- Policymakers should rely on independently reproduced evidence and avoid scaling up programs based on a single study
- Journals and research entities should expand open data and robustness disclosure mandates

CES ifo

ECONPOL

EconPol Europe is CESifo's economic policy platform. With key support from the ifo Institute, it seeks to leverage CESifo's globe-spanning network of more than 2,000 high-ranked economists – 16 of whom have won the Nobel Prize – and ifo's decades-deep research expertise to provide well-founded advice to European policymakers and to facilitate informed decisions. Drawing on the wide range of specializations of its members, EconPol's mission is to contribute to the crafting of evidence-based, effective economic policy in the face of the rapidly evolving challenges faced by the European economies and their global partners.

EconPol Policy Brief

Publisher and distributor: CESifo GmbH
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 89 9224-0, Email office@cesifo.de
EconPol Europe: www.econpol.eu

Editor: Andrea Balestracci

Editorial Board: Clemens Fuest, Oliver Falck, Lisandra Flach, Sarah Necker, Karen Pittel, Niklas Potrafke, Panu Poutvaara, Andreas Peichl, Helmut Rainer, Marcel Thum, and Ludger Woessmann

Reproduction permitted only if source is stated and copy is sent to EconPol.

Evidence Under Scrutiny: How Reliable Is the Research Backing Public Policy?

Abel Brodeur
Nikolai Cook
Derek Mikola
Maddalena Totarelli*

When governments design policies – whether it's raising the minimum wage, tightening environmental rules, or reforming welfare programs – they often rely on academic research to justify their decisions. The credibility of that evidence depends on two things: whether independent researchers can verify the results using the same data, and whether those results hold up when small but reasonable changes are made to the analysis.

These are not hypothetical concerns. A growing body of evidence across science suggests that many published findings are fragile – either because data and code are unavailable for others to check, or because results only hold under a specific set of assumptions. In economics and political science, where research directly informs high-stakes policy decisions, this matters enormously.

This brief summarizes findings from one of the largest verification exercises ever conducted in these fields. Led by the Institute for Replication (I4R), the project brought together hundreds of researchers worldwide to independently check 110 recently published studies from 12 of the most prestigious journals in economics and political science. All 12 journals require authors to share their data and code publicly – making this an optimistic upper bound. The results show that much is solid, but with important weaknesses that demand action.

* Brodeur (abrodeur@uottawa.ca), Mikola (derek.mikola@gmail.com): University of Ottawa; Nikolai Cook (ncook@wlu.ca): Wilfrid Laurier University; Totarelli (totarelli@ifo.de): ifo Institute. For an exhaustive list of the study's contributors, see the article on which this policy brief is based: Brodeur et al. (2026), "Reproducibility and Robustness of Economics and Political Science Research."

Why Reproducibility Matters for Policy, and What Is at Stake

The consequences of fragile evidence are not abstract. Across economics and political science, research findings directly inform legislation, regulatory design, and spending decisions in Europe and beyond. Studies on minimum wage effects, financial regulation, anti-poverty interventions, land reform, and labor market policy have all shaped public debates and, in some cases, concrete policy choices. When the findings underlying those decisions turn out to be sensitive to minor analytical choices that never appear in the published article, policy can end up built on sand.

The problem is compounded by an important asymmetry: Policymakers typically see only the final published result – a point estimate and a p-value – with no visibility into the hundreds of analytical choices that produced it. Those choices include which observations to include, how to define the main outcome, which control variables to add, and which estimation method to use. Each choice is defensible on its own, but the combination effectively gives researchers wide latitude to find statistically significant results even where the underlying signal is weak. This is sometimes called the "researcher degrees of freedom" problem (Simonsohn, Simmons and Nelson 2020), and it is one reason why results that look robust in a single study often fail to survive independent scrutiny.

The replication literature has already uncovered several high-profile cases where influential findings did not hold up. These are not necessarily cases of misconduct. Most reflect the cumulative effect of individually reasonable choices made in the absence of pre-specified analytical plans, combined with publication incentives that favor statistically significant and novel results. The large-scale exercise summarized in this brief provides, for the first time, a systematic picture of how widespread these problems are across two entire fields at their best-practice frontier.

Two distinct concepts matter here. Computational reproducibility asks whether independent researchers, using the same data and code, obtain the same results as the original authors. Robustness asks whether the core finding survives when equally defensible but different analytical choices are made – such as alternative control variables, sample restrictions, or estimation methods. Both matter for policy: A result that cannot be reproduced is meaningless, and one that evaporates under modest methodological variation provides a fragile basis for intervention.

At a Glance

Table 1 summarizes the five main dimensions of the study. Together, they show high computational reproducibility thanks to recent open-data reforms, moderate robustness of significant results, stable effect sizes on average, but meaningful rates of coding errors and strong evidence of publication bias.

Table 1 Summary of Main Findings

85%	72%	99%	~25%
Computational reproducibility	Robustness rate	Median replicated effect size	Studies with coding errors

Dimension	Finding
Computational reproducibility	85% of published claims were reproduced using the original data and code. The availability of replication packages rose from 59% in 2014 to ~90% in 2021–23, driven by mandatory data editor policies.
Robustness of significant results	72% of originally significant estimates remain significant under alternative analytical choices. However, the rate falls to 45% when the dependent variable is redefined.
Effect size stability	The median reanalyzed effect equals 99% of the published effect. Results are directionally stable; fragility manifests at significance margins. However, 31% of individual estimates fall well outside the plausible range.
Coding errors	~25% of studies contain non-trivial errors: duplicated observations, incomplete difference-in-differences specifications, mislabeled treatment variables, misspecified estimators. Most go undetected through standard editorial checks.
Publication bias	Published t-statistics cluster just above t = 1.96. This spike is absent from reanalyses. The average published p-value is 0.17, compared to 0.22 for reanalyses – a gap of over 10 percentage points in significance rates.

Source: Brodeur et al. (2026). © ifo Institute

Can Independent Researchers Get the Same Results?

The first test was straightforward: Take the original data and code, run them independently, and check whether the same numbers come out. This is called computational reproducibility, and it is the minimum standard for scientific transparency.

The result was encouraging but imperfect: 85 percent of findings were reproduced successfully (Brodeur et al. 2026). This is substantially better than earlier studies, which found that the majority of economics papers from leading journals could not be reproduced (Chang and Li 2022; Gertler, Galiani and Romero 2018). The improvement is largely driven by journals introducing mandatory data-sharing policies and appointing dedicated data editors to check submissions before publication. Across the 12 journals studied, the share of papers providing full replication packages rose from 59 percent in 2014 to over 90 percent by 2023 – with the biggest jumps in the year immediately after a data editor was appointed (Askarov, Doucouliagos, Doucouliagos and Stanley 2023; Brodeur, Cook and Neisser 2024).

Even so, an 85 percent rate means roughly one in seven published findings – even at the most transparent journals in the field – cannot be verified from the materials provided. It is reasonable to assume that in journals that do not require data sharing, the situation is almost certainly worse.

Do Results Hold Up Under Different Reasonable Choices?

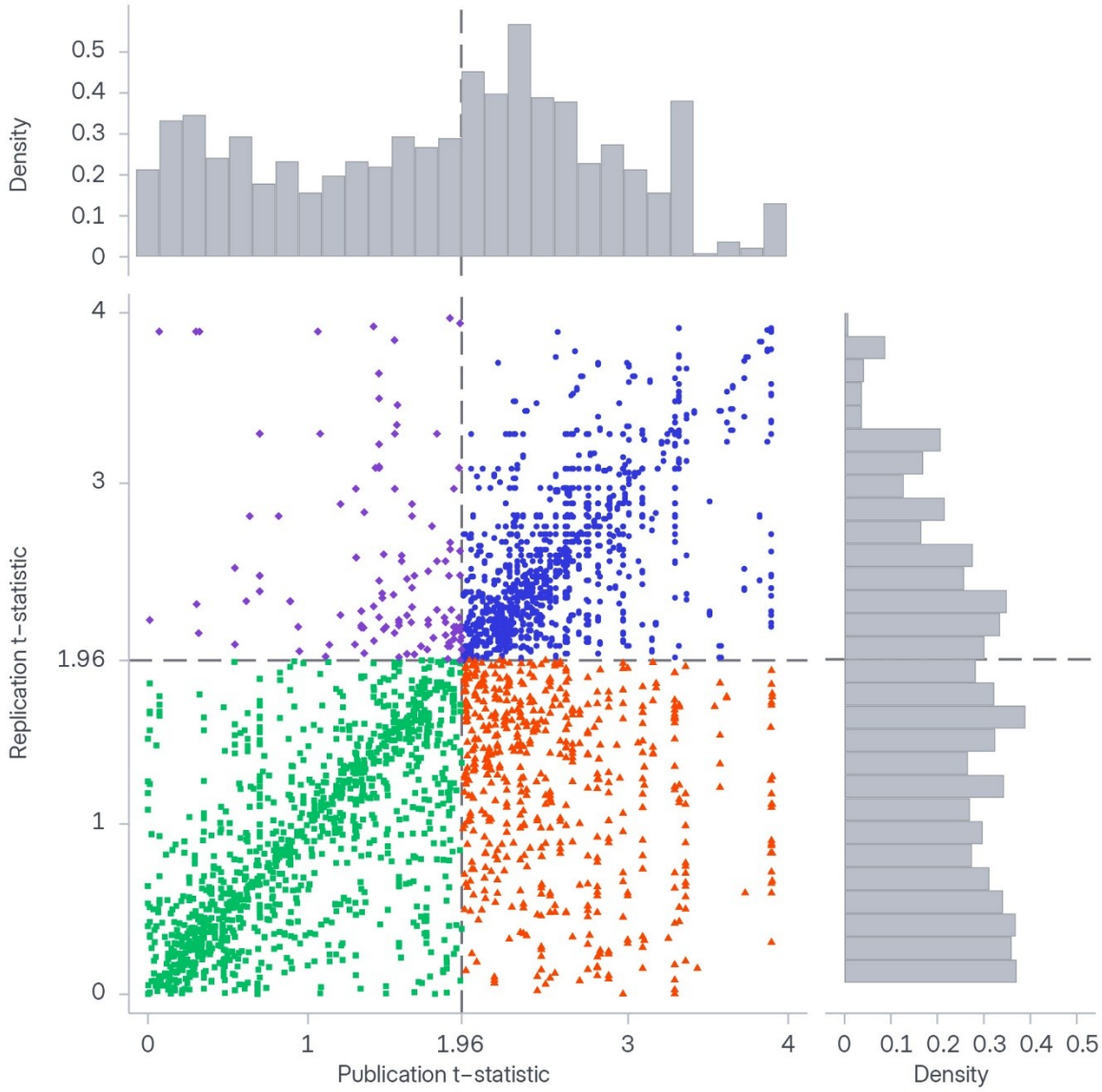
Reproducing the same number from the same code is just the first hurdle. The more important question is whether the finding itself is solid – or whether it depends on specific choices the original authors made, such as which control variables to include, which observations to keep, or how to measure the outcome of interest. Independent researchers tested this by re-running each analysis with different but equally defensible choices.

The overall robustness rate was 72 percent: Nearly three in four statistically significant results ($p \leq 0.05$) remained statistically significant in the same direction under reanalysis (Brodeur et al. 2026). The fragility varied by what was changed. Reanalyses that altered how the dependent variable was measured had the lowest robustness rate at just 45 percent – meaning more than half of results did not survive. Such changes are often seemingly innocuous: A study of an anti-poverty program might, for instance, measure household welfare by income rather than consumption, or define employment as any paid work rather than full-time work. These are choices that should not, in principle, overturn a genuine effect. Reanalyses that changed the independent variable had the highest rate at 78 percent. When researchers tested findings against entirely new data, 87 percent held up.

Effect sizes tell a similarly nuanced story. The median reanalysis effect was 99 percent of the originally published effect, suggesting the central tendency is stable. However, 6 percent of reanalyses produced an estimate of the opposite sign, 9 percent produced an estimate less than half the original magnitude, and 16 percent produced an estimate more than double the original. In all, roughly 31 percent of effect size estimates fell well outside the original range – a non-trivial share for evidence used in policy design.

Figure 1

Statistical Significance of Publication and Re-Analysis



Note: Each dot represents a pair of test statistics – the originally published t-statistic (horizontal axis) and the corresponding reanalysis t-statistic (vertical axis). The dashed lines mark t = 1.96, the conventional threshold for statistical significance at the 5 percent level. Blue circles (50 percent of the sample): The estimate is originally significant and remains significant in the reanalysis. Red triangles (14 percent of the sample): The estimate is originally significant and loses significance. Green squares (27 percent of the sample): The estimate is insignificant in both the original and the reanalysis. Purple diamonds (3 percent of the sample): The estimate becomes significant under reanalysis. The spike in published t-statistics just above 1.96 – visible in the top histogram but entirely absent from the reanalysis histogram – is a hallmark of publication bias: Results are disproportionately reported just above the significance threshold.
Source: Brodeur et al. (2026).

© ifo Institute

Figure 1 makes the publication bias story vivid. The top histogram shows a clear spike in published t-statistics just above 1.96 – the threshold for statistical significance at the 5 percent level ($p \leq 0.05$). Under a neutral reporting process, the distribution of t-statistics should be smooth. Instead, there are suspiciously many results just above the line and suspiciously few just below it. This pattern disappears entirely in the reanalysis histogram on the right, confirming that it reflects selective reporting in the original literature rather than a genuine feature of the data (Brodeur, Cook and Heyes 2020; Brodeur, Lé, Sangnier and

Zylberberg 2016). Across the full sample, 53 percent of published results were statistically significant, compared to just 43 percent of reanalyses – a gap of more than 10 percentage points.

Coding Mistakes Are Common and Often Go Undetected

Beyond checking whether results hold up, the independent researchers also examined the computer code used to produce the original findings. Setting aside minor technical issues like missing file paths, they found substantive coding errors in roughly 25 percent of the 110 studies (Brodeur et al. 2026). Economics papers had a higher error rate (26 percent) than political science papers (16 percent), probably because economics analyses tend to involve more complex code with more opportunities for mistakes.

The errors ranged widely: incorrect definition of the main outcome variable, mislabeled treatment variables, wrong control variables, flawed sample restrictions, and model misspecification (for example, estimation samples that inadvertently retained duplicated observations, or code that estimated a different model than the one described in the article).

Most of these errors went undetected through normal channels. Journal data editors check that code runs and produces matching output – but they do not check whether the code is actually doing what the paper claims, nor do they run alternative specifications (Vilhuber, Turrito and Welch 2020). The true error rate is also probably higher than 25 percent, since many studies did not provide raw data or data cleaning code, limiting what outside researchers could check.

Policy Implications

Treat Published Estimates as Uncertain Ranges, Not Fixed Facts

The 72 percent robustness rate means that for any given statistically significant result, there is roughly a one-in-four chance that a competent independent team making equally defensible choices would not confirm it at conventional significance levels. Commissioning evidence reviews should include requests for sensitivity analyses that vary key modeling choices – not just the point estimates reported in abstracts. Where a finding only holds under one specific set of assumptions, that fragility should be visible to decision-makers.

Prioritize Independently Reproduced Evidence

Findings that have been independently reproduced and confirmed robust provide a substantially stronger basis for policy. Institutions that compile evidence for policymakers should develop and maintain registers of reproduced evidence – analogous to systematic review registries in medicine – to help non-specialist decision-makers distinguish robust from fragile findings.

Avoid Scaling Up Programs on Single-Study Evidence

Several reproduction reports found that key results were sensitive to data errors or modeling choices not apparent from the published article alone. The risk is highest for non-experimental observational studies

with large researcher degrees of freedom. Where independent replication is absent, pilots with independent evaluation should precede full-scale rollout.

Extend Mandates for Open Data and Data Editors

The most actionable policy lever is editorial. The introduction of data editors and mandatory sharing policies appears to be a key driver of the improvement in reproducibility rates between 2014 and 2023. This improvement is concentrated in the top journals examined here; most economics and political science journals still lack such policies.

Incentivize Pre-Registration and Robustness Disclosure

Much of the gap between published significance and reanalysis robustness reflects researcher degrees of freedom: choices about variable construction, sample composition, and estimation made after observing the data. Pre-registration – publicly committing to hypotheses, variable definitions, and model specifications in a time-stamped registry before the data are analyzed – may substantially reduce this gap. For non-experimental observational studies where full pre-registration is impractical, journals and funders should require robustness disclosure: systematic reporting of how key results change under pre-specified alternative analytical choices.

References

Askarov, Z., A. Doucouliagos, H. Doucouliagos, and T. Stanley (2023), “The Significance of Data-sharing Policy”, *Journal of the European Economic Association* 21, 1191–1226.

Brodeur, A., N. Cook, and A. Heyes (2020), “Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics”, *American Economic Review* 110, 3634–3660.

Brodeur, A., N. Cook, and C. Neisser (2024), “P-Hacking, Data Type and Data-Sharing Policy”, *Economic Journal* 134, 985–1018.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016), “Star Wars: The Empirics Strike Back”, *American Economic Journal: Applied Economics* 8, 1–32.

Brodeur, A., D. Mikola, N. Cook et al. (2026), “Reproducibility and Robustness of Economics and Political Science Research”, *Nature* 652, 151–156.

Chang, A. C. and P. Li (2022), “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ‘Often Not’”, *Critical Finance Review* 11, 185–206.

Gertler, P., S. Galiani, and M. Romero (2018), “How to Make Replication the Norm”, *Nature* 554, 417–9.

Simonsohn, U., J. P. Simmons, and L. D. Nelson (2020), “Specification Curve Analysis”, *Nature Human Behaviour* 4, 1208–1214.

Vilhuber, L., J. Turrito, and K. Welch (2020), “Report by the AEA Data Editor”, *AEA Papers and Proceedings* 110, 764–75.