Article

Reproducibility and robustness of economics and political science research

https://doi.org/10.1038/s41586-026-10251-x

Received: 9 June 2025

Accepted: 6 February 2026

Published online: 1 April 2026

Check for updates

Science aspires to be cumulative. Reproducibility efforts strengthen science by testing the reliability of published findings, promoting self-correction, and informing policy-making[1]. Computational reproductions, whereby independent researchers reproduce the results of published studies, are an essential diagnostic tool[2–10]. Such efforts should have greater visibility[11–16]. However, little social science reproduction and robustness has been conducted at scale[10,13,17–23]. Here we reproduced original analyses and conducted robustness checks of 110 articles that were published in leading economics and political science journals with mandatory data and code sharing policies[17,18]. We found that more than 85% of published claims were computationally reproducible. In robustness checks, our reanalyses showed that 72% of statistically significant estimates remain significant and in the same direction, and the median reproduced effect size is nearly the same as the originally published effect size (that is, 99% of the published effect size). Additionally, 6 independent research teams examined 12 pre-specified hypotheses about determinants of robustness. Research teams with more experience found lower levels of robustness, and robustness did not correlate with author characteristics or data availability.

This project is a mega-reproduction, led by the Institute for Replication (I4R), that evaluates the reproducibility and robustness of 110 published studies in economics (79) and political science (31). Our focus is on studies published in 12 economics and political science journals between 2022 and 2023. Each of these journals has a data and code availability policy that requires authors to publicly share their materials upon publication, and most (but not all) also have a dedicated data editor. This editor is responsible for enforcing the journal's data and code policy and conducting internal computational reproducibility checks for accepted studies (Supplementary Information).

Not all studies from our targeted journals were chosen for reproduction and robustness analysis, and our sample is thus not a random representative sample of studies in economics and political science. Our approach leads to an over-representation of studies that use publicly available data[18]. Another feature of our sample is that the targeted journals have a data availability policy and enforce it, in contrast to many other journals in economics and political science. Our sample should thus be viewed as highly selective in terms of both impact and high data and code availability rates, and might present an optimistic upper bound on reproducibility rates. In fact, nearly all papers in our sample include replication packages with cleaned data and code to reproduce the results of the paper, and about 30% of the papers also provide the raw data and cleaning code used to generate the analytical data (Extended Data Fig. 1, levels 8–10).

This project relates to the broader reproducibility movement in psychology, neuroscience and biomedicine, and distinguishes itself from notable social science replication efforts along four key dimensions[24–26]. First, we are mostly reproducing non-experimental studies using the same data as the original authors. Second, we assess computational reproducibility and test the robustness of estimates to alternative specification choices. Because of the unique nature of the underlying studies—largely non-experimental work that uses observational data—we offer evidence about the general robustness of economics and political science. Third, we concentrate on recent studies in economics and political science. Finally, this is an ongoing initiative that aims to expand across disciplines, with the goal of mass reproduction of studies and reshaping of research norms at scale. This paper reports findings from the first 110 reproductions.

Definitions

We follow the nomenclature established in ref. 27 and define a claim to be 'computationally reproducible' if its results can be reproduced using the data and protocols from the original study. A claim is 'robust' if the results that support the claim are robust to alternative reasonable analytical decisions on the same data. Finally, a claim is 'replicable' if its results can be repeated using new data.

Teams and choice of study

The reproductions and replications in this project are generated in one of two streams. First, I4R has a board of editors who recommend potential reproducers. Second, I4R held 11 events called replication games (hereafter games)[28]. Games are one-day events that are open to faculty, postdocs, graduate students and other researchers. Participants are assigned to a small team of three to five other researchers working in the same subfield (for example, development economics).

Participating teams are offered a shortlist of studies (an average of five) in their subfield of interest about three weeks before the games. They are asked to select a paper as a team and familiarize themselves with the data and codes publicly posted by the original authors (the 'replication package') before the games. After the games, teams submit a standardized reproduction report summarizing their results.

A list of authors and their affiliations appears online. ✉e-mail: abrodeur@uottawa.ca

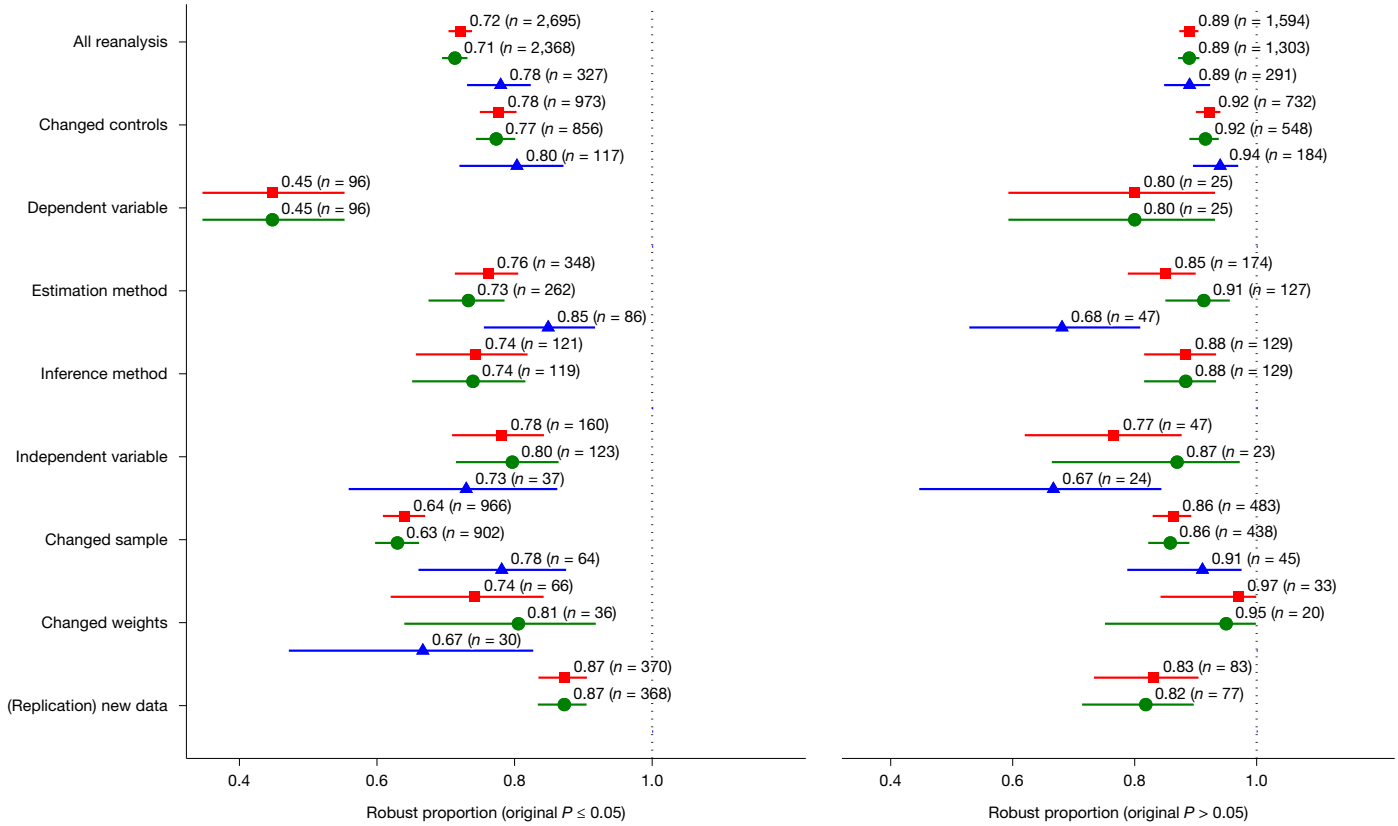


Fig. 1 | Robustness rate. Robustness rate for originally statistically significant research (left) and originally statistically insignificant research (right) in the full sample (red squares), the economics subsample (green circles) and the political science subsample (blue triangles). Squares, circles and triangles represent proportions, with 95% Clopper–Pearson confidence intervals presented as whiskers. Each group of three estimates represents different types of reanalysis that are non-mutually exclusive: all reanalysis, all types of reanalysis for all data, all types of reanalysis in economics and all types of reanalysis in political science; changed controls, reanalyses that changed the control variables—for example, by adding or redefining them; dependent variable, reanalyses that changed the dependent variable—for example, by using a different standardization or binarization; estimation method, reanalyses that changed the estimation method—for example, by adjusting a matching procedure; inference method, reanalyses that changed the inference method—for example, the level on which standard errors are clustered; independent variable, reanalyses that changed the main independent variable—for example, by taking into account treatment intensity; changed sample, reanalyses that changed the sample—for example, by excluding outliers; changed weights, reanalyses that changed the weights applied, or applied weights for the first time; (replication) new data, reanalyses that use new data (replication), representing replicability rates for reanalyses that introduced new data—for example, comparable outcomes from more recent survey waves.

I4R emphasizes to reproducers that the goal is not to show that the results are not reproducible. Instead, the goal is to test whether the claims are reproducible and robust. This is key, as some reproducers might engage in reverse specification searching (selective reporting of insignificant results). I4R stresses the importance of reasonable robustness checks and recoding[29]. Reanalyses are sensible tests of the research question and are expected to be statistically valid and theoretically informed.

We survey the reasons why teams selected their paper (Extended Data Fig. 2): 13.6% of teams were assigned a study (that is, they did not choose which study to work on), about 45% of teams report 'methods used' as the reason for selection, 36% of teams reported 'because of the journal of publication' and about 25% cited the 'length of time to reproduce results'.

If a large portion of reproducers select papers based on the assumption that their findings are questionable, it could skew reproducibility rates downwards, as such studies might be more prone to revealing problematic outcomes. However, in this project, only a minimal fraction of teams indicated that they chose their paper because of ex ante beliefs that main results are not replicable (3.6%). We found that selecting a paper due to the reproducers' belief that the paper is not robust is inversely correlated with reproducer experience ($\rho = -0.19$, $P < 0.001$). Five per cent of teams indicated that their choice was based on statistical power, sample size and trust of original authors.

Data and computational reproducibility

We find a computational reproducibility rate of 85%. That is, when provided with the original data and code, independent researchers are able to reproduce the published results in economics and political science studies 85% of the time using either the raw and analytical data or the analytical data when the raw data were not provided. The remaining 15% of cases involved studies with only partial availability of code or data, or instances in which code did not run or produced inconsistent results (Extended Data Fig. 1 and Supplementary Information). Fixing paths, missing packages and software requirements were not considered failures of computational reproducibility. In those instances, we fixed paths and added missing packages and software requirements.

Our findings suggest high rates of—but far from perfectly—computationally reproducible results for leading journals. Our results are in contrast with several studies that document low computational reproducibility rates in economics[13,19,22]. This may in part reflect the effectiveness of editorial policies in journals that have introduced data editors and mandatory sharing of replication packages.

To provide context to these findings, we mapped data and code availability in all of our target journals between 2014 and 2023. As discussed in the Supplementary Information, data and code-sharing practices have markedly improved during this period. Replication folders were

attached to 59% of papers in 2014, and this increased to an apparently stable value close to 90% in 2021–2023 (Extended Data Figs. 3–6). Additionally, for journals that introduced data editors during this period, much of this improvement occurred during the first year following this change.

Robustness

For robustness, we directly compare original point estimates to the revised point estimates. This one-on-one comparison enables us to assess the robustness of a specific hypothesis test, in addition to the robustness of our entire sample. We thus examine several claims within a study and conduct robustness reproducibility and robustness for multiple claims.

Reproducers are then free to conduct any robustness or recoding exercises. They focus on the reproducibility of the claims and have access to the replication package, enabling them to directly test the robustness of the main results. This is a crucial advantage over the traditional review process, as reproducers may uncover coding errors and discrepancies between the paper and the provided code. They may also uncover coding decisions that were not discussed (or are difficult to find) in the article.

However, this flexibility also brings some disadvantages. As with the role of reviewers in the journal review process, reproducers spend different amounts of time and effort on their respective replication. Some reproducers are more experienced at coding, whereas others are more familiar with methods, or are simply unable to implement robustness checks owing to a lack of raw data (Extended Data Fig. 7). This means that reproducibility efforts and type of reanalysis vary across teams. Teams worked 13 ± 24 (mean ± s.d.) active days on the reproductions and robustness, and reports were 19 ± 14 (mean ± s.d.) pages long.

Figure 1 (all reanalysis) shows a robustness rate of 72%. This result means that when alternative analytical decisions were made on the same data, 72% of originally statistically significant estimates ($P < 0.05$) remained statistically significant ($P < 0.05$) in the original direction.

We find large differences by reanalysis type. The reanalysis type that has the highest robustness rate (78%) is changing the independent variable measure (examples include log transformations and discretization). The reanalysis type that has the lowest robustness rate (45%) is any that includes changing the dependent variable measure (for example, categorizing the variable or log transformation). When a replication (addition of new data—for example, from more recent survey waves or an alternative source) is applied, the replication rate is 87%.

The average robustness rates are 71% and 78% for economics and political science, respectively; the 6.7% difference between fields is statistically significant (two-sample difference in proportions $z = -2.52$, $P = 0.012$, $n_1 = 2{,}368$, $n_2 = 327$). The general pattern of the robustness rates is similar between economics and political science (with the exception of dependent variable and inference method, which were not applied by any of the political science reanalyses). Focusing on robustness rates for originally statistically insignificant findings, we find a robustness rate of 89%.

Supplementary Table 1 shows shifts in statistical significance between all significance regions. We find that 7.44% of reanalyses find an effect with the opposite valence or direction from the original result. By contrast, of the 62.84% of original analyses that were statistically significant, 70.5% remained significant and retained the same valence. Of particular note is the 15.06% of reanalyses that exhibited a statistically insignificant result for originally statistically significant analyses.

Figure 2 shows the distribution of test statistics for the original point estimates and the reanalyses. We find that 53% of the originally published test statistics are statistically significant (to the right of the statistical significance threshold). By contrast, 43% of reanalyses are statistically significant (above the statistical significance threshold in the vertical axis histogram). The simple difference in proportions

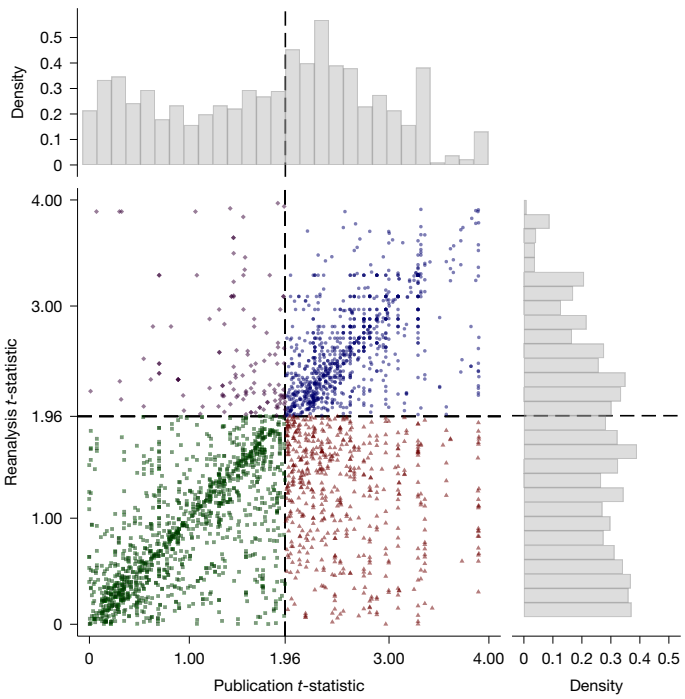


Fig. 2 | Statistical significance of publication and reanalysis. Top histogram, distribution of publication tests of significance. *t*-statistics greater than 4 are truncated for exposition. The bars of the histogram are of width 0.14, with exactly 14 bars between 0 and the statistical threshold of $t = 1.96$ (corresponding to statistical significance at the 5% level). Right histogram, distribution of reanalysis tests of significance. *t*-statistics greater than 4 are truncated for exposition. In the scatter plot, each marker is a pair of test statistics: an originally published test statistic (horizontal value) and an associated reanalysis test statistic (vertical value). If the original and reanalysis test statistics were identical, this scatterplot would follow the 45-degree line. As both axes represent statistical significance, we have bifurcated them with a line at $t = 1.96$, representing statistical significance at the 5% threshold. Blue circles indicate originally statistically significant statistics that are also statistically significant under reanalysis—this represents 50% of the sample. Red triangles indicate originally significant test statistics that are no longer statistically significant under reanalysis—this represents 14% of the sample. Green squares indicate originally statistically insignificant test statistics that remain insignificant under reanalysis—this represents 27% of the sample. Purple diamonds indicate originally statistically insignificant test statistics that become statistically significant under reanalysis—this represents 3% of the sample. Not displayed are the 6% of test statistics that represent a direction reversal between the originally estimated effect and the effect estimated under reanalysis.

is statistically significant (difference of 10.4%, McNemar's $\chi^2 = 264.11$, $P < 0.001$, $n = 4{,}750$).

The average originally published *t*-statistic is 1.797, whereas the average reanalysis *t*-statistic is 1.544. The difference between the paired original study estimates and reanalysis estimates is statistically significant (Wilcoxon signed-rank test $z = 15.477$, $P < 0.001$, $n = 3151$). Indeed, we reject the null hypothesis of a two-sample Kolmogorov–Smirnov test that the two distributions come from the same probability distribution ($P < 0.001$). We also note the large increase in test statistic density immediately after the statistical significance threshold (Extended Data Figs. 8 and 9), which offers strong evidence of publication bias in originally published research[30,31]. By contrast, this increase at the significance level threshold is missing from the vertical axis histogram depicting the distribution of reanalyses.

When expressed as *P* values, the average originally published *P* value is 0.167 whereas the average reanalysis *P* value is 0.219; the difference is statistically significant (Wilcoxon signed-rank test $z = -16.007$, $P < 0.001$, $n = 4{,}063$).

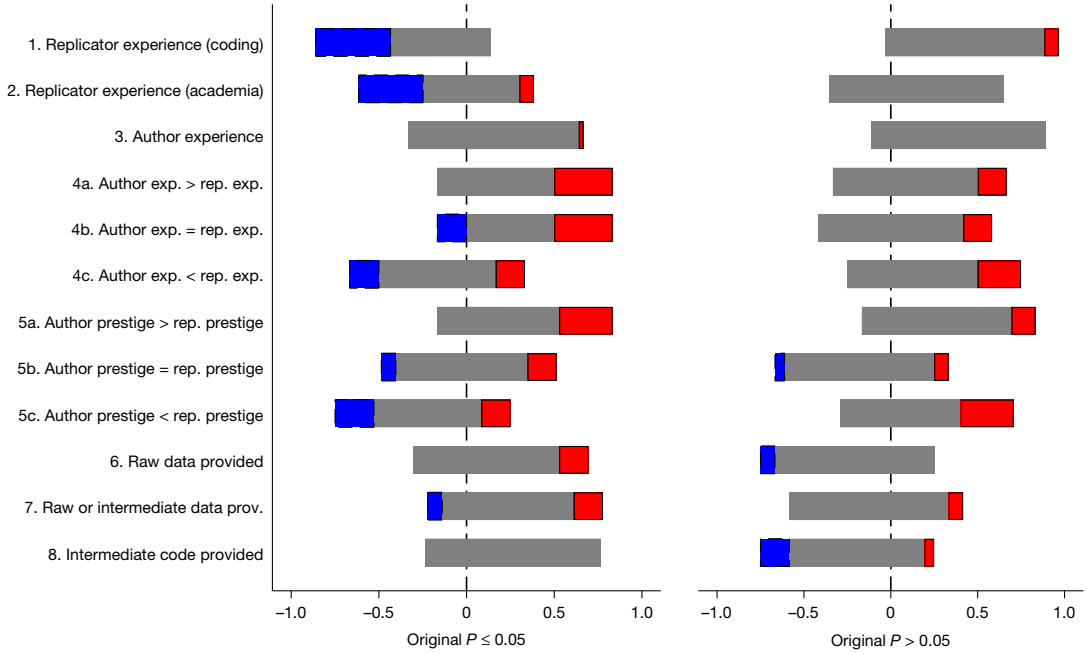


Fig. 3 | Robustness rate determinants. Six independent teams answered 12 questions about the reanalysis database. Each bar represents a different question: 'Does reproducibility of an originally statistically significant result depend on…' (left); and 'Does reproducibility of an originally statistically insignificant result depend on…' (right). Blue indicates the proportion of teams that indicated a negative and statistically significant relationship, in whichever manner the team defined in their analysis. Grey indicates the proportion of teams that indicated a statistically insignificant relationship; left of the zero line indicates negative and right of the zero line indicates positive. Red indicates the proportion of teams that indicated a statistically significant and positive relationship. All teams were equally weighted. Exp., experience; prov., provided; rep., replicator.

In this project, we conduct multiple reanalyses per original study, and so it is possible that much of the difference between original studies and their reanalyses is driven or characterized by large changes in a small subset of studies rather than indicative of more general shifts between original and reanalysis. In fact, we find evidence of general shifts. The proportion of original studies that have at least one statistically significant result is 95.3%, whereas for the corresponding reanalyses this is 92.9% (difference of 2.4%, McNemar's $\chi^2 = 1.00$, $P < 0.625$, $n = 86$). Only 3.6% of articles did not lose any statistical significance under replication, and the average replication lost statistical significance for 29% of replication tests (median of 22%). In only three original studies that reported statistically significant results, the reanalysis found that all test statistics were not statistically significant.

Determinants of robustness

This section examines what, if any, characteristics of the authors, reproducers or the original articles are informative of the robustness rate.

Although this analysis is merely exploratory, this project applied both pre-registration and many-analysts approaches[32–36]. By pre-specifying which research questions would be examined and averaging the responses to those research questions over multiple independent teams, the results here are guarded against specification searching and confirmation bias.

About 110 co-authors were invited to participate regarding the proposed determinants of robustness. We received answers from ten individuals and formed six many-analysts teams. Each team answered several research questions. The results are displayed in Fig. 3.

The teams began by analysing originally statistically significant results and answering the first question: "Does reproducibility/replicability rate depend on reproducers' experience coding?" Specifically, most of the teams estimated a negative coefficient in a regression with reproducibility as the dependent variable and a measure of their choosing for reproducers' experience as the primary independent variable—that is, the relationship is far more likely to be negative than positive. We interpret this result to mean that reproducers who are more experienced (broadly defined, as each of the many analysts defined experience independently) are better able to detect non-robust results in their chosen paper, similar to the notion of the 'trained eye' of a detective finding subtle clues the untrained eye may miss at the scene. The remaining 11 pre-specified hypotheses that the analysts tested were whether reproducibility is associated with: (2) reproducers' experience in academia; (3) the original authors' experience in academia; (4) whether authors have more (4a), similar (4b) or less (4c) experience than reproducers; (5) whether authors have more (5a), similar (5b), or less (5c) prestige (their institution, defined independently by the analysts) than reproducers; (6) whether raw data were provided; (7) whether raw or intermediate data were provided; and (8) whether cleaning code was provided .

Among results that were originally statistically significant, the first hypothesis yielded the clearest finding: the more experience a reproducer team had, the lower the robustness rate they found. Thus, one plausible interpretation of our main results is that robustness in our full sample would probably have been lower if equally highly qualified replicator teams had been assigned to each paper. However, according to the results ('Determinants of robustness'), the provision of raw or intermediate data or the necessary cleaning codes does not seem to affect the robustness of research.

When analysts examined these same 12 hypotheses for originally statistically insignificant results, the relationships were far more likely to be positive than negative, but (as indicated by the proportion in grey) the relationships were often not statistically significant.

Effect size

Figure 4 displays publication and reanalysis effect sizes. In economics and political science, effect sizes are largely reported as regression coefficients with units, whereas in other sciences, effect sizes are often reported using more comparable measures such as Cohen's *d*. Because

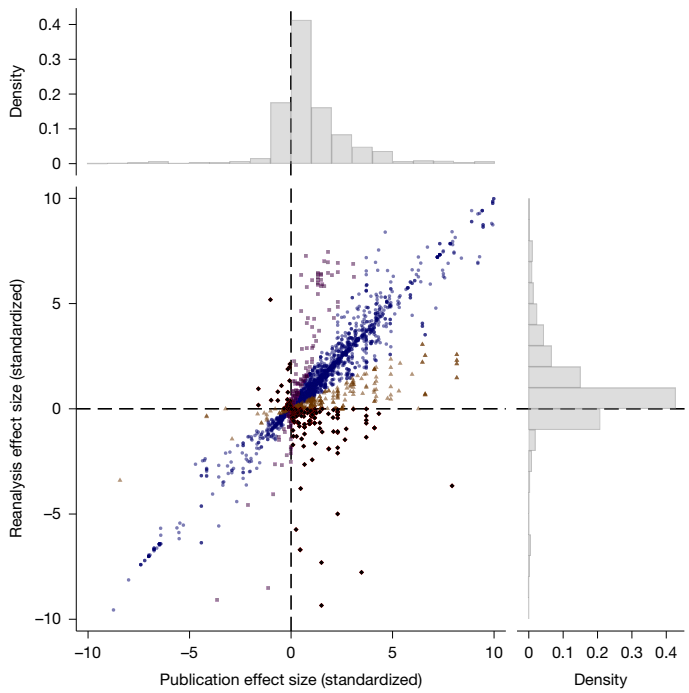


Fig. 4 | Effect size of publication and reanalysis. Top histogram, distribution of originally published effect sizes standardized by the average effect size within a published article. Right histogram, distribution of reanalysis published effect sizes standardized by the average effect size within a published article. Scatter plot, each marker is a pair of effect sizes: the originally published effect size (horizontal value) and an associated reanalysis effect size (horizontal value). If an originally estimated and reanalysis effect size were of similar magnitude (and direction), the markers would gather tightly around the 45-degree line that passes through the origin. Blue circles indicate effect sizes that are similar (between 50% to 200% of the original effect size) under reanalysis—this represents 69% of the sample. Red diamonds indicate effect size estimates that switch direction under reanalysis—this represents 6% of the sample. Orange triangles indicate effect size estimates that are 50% or less of the original magnitude under reanalysis—this represents 9% of the sample. Purple squares indicate effect size estimates that are double or larger of the original magnitude under reanalysis—this represents 16% of the sample.

raw effect sizes vary widely between original studies, each of the markers is standardized by the within-article average published effect size (for example, estimated effects of 2, 4 and 6 are standardized within a publication to 0.5, 1.0 and 1.5).

We find that, on average, the median effect size of a reanalysis is equivalent to the published effect size (that is, 99% the size of the published effect) and the mean replicated effect is 9% larger than the original. Extended Data Fig. 10 illustrates the distribution effect sizes of reanalyses. This result is in stark contrast to previous projects focused on replication with new data in psychology or social science experiments[24–26], which uncovered replication rates ranging from 50% to 66%. Three major differences between our project and these replication efforts are that we focus on robustness as opposed to replication with new data, we focus on recently published articles, and our sample is composed mostly of non-experimental studies that utilize secondary data.

Coding errors and recoding

We also investigated the prevalence of coding errors and discrepancies between the code and article. Computational reproducibility pertains to the ability of the provided replication folder to reproduce the exhibits and statistics displayed in the research (that is, manuscripts and appendices). Reproducers may be able to reproduce all exhibits exactly as they appear (computationally reproducible), but the exhibits may have been constructed with coding errors or discrepancies.

Except for minor inconveniences (such as missing packages or broken pathways), we identify coding errors in approximately 25% of the studies, with some studies containing multiple errors (Supplementary Information). The prevalence of coding errors is larger for economics (26%) than for political science (16%). Types of error include defining the dependent variable, defining the main independent variable, defining control variables and mis-specification of the estimation, model, inference or sample. Although not all of these coding errors affect the conclusions of the original studies, we uncover several significant errors that warrant discussion. These major errors include instances of duplicated observations on a large scale, incomplete interaction in a difference-in-differences model, mislabelling of the main treatment variable for a substantial number of (or all) observations, and using different models or estimators from the ones reported in the article.

It is important to note that this 25% figure is likely to underestimate the true prevalence of coding errors. Reproducers may have missed some errors, and many replication packages do not include raw data or data-cleaning code, limiting the ability to detect additional issues.

A number of reproducers also recoded the analysis using different statistical software. Out of 23 recoding exercises, we find major differences for 3 studies and minor differences for 10 studies. Two of the major differences were uncovered when using different software and examining the code from the original study. Additionally, one team who computationally reproduced the results using a different version of the software used in the original study uncovered noteworthy differences in the magnitude and significance of the estimates (Supplementary Information).

Communication with original authors

I4R shares completed reproduction reports with original authors before public release[28]. Reports are typically reviewed by A. Brodeur or another board member, mainly for tone and structure. I4R then disseminates the report and any author response simultaneously (a full list of reports is provided in the Supplementary Information). Reproducers may revise their reports after receiving feedback from original authors.

About 95% of contacted authors responded (including one case in which an author was unreachable after leaving academia). Among respondents, 11% provided only brief notes or indicated that they could not respond, 59% offered informal feedback, and 30% supplied a formal response. For comparison, a previous study reported that around 25% of authors in their sample provided a formal response[37].

Around two-thirds of reproducers indicated that interactions with original authors improved their reports, often by clarifying variables or procedures, supplying data or data-access instructions, or helping to adjust tone. In one case, the original authors conducted additional robustness checks in their non-public files at the reproducers' request.

Finally, we assess agreement between authors and reproducers. Authors' final responses were coded for whether disagreements remained after mediation. Only 23% of articles showed any remaining disagreement. Further details are presented in the Supplementary Information.

Discussion

A substantial information asymmetry exists between authors and the broader academic community, including reviewers and editors[30]. Reviewers rarely see the underlying data and code and may be unaware of crucial coding decisions, even as journals routinely request multiple robustness checks. This limited visibility means that major errors or inconsistencies can go undetected.

Large-scale reproducibility initiatives offer a promising way to address these challenges in the social sciences and beyond. Our project provides a systematic, scalable approach to evaluating reproducibility and robustness, with the goal of increasing transparency and improving the credibility of published research. Although stronger incentives to conduct reproducibility and robustness remain necessary, we do not attempt to evaluate which specific incentives would be most effective, as doing so would require speculation beyond the scope of our data. Identifying the most effective incentives is an important research question that we hope will be addressed in future work.

Given the low prevalence of diagnostic replication in published work[38], the scale of this ongoing effort could shift research norms. By encouraging more rigorous methodologies, deterring questionable research practices and emphasizing collaboration, it may help place greater weight on the reliability of results in publication decisions.

Although our journal sample is selective, the findings are encouraging and suggest a high level of computational reproducibility. These patterns—and the existence of a large-scale, community-driven effort—may strengthen trust in published results.

We also asked reproducers about the quality of the replication packages that they examined. More than 40% reported gaining a more optimistic view of the discipline, whereas approximately 5% developed a more negative opinion. This suggests that mass reproduction of studies accompanied by replication packages can directly enhance researchers' trust in scientific findings.

The success and scalability of this initiative have been driven by the intrinsic motivation of participating researchers to support open science and improve their technical skills. By late 2025, I4R had organized 80 replication games involving more than 3,500 researchers, with events held every other week. These efforts show that the skilled labour needed for large-scale reproduction can emerge organically from an engaged research community.

The project also has the potential to advance science and improve equity. Publicly posting data and code facilitates learning, speeds methodological diffusion and enables independent verification. Reproducing analyses in open-source software can also help to level the playing field for researchers who lack access to expensive licenses.

Our results have limitations. Only a small number of economics and political science journals currently require data and code[17,18], and fewer still check reproducibility[39]. Thus, our findings largely reflect leading journals with strong data-sharing norms. Future research should assess reproducibility more broadly by examining a random sample of papers from journals with and without data availability policies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-026-10251-x.

1. Vazire, S. Quality uncertainty erodes trust in science. *Collabra Psychol.* **3**, 1 (2017).
2. Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M. & Stodden, V. Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* **11**, 8–18 (2008).
3. King, G. Replication, replication. *Polit. Sci. Polit.* **28**, 444–452 (1995).
4. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).
5. Marcoci, A. et al. Predicting the replicability of social and behavioural science claims in COVID-19 preprints. *Nat. Hum. Behav.* **9**, 287–304 (2025).
6. Milkowski, M., Hensel, W. M. & Hohol, M. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *J. Comput. Neurosci.* **45**, 163–172 (2018).
7. Moonesinghe, R., Khoury, M. J. & Janssens, A. C. J. W. Most published research findings are false—but a little replication goes a long way. *PLoS Med.* **4**, e28 (2007).
8. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science* (National Academies Press, 2019).
9. Peterson, D. & Panofsky, A. Self-correction in science: the diagnostic and integrative motives for replication. *Soc. Stud. Sci.* **51**, 583–605 (2021).
10. Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R. & Debonnel, E. Certify reproducibility with confidential data. *Science* **365**, 127–128 (2019).
11. Brandon, A. & List, J. A. Markets for replication. *Proc. Natl Acad. Sci. USA* **112**, 15267–15268 (2015).
12. Freese, J. & Peterson, D. Replication in social science. *Annu. Rev. Sociol.* **43**, 147–165 (2017).
13. Gertler, P., Galiani, S. & Romero, M. How to make replication the norm. *Nature* **554**, 417–9 (2018).
14. Maniadis, Z. & Tufano, F. The research reproducibility crisis and economics of science. *Econ. J.* **127**, F200–F208 (2017).
15. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
16. Nosek, B. A. et al. Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
17. Askarov, Z., Doucouliagos, A., Doucouliagos, H. & Stanley, T. The significance of data-sharing policy. *J. Eur. Econ. Assoc.* **21**, 1191–1226 (2023).
18. Brodeur, A., Cook, N. & Neisser, C. P-hacking, data type and data-sharing policy. *Econ. J.* **134**, 985–1018 (2024).
19. Chang, A. C. & Li, P. Is economics research replicable? Sixty published papers from thirteen journals say 'often not'. *Crit. Finance Rev.* **11**, 185–206 (2022).
20. Christensen, G. & Miguel, E. Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **56**, 920–80 (2018).
21. Dafoe, A. Science deserves better: the imperative to share complete replication files. *Polit. Sci. Polit.* **47**, 60–66 (2014).
22. McCullough, B., McGeary, K. A. & Harrison, T. D. Do economics journal archives promote replicable research?. *Can. J. Econ.* **41**, 1406–1420 (2008).
23. Pérignon, C. et al. Computational reproducibility in finance: evidence from 1,000 tests. *Rev. Financ. Stud.* **37**, 3558–3593 (2024).
24. Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433–1436 (2016).
25. Camerer, C. F. et al. Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).
26. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
27. Dreber, A. & Johannesson, M. A framework for evaluating reproducibility and replicability in economics. *Econ. Inq.* **63**, 338–356 (2025).
28. Brodeur, A., Dreber, A., Hoces de la Guardia, F. & Miguel, E. Replication games: how to make reproducibility research more systematic. *Nature* **621**, 684–686 (2023).
29. Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nat. Hum. Behav.* **4**, 1208–1214 (2020).
30. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star Wars: the empirics strike back. *Am. Econ. J.* **8**, 1–32 (2016).
31. Brodeur, A., Cook, N. & Heyes, A. Methods matter: p-hacking and publication bias in causal analysis in economics. *Am. Econ. Rev.* **110**, 3634–3660 (2020).
32. Botvinik-Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
33. Breznau, N. et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl Acad. Sci. USA* **119**, e2203150119 (2022).
34. Huntington-Klein, N. et al. The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* **59**, 944–960 (2021).
35. Menkveld, A. J. et al. Nonstandard errors. *J. Finance* **79**, 2339–2390 (2024).
36. Silberzahn, R. et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
37. Fišar, M. et al. Reproducibility in *Management Science*. *Manag. Sci.* **70**, 1343–2022 (2024).
38. Ankel-Peters, J., Fiala, N. & Neubauer, F. Do economists replicate?. *J. Econ. Behav. Org.* **212**, 219–232 (2023).
39. Vilhuber, L., Turrito, J. & Welch, K. Report by the AEA Data Editor. *AEA Pap. Proc.* **110**, 764–765 (2020).

Abel Brodeur[1]✉, Derek Mikola[1], Nikolai Cook[2], Lenka Fiala[3,4], Thomas Brailey[5], Ryan Briggs[6], Alexandra de Gendre[7], Yannick Dupraz[8], Jacopo Gabani[9,10], Romain Gauriot[11], Joanne Haddad[12], Goncalo Lima[13,14], Jörg Ankel-Peters[15], Anna Dreber[16], Douglas Campbell[17], Lamis Kattan[18], Diego Marino Fages[19], Fabian Mierisch[20], Pu Sun[21], Taylor Wright[22], Marie Connolly[23], Fernando Hoces de la Guardia[24], Magnus Johannesson[16], Edward Miguel[25], Lars Vilhuber[26], Alejandro Abarca[27], Mahesh Acharya[28], Sossou Simplice Adjisse[29], Ahwaz Akhtar[30], Eduardo Alberto Ramirez Lizardi[31], Sabina Albrecht[32], Synøve Nygaard Andersen[31], Zubaria Andlib[33,34], Falak Arrora[35], Thomas Ash[36], Etienne Bacher[37], Sebastian Bachler[38], Félix Bacon[39], Manuel Bagues[35], Timea Balogh[40], Alisher Batmanov[41], Mara Barschkett[42,43], Barış Kaan Basdil[44], Jaromír Baxa[45,46], Sascha O. Becker[35,47], Monica Beeder[48], Louis-Philippe Beland[49], Abdel-Hamid Bello[50], Daniel Benenson Markovits[51], Grant Benjamin[52], Thomas Bergeron[50], Moussa P. Blimpo[52], Marco Binetti[53], Carl Bonander[54], Joseph Bonneau[55], Endre Borbáth[56], Nicolai Borgen[57,58], Solveig Topstad Borgen[31], Jonathan Borowsky[59], Elisa Brini[60], Myriam Brown[39], Martin Brun[61], Stephan Bruns[62,63,64], Nino Buliskeria[65], Andrea Calef[66], Alistair Cameron[67], Pamela Campa[68], Santiago Campos-Rodríguez[69], Giulio Giacomo Cantone[70], Fenella Carpena[71], Perry Jess Carter[72], Paul Castañeda Dower[73], Ondrej Castek[74], Jill Caviglia-Harris[75], Gabriella Chauca Strand[76], Shi Chen[77], Sya In Chzhen[78], Jong Chung[79], Jason Collins[80], Alexander Coppock[81], Hugo Cordeau[52], Ben Couillard[52], Jonathan Crechet[1], Lorenzo Crippa[40], Jing Cui[82], Christian Czymara[83], Haley Daarstad[55], Danh Chi Dao[84], Daniel Dao[5], Marco David Schmandt[85], Astrid de Linde[31], Lucas De Melo[86], Lachlan Deer[87], Micole De Vera[88], Velichka Dimitrova[89], Jan Fabian Dollbaum[90], Jan Matti Dollbaum[91,92], Michael Donnelly[52], Luu Duc Toan Huynh[93], Tsvetomira Dumbalska[5], Jamie Duncan[52], Kiet Tuan Duong[94], Thibaut Duprey[95], Christoph Dworschak[94,96], Sigmund Ellingsrud[97], Ali Elminejad[65], Yasmine Eissa[98], Andrea Erhart[38], Giulian Etingin-Frati[99], Elaheh Fatemipour[35], Alexa Federice[55], Jan Feld[100], Guidon Fenig[1], Mojtaba Firouzjaeiangalougah[74], Erlend Fleisje[101], Alexandre FortiFriter-Chouinard[39], Julia Francesca Engel[102], Nadjim Fréchet[103], Reid Fortier[104], Tilman Fries[92], Michael James Frith[105], Thomas Galipeau[106], Sebastian Gallegos[107], Areez Gangji[108], Xiaoying Gao[94], Cloé Garnache[109], Attila Gáspár[110,111], Evelina Gavrilova[112], Arijit Ghosh[15], Garreth Gibney[113], Grant Gibson[114], Geir Godager[31], Leonard Goff[28], Da Gong[115], Javier González[116], Jeremy D. Gretton[117], Cristina Griffa[118], Idaliya Grigoryeva[41], Maja Grøtting[119], Eric Guntermann[25], Jiaqi Guo[120], Alexi Gugushvili[31], Hooman Habibnia[121], Sonja Häffner[122], Jonathan D. Hall[123], Olle Hammar[124,125], Amund Hanson Kordt[31], Barry Hashimoto[20], Jonathan S. Hartley[126], Carina I. Hausladen[127], Tomáš Havránek[45,128], Harry He[129], Matthew Hepplewhite[5], Mario Herrera-Rodriguez[130], Felix Heuer[15], Anthony Heyes[120], Anson T. Y. Ho[131], Jonathan Holmes[1], Armando Holzknecht[38], Yu-Hsiang Dexter Hsu[55], Shiang-Hung Hu[132], Yu-Shiuan Huang[133], Mathias Huebener[134], Christoph Huber[135], Kim P. Huynh[136,137], Zuzana Irsova[45,138], Ozan Isler[139], Niklas Jakobsson[140,141], Raphaël Jananji[50], Tharaka A. Jayalath[142], Michael Jetter[143], Jenny John[1], Rachel Joy Forshaw[144], Felipe Juan[145], Valon Kadriu[146], Sunny Karim[49], Edmund Kelly[5], Duy Khanh Hoang Dang[147], Tazia Khushboo[28], Jin Kim[148], Gustav Kjellsson[149,150], Anders Kjelsrud[109], Andreas Kotsadam[151], Jori Korpershoek[51], Lewis Krashinsky[52], Suranjana Kundu[152], Alexander Kustov[153], Nurlan Lalayev[35], Audrée Langlois[39], Jill Laufer[154], Blake Lee-Whiting[155], Andreas Leibing[156], Gabriel Lenz[25], Joel Levin[129], Peng Li[157], Tongzhe Li[6], Yuchen Lin[35], Ariel Listo[158], Dan Liu[159], Xuewen Lu[28], Elvina Lukmanova[160], Alex Luscombe[161], Lester R. Lusher[162], Ke Lyu[163], Hai Ma[164], Nicolas Mäder[165], Clifton Makate[166], Alice Malmberg[55], Adit Maitra[7], Marco Mandas[167], Jan Marcus[168], Shushanik Margaryan[169], Lili Márk[168], Andres Martignano[170], Abigail Marsh[171], Isabella Masetto[172], Anthony McCanny[52], Emma McManus[173], Ryan McWay[59], Lennard Metson[172], Jonas Minet Kinge[31], Sumit Mishra[174], Myra Mohnen[1], Jakob Moeller[121], Rosalie Montambeault[39], Sébastien Montpetit[35], Louis-Philippe Morin[1], Todd Morris[32], Scott Moser[175], Fabio Yoshio Suguri Motoki[176], Lucija Muehlenbachs[28,177], Andreea Musulan[50,178,179], Marco Musumeci[180], Munirul Nabin[11], Karim Nchare[181], Florian Neubauer[15], Quan M. P. Nguyen[182], Tuan Nguyen[62], Viet Nguyen-Tien[172], Ali Niazi[28], Giorgi Nikolaishvili[183], Ardyn Nordstrom[49], Patrick Nüß[184], Angela Odermatt[185], Matt Olson[186], Henning Øien[187], Tim Ölkers[188], Miquel Oliver i Vert[189], Emre Oral[190], Christian Oswald[191], Ali Ousman[164], Ömer Özak[192,193,194], Shubham Pandey[195], Alexandre Pavlov[50], Martino Pelli[196], Romeo Penheiro[197], RyuGyung Park[198], Eva Pérez Martel[12], Tereza Petrovičová[129], Linh Phan[55], Alexa Prettyman[199], Jakub Procházka[173], Aqila Putri[158], Julian Quandt[121], Kangyu Qiu[200], Loan Quynh Thi Nguyen[201], Andaleeb Rahman[26], Carson H. Rea[202], Adam Reiremo[203], Laëtitia Renée[204], Joseph Richardson[33], Nicholas Rivers[1], Bruno Rodrigues[205], William Roelofs[52], Tobias Roemer[5], Ole Rogeberg[151], Julian Rose[15], Andrew Roskos-Ewoldsen[55], Paul Rosmer[206], Barbara Sabada[95], Soodeh Saberian[207], Nicolas Salamanca[7], Georg Sator[170,208], Daniel Scates[55], Elmar Schlüter[209], Cameron Sells[20], Sharmi Sen[67], Ritika Sethi[210], Anna Shcherbiak[121], Moyosore Sogaolu[211], Matt Soosalu[49], Erik Ø. Sørensen[112], Manali Sovani[212], Noah Spencer[52], Stefan Staubli[28], Renske Stans[213], Anya Stewart[55], Felix Stips[214], Kieran Stockley[170], Stephenson Strobel[200], Ethan Struby[215,216,217], John P. Tang[218], Idil Tanrisever[25], Thomas Tao Yang[159], Ipek Tastan[28], Dejan Tatić[121], Benjamin Tatlow[170], Féraud Tchuisseu Seuyong[50], Rémi Thériault[219], Vincent Thivierge[1], Wenjie Tian[1], Filip-Mihai Toma[220], Maddalena Totarelli[221], Van-Anh Tran[67], Hung Truong[1], Nikita Tsoy[222], Kerem Tuzcuoglu[223], Diego Ubfal[9], Laura Villalobos[75], Julian Walterskirchen[224], Joseph Tao-yi Wang[225], Vasudha Wattal[226], Matthew D. Webb[49], Bryan S. Weber[227], Reinhard Weisser[228], Wei-Chien Weng[55], Christian Westheide[229,230], Kimberly White[92], Jacob Winter[52], Timo Wochner[231], Matt Woerman[232], Jared Wong[233], Ritchie Woodard[78], Marcin Wroński[234], Myra Yazbeck[1], Gustav Chung Yang[235], Luther Yap[236], Kareman Yassin[237], Hao Ye[238], Jin Young Yoon[84], Chris Yurris[164], Tahreen Zahra[49], Mirela Zaneva[5], Aline Zayat[1], Jonathan Zhang[239], Ziwei Zhao[240] & Yaolang Zhong[35]

[1]University of Ottawa, Ottawa, Ontario, Canada. [2]Wilfrid Laurier University, Waterloo, Ontario, Canada. [3]Institute for Replication, University of Ottawa, Ottawa, Ontario, Canada. [4]Institute for Replication, Tilburg University, Tilburg, The Netherlands. [5]University of Oxford, Oxford, UK. [6]University of Guelph, Guelph, Ontario, Canada. [7]The University of Melbourne, Carlton, Victoria, Australia. [8]Paris Dauphine University, PSL University, LEDA, CNRS, IRD, Paris, France. [9]World Bank, Washington, DC, USA. [10]Centre for Health Economics, University of York, York, UK. [11]Deakin University, Melbourne, Victoria, Australia. [12]Universitat Autonoma de Barcelona, Bellaterra, Spain. [13]European University Institute, Florence, Italy. [14]University of Bologna, Bologna, Italy. [15]RWI-Leibniz Institute for Economic Research, Essen, Germany. [16]Stockholm School of Economics, Stockholm, Sweden. [17]American University of Armenia, Santa Monica, CA, USA. [18]School of Foreign Service, Georgetown University Qatar, Doha, Qatar. [19]University Business Centre, Durham University, Durham, UK. [20]Independent Researcher, Bavaria, Germany. [21]Dongbei University of Finance and Economics, Dalian, China. [22]Brock University, St Catharines, Ontario, Canada. [23]UQAM, Montreal, Quebec, Canada. [24]Berkeley Initiative for Transparency in the Social Sciences, Berkeley, CA, USA. [25]University of California Berkeley, Berkeley, CA, USA. [26]Cornell University, Ithaca, NY, USA. [27]Texas Tech University, Lubbock, Texas, USA. [28]University of Calgary, Calgary, Alberta, Canada. [29]Oregon State University, Bloomfield, NM, USA. [30]George Washington University, Washington, DC, USA. [31]University of Oslo, Oslo, Norway. [32]University of Queensland, Brisbane, Queensland, Australia. [33]Lancaster University, Lancaster, UK. [34]Federal Urdu University of Arts Science and Technology, Islamabad, Pakistan. [35]University of Warwick, Coventry, UK. [36]Anderson School of Management UCLA, Los Angeles, CA, USA. [37]Luxembourg Institute of Socio-Economic Research (LISER), Esch-sur-Alzette, Luxembourg. [38]University of Innsbruck, Innsbruck, Austria. [39]Université Laval, Quebec City, Quebec, Canada. [40]University of Strathclyde, Glasgow, UK. [41]University of California San Diego, La Jolla, CA, USA. [42]University of Bonn, Bonn, Germany. [43]IZA Berlin, Berlin, Germany. [44]Risk Software Technologies, Istanbul, Turkey. [45]Institute of Economic Studies, Faculty of Social Sciences, Charles University, Prague, Czech Republic. [46]Institute of Information Theory and Automation, Czech Academy of Sciences, Prague, Czech Republic. [47]Monash University, Clayton, Victoria, Australia. [48]University of Southampton, Southampton, UK. [49]Carleton University, Ottawa, Ontario, Canada. [50]University of Montreal, Montreal, Quebec, Canada. [51]Columbia University, New York, NY, USA. [52]University of Toronto, Toronto, Ontario, Canada. [53]Institute of Intercultural and International Studies, University of Bremen, Bremen, Germany. [54]Karlstad Business School, Karlstad University, Karlstad, Sweden. [55]University of California Davis, Davis, CA, USA. [56]Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany. [57]Centre for Research on Equality in Education, University of Oslo, Oslo, Norway. [58]Department of Special Needs Education, University of Oslo, Oslo, Norway. [59]University of Minnesota, St Paul, MN, USA. [60]University of Florence, Florence, Italy. [61]Finnish Centre of Excellence in Tax Systems Research, Tampere University, Tampere, Finland. [62]Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium. [63]INCHER Kassel, Kassel, Germany. [64]METRICS Stanford, Stanford University, Stanford, CA, USA. [65]Nazarbayev University, Astana, Kazakhstan. [66]UCL School of Management, University College London, London, UK. [67]Laterite Ltd, Kigali, Rwanda. [68]Stockholm Institute of Transition Economics, Stockholm, Sweden. [69]University of California Irvine, Irvine, CA, USA. [70]"Magna Graecia" University of Catanzaro, Catanzaro, Italy. [71]Oslo Business School, Oslo Metropolitan University, Oslo, Norway. [72]NYU Abu Dhabi, Abu Dhabi, United Arab Emirates. [73]University of Wisconsin-Madison, Madison, WI, USA. [74]Masaryk University, Brno, Czech Republic. [75]Salisbury University, Salisbury, MD, USA. [76]Institute of Medicine, University of Gothenburg, Gothenburg, Sweden. [77]Zhejiang University, Hangzhou, China. [78]University of East Anglia, Norwich, UK. [79]Auburn University, Auburn, AL, USA. [80]University of Technology Sydney, Ultimo, New South Wales, Australia. [81]Northwestern University, Evanston, IL, USA. [82]Beijing Normal University, Beijing, China. [83]Netherlands Interdisciplinary Demographic Institute, Den Haag, Netherlands. [84]Queen's University, Kingston, Ontario, Canada. [85]TU Berlin, Berlin, Germany. [86]Nottingham Interdisciplinary Centre for Economic and Political Research, University of Nottingham, Nottingham, UK. [87]University of Melbourne, Carlton, Victoria, Australia. [88]Banco de España, Madrid, Spain. [89]Social Research Institute, University College London, London, UK. [90]University College Dublin, Dublin, Ireland. [91]University of Fribourg, Fribourg, Switzerland. [92]LMU Munich, Munich, Germany. [93]Queen Mary University of London, London, UK. [94]University of York, York, UK. [95]Bank of Canada, Ottawa, Ontario, Canada. [96]German Institute for Development Evaluation, Bonn, Germany. [97]BI Norwegian Business School, Oslo, Norway. [98]The American University in Cairo, Cairo, Egypt. [99]ETH Zurich, Zurich, Switzerland. [100]Victoria University of Wellington, Wellington, New Zealand. [101]Oslo Economics, Oslo, Norway. [102]Kiel University, Kiel, Germany. [103]Concordia University, Montreal, Quebec, Canada. [104]VisualAIM, Calgary, Alberta, Canada. [105]University of Edinburgh, Edinburgh, UK. [106]University of Toronto, Montreal, Quebec, Canada. [107]UAI Business School, Santiago, Chile. [108]Independent Researcher, Ottawa, Ontario, Canada. [109]Oslo Metropolitan University, Oslo, Norway. [110]ELTE Centre for Economic and Regional Studies, Budapest, Hungary. [111]Central European University, Vienna, Austria. [112]NHH Norwegian School of Economics, Bergen, Norway. [113]University of Galway, Galway, Ireland. [114]Canadian Research Data Centre Network, McMaster University, Hamilton, Ontario, Canada. [115]State University of New York Geneseo, Geneseo, NY, USA. [116]Southern Methodist University, Dallas, TX, USA. [117]Behavioural Insights Unit, Ontario Public Service, Toronto, Ontario, Canada. [118]University of Chile, Santiago, Chile. [119]The Norwegian Institute of Public Health, Oslo, Norway. [120]University of Birmingham, Birmingham, UK. [121]Vienna University of Economics and Business, Vienna, Austria. [122]Peace Research Institute Oslo, Oslo, Norway. [123]University of Alabama, Tuscaloosa, AL, USA. [124]Linnaeus University, Växjö, Sweden. [125]Institute for Futures Studies, Stockholm, Sweden. [126]Stanford University, Stanford, CA, USA. [127]University of Konstanz, Konstanz, Germany. [128]Faculty of International Relations, Prague University of Economics and Business, Prague, Czech Republic. [129]University of California San Diego, San Diego, CA, USA. [130]Programa Estado de la Nacion, CREST-Ecole Polytechnique, Palaiseau, France. [131]Toronto Metropolitan University, Toronto, Ontario, Canada. [132]California Institute of Technology, Pasadena, CA, USA. [133]Department of Political Science, National Chengchi University, Taipei, Taiwan. [134]Federal Institute for Population Research BiB, Berlin, Germany. [135]Aalto University, Espoo, Finland. [136]Department of Economics, Indiana University, Bloomington, IN, USA. [137]Laboratoire d'Économie d'Orléans, Université d'Orléans, Orléans, France. [138]Anglo-American University

Prague, Prague, Czech Republic. [139]The University of Queensland, St Lucia, Queensland, Australia. [140]Karlstad University, Karlstad, Sweden. [141]FBK-IRVAPP, Trento, Italy. [142]Global Water Security Center, Tuscaloosa, AL, USA. [143]University of Western Australia, Perth, Western Australia, Australia. [144]Heriot-Watt University, Edinburgh, UK. [145]Howard University, Washington, DC, USA. [146]University of Kassel and INCHER, Kassel, Germany. [147]University College London, London, UK. [148]Chinese University of Hong Kong, Hong Kong, China. [149]Centre for Health Governance, University of Gothenburg, Gothenburg, Sweden. [150]HEPER School of Public Health and Community Medicine, University of Gothenburg, Gothenburg, Sweden. [151]Ragnar Frisch Centre for Economic Research, Oslo, Norway. [152]World Inequality Lab, Paris School of Economics, Paris, France. [153]University of Notre Dame, Notre Dame, IN, USA. [154]UC Center Sacramento, University of California Davis, Sacramento, CA, USA. [155]University of Western Ontario, London, Ontario, Canada. [156]Dresden University of Technology, Dresden, Germany. [157]University of Bath, Bath, UK. [158]University of Maryland, College Park, MD, USA. [159]Australian National University, Canberra, Australian Capital Territory, Australia. [160]New Economic School, Moscow, Russia. [161]Government of Canada, Ottawa, Ontario, Canada. [162]University of Pittsburgh, Pittsburgh, PA, USA. [163]University of Nevada Reno, Reno, NV, USA. [164]McGill University, Montreal, Quebec, Canada. [165]Knauss School of Business, University of San Diego, San Diego, CA, USA. [166]Norwegian University of Life Sciences, Ås, Norway. [167]University of Cagliari, Cagliari, Italy. [168]Freie Universität Berlin, Berlin, Germany. [169]University of Potsdam, Potsdam, Germany. [170]University of Nottingham, Nottingham, UK. [171]Finance Canada, Ottawa, Ontario, Canada. [172]London School of Economics and Political Science, London, UK. [173]Health Organisation, Policy and Economics Research Group, The University of Manchester, Manchester, UK. [174]Krea University, Sri City, India. [175]School of Politics and International Relations, University of Nottingham, Nottingham, UK. [176]University of Texas Rio Grande Valley, Edinburg, TX, USA. [177]Resources for the Future, Washington, DC, USA. [178]IVADO, Montreal, Quebec, Canada. [179]Mila, Montreal, Quebec, Canada. [180]University of Padova, Padova, Italy. [181]Vanderbilt University, Nashville, TN, USA. [182]University of Sussex, Brighton, UK. [183]Wake Forest University, Winston-Salem, NC, USA. [184]IWH Halle, Halle, Germany. [185]Princeton University, Princeton, USA. [186]University of Pennsylvania Wharton, Chicago, IL, USA. [187]Department of Health Management and Health Economics, University of Oslo, Oslo, Norway. [188]Humboldt University zu Berlin, Berlin, Germany. [189]Universitat de Girona, Girona, Spain. [190]University of Mannheim, Mannheim, Germany. [191]University of the Bundeswehr Munich, Neubiberg, Germany. [192]Department of Economics, Southern Methodist University, Dallas, TX, USA. [193]IZA, Bonn, Germany. [194]GLO, Essen, Germany. [195]Institute of Psychology, Osnabrück University, Osnabrück, Germany. [196]Asian Development Bank, Manila, Philippines. [197]University of Houston, Houston, TX, USA. [198]William and Mary Department of Government, Williamsburg, VA, USA. [199]Towson University, Towson, MD, USA. [200]McMaster University, Hamilton, Ontario, Canada. [201]National Economics University, Hanoi, Vietnam. [202]Emory University, Atlanta, GA, USA. [203]Norwegian School of Economics, Bergen, Norway. [204]Université de Montréal, Montreal, Quebec, Canada. [205]Ministry of Research and Higher Education Luxembourg, Luxembourg, Luxembourg. [206]Berlin School of Economics, Humboldt University of Berlin, Berlin, Germany. [207]University of Manitoba, Winnipeg, Manitoba, Canada. [208]Institute for Advanced Studies Vienna, Vienna, Austria. [209]Justus Liebig University Giessen, Giessen, Germany. [210]University of Chicago, Chicago, IL, USA. [211]GATE, Rotman School, University of Toronto, Toronto, Ontario, Canada. [212]Tufts University, Medford, MA, USA. [213]The Netherlands Court of Audit, The Hague, The Netherlands. [214]Institute for Employment Research IAB, Nuremberg, Germany. [215]Carleton College, Northfield, MN, USA. [216]Boston College, Chestnut Hill, MA, USA. [217]Minnesota Supercomputing Institute, Minneapolis, MN, USA. [218]Utrecht University, Utrecht, Netherlands. [219]New York University, New York, NY, USA. [220]Bucharest University of Economic Studies, Bucharest, Romania. [221]Ifo Institute, Ludwig Maximilian University of Munich, Munich, Germany. [222]INSAIT Sofia University, Sofia, Bulgaria. [223]Amazon, Seattle, WA, USA. [224]University of Gothenburg, Gothenburg, Sweden. [225]Department of Economics and Taiwan Social Resilience Research Center, National Taiwan University, Taipei, Taiwan. [226]The University of Manchester, Manchester, UK. [227]College of Staten Island-CUNY, Staten Island, NY, USA. [228]University of the West of England, Bristol, UK. [229]Stockholm Business School, Stockholm University, Stockholm, Sweden. [230]Leibniz Institute for Financial Research SAFE, Frankfurt, Germany. [231]KOF Institute, ETH Zurich, Zürich, Switzerland. [232]Colorado State University, Fort Collins, CO, USA. [233]Yale University, New Haven, CT, USA. [234]SGH Warsaw School of Economics, Warsaw, Poland. [235]Harvard University, Cambridge, MA, USA. [236]National University of Singapore, Singapore, Singapore. [237]Hitotsubashi University, Tokyo, Japan. [238]Community for Rigor, University of Pennsylvania, Philadelphia, PA, USA. [239]Sanford School of Public Policy, Duke University, Durham, NC, USA. [240]Swiss Finance Institute, University of Lausanne, Lausanne, Switzerland.

Methods

Our focus is on 12 journals. The journals are the following for economics: *American Economic Review*, *American Economic Review: Insights*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Journal: Macroeconomics*, *The Economic Journal*, *Journal of Political Economy*, *Quarterly Journal of Economics* and *Review of Economic Studies*. For political science, the journals are: *American Journal of Political Science*, *American Political Science Review* and *Journal of Politics*.

We have two streams to generate reproductions.

The I4R board

First, I4R has a board of editors who recommend potential reproducers. All board members are nominated by the lead author, A. Brodeur, who then reaches out to the board for suggestions of reproducers who could be a good fit for the studies in the targeted journals.

Replication games

Our second stream to generate reproductions and replications is the replication games. Games are one-day meet-ups open to faculty, postdocs, graduate students and other researchers. Participants join a small team of about three to five researchers all working in the same subfield (for example, development economics).

Types of reanalyses. We group reanalyses into eight groups: (1) alternative control variables; (2) change the sample; (3) change (coding of) the dependent variable; (4) change (coding of) the main independent variable; (5) change estimation method; (6) change inference method; (7) change weighting scheme; and (8) replication using new data.

Robustness for figures. While the bulk of our analysis compares coefficients and statistical significance from the original study and the work of reproducers, many results in papers are also displayed in figures. For those that are plots of coefficients (that is, event studies) we encouraged reproducers to provide the underlying statistics used to create the graph. This was often at the discretion of the reproducers: it could be taxing to write new code to compare and extract those values. In one example, the underlying programs which were written by the original authors were too complicated to modify with robustness checks. Excepting anecdotal examples, many teams found it feasible to reproduce a figure as part of a robustness check or direct replication. In those circumstances, we (A. Brodeur and D. Mikola) tried to subjectively describe whether we believed the results were the same. This was usually taken with the discussion of the reproducers and reading the original paper. We find that 189 out of 263 figures (71.9%) display the same result as the original paper and can be reasonably compared.

Non comparable reanalyses. As mentioned earlier, a direct comparison is not possible between the original analysis and the reproducers' analysis for about 15% of reanalyses. In applied microeconomics and politics papers, this may be owing to a change in the estimator or a change in the scale of the dependent or main independent variable. There are also scenarios where the original paper uses methods where coefficient estimates and *P* values are not the objective of the analysis. This is apparent in a few empirical macroeconomics papers teams looked at. A common 'robustness check' would be to adjust parameters that enter a model, possibly using accepted values in the field or estimated from an alternative dataset.

Eighty-two articles have at least one non-comparable estimate. Only a small proportion (ten reanalyses) were not directly comparable for all reported reanalysis estimates. For reanalyses that were not directly comparable, we report the proportion that reproducers indicated were of the same statistical significance as the original and same valence or direction. For our four definitions of reproducibility and replication rates these are: when the original estimate is statistically significant at the 5% level, 85% of those we considered not directly comparable indicated their reanalysis was of the same significance (93% for the 10% level). When the original estimate was not statistically significant at the 5% level, 88% of those we considered not directly comparable indicated their reanalysis was of the same (non)significance (92% for the 10% level).

Study selection. Not all studies from our targeted journals have been reproduced or replicated. Our approach leads to an over-representation of studies using publicly available data. Another feature of our sample is that the targeted journals have a data availability policy and enforce it. This is in contrast to many top journals in economics and political science. Our sample should thus be viewed as very selected both in terms of impact and high data and code availability rates. In fact, approximately 45% of replication packages in our sample included raw data and complete cleaning code. An additional 13.5% provided partial cleaning code.

Journal policy. The *American Journal of Political Science* does not have a data editor. Instead, computational reproducibility checks are carried out by staff at the Odum Institute for Research in Social Science, at the University of North Carolina, Chapel Hill. The journals that do not conduct reproducibility checks are the *American Political Science Review*, the *Journal of Political Economy* and the *Quarterly Journal of Economics*. The other journals conduct computational reproducibility checks internally.

Data editors ensure that replication packages include data and codes, and that the documentation (for example, a readme file) is complete. In the event that the authors cannot share some or all the data, they request that information is provided on how other researchers could obtain the datasets. Their teams also run the codes and ensure that the output is similar to what is reported in the article. They do not look for coding errors or run robustness checks.

Many-analysts approach

Our approach and research questions, which we detail below, were pre-registered. Our preanalysis plan was pre-registered at the Open Science Framework (https://osf.io/8wsqx/). The preanalysis plan was pre-registered prior to sharing the meta database with analysts. The Supplementary Information contains more information on the meta database.

The six analyst teams tackled the following eight questions:

1. Does reproducibility/replicability rate depend on replicators' experience coding?
2. Does reproducibility/replicability rate depend on replicators' academic experience?
3. Does reproducibility/replicability rate depend on the authors' experience?
4. Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular:
   (a) Are reproducibility/replicability rate higher when authors' experience is high, and replicators' experience is low (in comparison to similar levels)?
   (b) Are reproducibility/replicability rate higher when authors' experience and replicators' experience is similar (in comparison to dissimilar levels)?
   (c) Are reproducibility/replicability rate higher when authors' experience is low, and replicators' experience is high (in comparison to similar levels)?
5. Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular:
   (a) Are reproducibility/replicability rate higher when authors' have high prestige, and replicators' experience have low prestige (in comparison to similar levels)?

(b) Are reproducibility/replicability rate higher when authors' and replicators' prestige is similar (in comparison to dissimilar levels)?
(c) Are reproducibility/replicability rate higher when authors' have low prestige, and replicators' experience have high prestige (in comparison to similar levels)?
6. Does reproducibility/replicability rate depend on the original authors providing raw data?
7. Does reproducibility/replicability rate depend on the original authors providing raw or intermediate data?
8. Does reproducibility/replicability rate depend on the original authors providing cleaning code?

Data for analysts. Analysts were not given access to raw data (database, team leader surveys and individual surveys). Rather, they were given access to intermediate or analytical data that were cleaned and merged in a manner that was consistent with their analysis. Giving researchers a downstream dataset allowed A. Brodeur and D. Mikola to place restrictions on what the analysts could do. The clearest example of this would be defining dependent variables that were not allowed to be changed—providing a consistent definition between analysts. Asking certain research questions also restricted the data given to the analysts. These restrictions were done in ways so that any analysis done would be more comparable.

The backbone of the data provided to analysts was the meta database, to which questions from the team leader surveys and individual surveys were added. Much of the information from the individual surveys was aggregated to the report level.

The data given to the analysts changed as reproduction reports, team leader and individual surveys were completed. In total, we provided 13 updated databases for analysts between 6 November 2023 and 12 February 2024. We did this to give analysts time to create scripts which would work with partial datasets as we worked to gather reports and surveys. This allowed analysts to expedite their analysis once the full dataset was constructed.

The goal was to have each team answer each research question independently. Each team received the same instructions and data. We allowed full flexibility to all teams. Teams were allowed to use any statistics package, statistical model, inference, weighting scheme and so on. Teams were free to choose the independent variables and how to code them. Teams were also free to construct their own derived variables from the dataset given to them.

We provided the four dependent variables and the database to all teams. They were allowed to use any of the provided variables and new data. The only restriction imposed on teams is that they needed to use our four main dependent variables.

Team construction. We asked a subset of coauthors on this paper (reproducers) if they would like to help to analyse our database. We informed them that we would "have different teams independently working together at answering the same research questions (for example, what is the reproducibility/replicability rate for each specific type of robustness checks/recoding)." The subset of coauthors who received an invitation to volunteer were: (1) contacted between 21 September and 8 October 2023; and (2) had completed, or were near completion of, their reproduction report. We sent invitations (a simple sign-up form) in an email which also asked the reproducers to respond to individual and team leader surveys which formed parts of our previous analysis. About 110 co-authors were invited between 21 September and 8 October. Ten individuals ultimately signed-up as 'many-analysts'.

In our request for volunteers, we asked volunteers if they: (1) had a team who wanted to do research on the project; (2) wanted to be added to a team; or (3) wanted to work on the analysis alone. No one joined as teams, most people wanted to be added to a team, and the remainder wanted to work alone. For those that wanted to work together, we assembled teams as best we could so they were close enough in time zones. We had two teams of three, one team of two, and two individuals. A. Brodeur and D.M. also acted as a team of two, yielding six teams in total. No members of any teams left during the analysts phase.

Although the principal investigator (A. Brodeur) ultimately provided each volunteer with a payment of 3,000 Canadian dollars, this compensation was not disclosed or anticipated at the time they agreed to participate.

Sample composition of the database. The database described above provides 6,583 reanalysed test statistics from 103 reproduction reports (seven reports did not include robustness checks). The other test statistics are estimates obtained by recoding the analysis.

Supplementary Table 11 provides summary statistics for the full sample and by journal. In total, 83 reproduction reports were completed through games compared with 27 through the editorial board stream. Seventy-nine reproduction reports are for the field of economics versus 31 for political science.

There is no universally agreed upon criterion for reproduction. As a first criterion, we follow much of the literature and define reproducibility as obtaining a statistically significant effect in the same direction (positive or negative) as the original study. Throughout, we rely on four main dependent variables:

- First dependent variable: dummy variable indicating whether the reanalysis is statistically significant at 5% level and same valence. For this dependent variable, we only keep original estimates statistically significant at the 5% level.
- Second dependent variable: dummy variable indicating whether the reanalysis is statistically significant at 10% level and same valence. For this dependent variable, we only keep original estimates statistically significant at the 10% level.
- Third dependent variable: dummy variable indicating whether the reanalysis remains not statistically significant at 5% level. For this dependent variable, we only keep original estimates statistically insignificant at the 5% level.
- Fourth dependent variable: dummy variable indicating whether the reanalysis remains not statistically significant at 10% level. For this dependent variable, we only keep original estimates statistically insignificant at the 10% level.

The average number of reanalysed test statistics per article is about 60. The s.d. is very high (73), with a maximum of 421. This is unsurprising given that some teams, for instance, focused most of their attention to (blindly) recoding using the raw data (either provided by the authors or re-downloaded by the reproducers), whereas other teams focused solely on conducting robustness checks for multiple central hypotheses. As an illustrative example, imagine that an original article has three main outcome variables and relies on two main specifications. If the reproducers conduct 5 different robustness checks for each outcome variable and specification, then this would lead to 30 reanalysed test statistics.

As a robustness check, we deal with this issue by adjusting the weight of each test statistics by the inverse number of such statistics in the reproduction report such that each reproduction report has the same weight.

Supplementary Table 2 provides descriptive statistics. The articles in our sample are all recently published with a relatively small number of Google Scholar citations (44 on average) as of the completion of a reproduction report. The original authors are more experienced than reproducers with 11 years of experience (that is, years since completing their PhD) against 3. Original authors have on average 4,269 Google Scholar citations in comparison to 478 for reproducers. Those differences are mostly driven by the larger share of graduate students among reproducers than for original authors (49% against 6%). There are about 2.6 original authors per article in comparison to 3.2 for reproducers.

About 15% of reproducers have recently published in a top 5 or one of the three leading political science journals in our sample. Approximately 30% have published in those journals or in one of the other journals in our sample.

Although reproducers have less academic experience than original authors on average, their level of expertise as programmers is quite advanced. About 10%, 48% and 33% of reproducers report that their level of expertise is 'expert', 'proficient' and 'competent,' respectively. Moreover, about 55% of reproducers had already produced a replication package for their own work or journal publication.

Computational reproducibility. We rely on the Social Science Reproduction Platform (SSRP) ten-point scale to document computational reproducibility. This scale is useful, as it is standardized and offers more details than a simple indicator for whether the results are computationally reproducible (https://bitss.github.io/ACRE/assessment.html#score provides more details on SSRP and this scale). On this scale, a rating of 1 signifies the incapacity to reproduce results due to the absence of data or code, and a rating of 10 indicates the capability to faithfully reproduce results from the raw data (unaltered files obtained by the authors from the sources cited in the paper) to the final numerical results as published in the paper.

The following is a direct reproduction from the Guide for Accelerating Computational Reproducibility in the Social Sciences.

- Level 1 (L1): No data or code are available. Possible improvements include adding: raw data, analysis data, cleaning code and analysis code.
- Level 2 (L2): Code scripts are available (partial or complete), but no data are available. Possible improvements include adding: raw data and analysis data.
- Level 3 (L3): Analytic data and code are partially available, but raw data and cleaning code are missing. Possible improvements include: completing analysis data and/or code, adding raw data and adding analysis code.
- Level 4 (L4): All analytic datasets and analysis code are available, but the code fails to run or produces results inconsistent with the paper (not computationally reproducible from analytic data (CRA)). Possible improvements include: debugging the analysis code or obtaining raw data.
- Level 5 (L5): Analytic datasets and analysis code are available and they produce the same results as presented in the paper (CRA). The reproducibility package may be improved by obtaining the original raw data. Note: This is the highest level that most published research papers can attain currently. Computational reproducibility from raw data is required for papers that are reproducible at Level 6 and above.
- Level 6 (L6): Cleaning code scripts are available (partial or complete), but raw data are missing. Possible improvements include: adding raw data.
- Level 7 (L7): Cleaning code is available and complete and raw data are partially available. Possible improvements: adding raw data.
- Level 8 (L8): All the materials (raw data, analytic data, cleaning code and analysis code) are available. However, the cleaning code fails to run or produces different results from those presented in the paper (not computationally reproducible from raw data (CRR)) or the analysis code fails to run or produces results inconsistent with the paper (not CRA). Possible improvements: debugging the cleaning or analysis code.
- Level 9 (L9): All the materials (raw data, analytic data, cleaning code and analysis code) are available. The analysis code produces the same output as presented in the paper (CRA). However, the cleaning code fails to run or produces different results from those presented in the paper (not CRR). Possible improvements: debugging the cleaning code.
- Level 10 (L10): All necessary materials are available and produce consistent results with those presented in the paper. The reproduction involves minimal effort and can be conducted starting from the analytic data (CRA) and the raw data (CRR). Note that level 10 is aspirational and may be unattainable for most research published today.

Each team was asked to assign a reproducibility score on a scale of one to ten to the paper reproduced. This involved documenting the completeness of the data and code, and whether the materials produce results consistent with those in the article. Their focus for computational reproducibility is only for the claims that they have investigated rather than all exhibits in the article.

The results are presented in Extended Data Fig. 1. This figure shows the variation across papers, with the highest concentration of scores concentrated at levels 10 and 5. Indeed, over 85% (Levels 5 and 10) of results examined in our sample were fully reproducible using either: (1) the raw and analytical data, or; (2) the analytical data when the raw data were not provided. Level 10 means that all necessary materials are available and produce consistent results with those presented in the paper. Level 5 means that analytic datasets and analysis code are available, and they produce the same results as presented in the paper. In other words, level 5 indicates that the reproducers successfully (computationally) reproduced the numerical results using the analytical data, but the raw data were not provided, whereas level 10 indicates that the reproducers successfully (computationally) reproduced the numerical results using the raw data and cleaning and analytical codes.

The remaining 15% includes studies for which analytic code and data are partially available and studies for which some of the codes (cleaning or analytic) fail to run or produce results inconsistent with the paper. These findings suggest very high rates of computationally reproducible results.

Our results are in stark contrast with several studies documenting low computational reproducibility rates[13,19,40]. This is perhaps unsurprising given that most of the articles in our sample were already computationally reproduced by data editors. This highlights that the open science movement has improved computational reproducibility of research findings in leading economics and political science journals. Our approach is also different as we are targeting newer studies and only articles for which (at least) analytical data were available to the teams of reproducers. A more comparable (and recent) study is ref. 37, which assess the reproducibility of nearly 500 articles published in the journal *Management Science*. They find that more than 95% of articles could be reproduced if data accessibility and software requirements were not an obstacle for reproducers.

Recoding. We now turn to recoding exercises conducted by a subset of teams. Those teams either recoded using a different software language or used the same software without looking at the original authors' code. In total, 19 teams of reproducers engaged in computationally reproducing and checking for coding errors using a different statistical software than the original authors. This may be due to reproducers being more comfortable in another software language or the availability of specific commands (to run a robustness check). Five teams also recoded the empirical analysis without looking at the authors' code/programs.

Recoding in a different software opens up the ability for others to benefit and understand the empirical foundations of published articles in ways that the original authors may not have been able to convey. For instance, verifying reproducibility by translating it into R or Python makes the study itself accessible to many more researchers.

Recoding also helps to assess the importance of differing assumptions embedded within programming languages (for example, different types of random number generations, rounding rules and numerical precision). We categorized recoding exercises done by reproducers into three categories: (1) identical numerical results; (2) minor differences; and (3) major differences. Minor differences involve small numerical discrepancies between the authors' estimates and those obtained by the reproducers. Those differences do not lead to important changes

in significance or magnitude. By contrast, major differences lead to major differences in one or multiple claims.

Coding errors and discrepancies. We now turn to documenting the prevalence of coding errors and discrepancies between the code and the published article. Of note, a paper might be fully reproducible, but the programs may contain coding errors. Similarly, there might be important discrepancies between what the article states and what the programs compute, while remaining computationally reproducible.

We do not document trivial coding errors such as versioning issues and missing packages/paths. Those coding errors are typically easily fixed by the reproducers. We instead focus on coding errors which could have had an impact on claims and conclusions of articles.

We uncover minor or major coding errors in 26 of the 110 studies in our sample, with some studies containing multiple errors. The errors can be broadly categorized into errors of the dependent variable (4 articles), main independent variable (5), control variables (10), estimation (2), inference (2), sample/observations (8) and other (5). While not all coding errors lead to changes in the conclusions of the original study, we uncovered several major coding errors worth discussing. Some examples of major errors include: a very large number of duplicated observations, failing to fully interact a difference-in-differences regression specification, miscoding the treatment variable for a large number of (or all) observations, and clear model misspecification.

The prevalence of coding errors is larger for economics (26%) than political science (16%). A plausible explanation is that replication packages from economic articles have more lines of code than those in political science, mechanically increasing the likelihood of at least one coding error.

We also uncovered transcription issues for 13 studies, typically involving small numerical differences or rounding errors that do not affect the claims or conclusions of the article.

Time trends in data and code availability. To document time trends in data and code availability in economics and political science between 2014 and 2023, we randomly sampled 10 empirical articles per year for each of our 12 target journals. We define an article as empirical if it relies on real or simulated data at any point in the text. Thus, a theoretical article that is motivated with a descriptive analysis of labour market trends, or an econometric paper showing properties of an estimator on synthetic data would both be classified as empirical for the purposes of our study.

To randomly select papers, we proceeded as follows. First, we noted the number of issues per journal per year. Second, we drew ten issues (with replacement) for each year. Third, for each issue, we generated a random permutation of numbers between 1 and 35, giving us the order in which papers from a given issue should be considered. So, for example, if the first issue drawn was 4, and the first number in our permutation sequence was 10, we would consider the tenth article in the fourth issue for coding. We skipped an article and proceeded with the next number in the permutation if the article in question: (1) was not empirical; (2) was not a standard article (we excluded comments, replies and corrections, retraction notices and editor notes, even if they were empirical in nature); (3) was a duplicate that had already been considered (for example, issue number one, article number five, was drawn twice in a row); or (4) did not exist (our chosen journals typically publish around ten articles per issue, so higher numbers in the permutation often went unused).

In our coding we considered whether the journal website or the article pdf contain a link to a replication package, whether this package is accessible, and what the contents of the package are. We tracked the availability of a readme file, cleaning and analytical code, and raw, intermediate and final data. Note that our coding of code availability is optimistic in the sense that we only note whether a particular type of code exists; we did not verify its completeness or correctness. However, when authors explicitly indicated that a code was incomplete, we noted this information.

Of note, the *American Economic Review: Insights* only formally became a journal in 2019. For the previous five years, we did not collect data for this journal, leading to ten fewer papers per year.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data are available on Zenodo (https://zenodo.org/records/17792605 (ref. 41)) and OSF (https://osf.io/8wsqx/). See OSF for our preanalysis plan.

Code availability

Code is available on Zenodo (https://zenodo.org/records/17792605 (ref. 41)) and OSF (https://osf.io/8wsqx/).

40. Wood, B. D., Müller, R. & Brown, A. N. Push button replication: is impact evaluation evidence for international development verifiable?. *PLoS ONE* **13**, e0209416 (2018).
41. Brodeur, A. Replication package for "Computational reproducibility and robustness of empirical economics and political science research between 2022 and 2023" [Data set]. *Zenodo* https://doi.org/10.5281/zenodo.17792605 (2025).

Acknowledgements We acknowledge support from Coefficient Giving and the Social Sciences and Humanities Research Council. Any views expressed herein are the authors' personal opinions and not those of Ontario Public Service. The work by J.D.G. was not undertaken under the auspices of the Ontario Public Service as part of his employment responsibilities. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. The findings, interpretations, and conclusions expressed in this work are entirely those of the authors and do not necessarily reflect the views of the World Bank or its Board of Directors. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Banco de España or the Eurosystem. All remaining errors are the authors' responsibility.

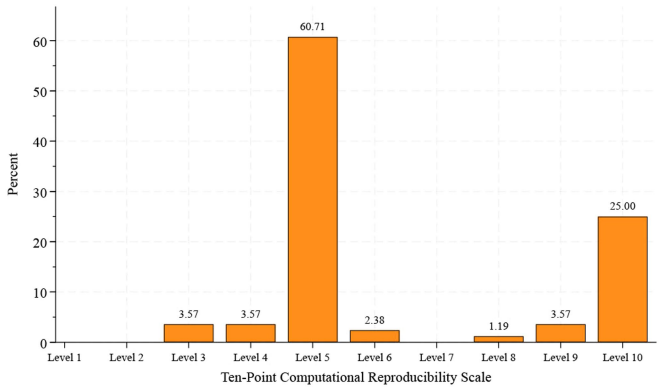
Author contributions Preparation of tables, figures and manuscript: A. Brodeur, N. Cook, D. Mikola and L. Fiala. Conception or design of the work: J. Ankel-Peters, A. Brodeur, M. Connolly, N. Cook, A. Dreber, F. Hoces de la Guardia, M. Johannesson, E. Miguel, D. Mikola and L. Vilhuber. Analysis or interpretation of the reproducibility data: T. Brailey, R. Briggs, A. Brodeur, N. Cook, A. de Gendre, Y. Dupraz, J. Gabani, R. Gauriot, G. Lima and D. Mikola. Analysis or interpretation of data and generating data and conception of a reproduction: D. Campbell, N. Cook, J. Haddad, L. Kattan, D. M. Fages, F. Mierisch, P. Sun, T. Wright, A. Abarca, M. Acharya, S. S. Adjisse, A. Akhtar, E. A. Ramirez Lizardi, S. Albrecht, S. Nygaard Andersen, Z. Andlib, F. Arrora, T. Ash, E. Bacher, S. Bachler, F. Bacon, M. Bagues, T. Balogh, A. Batmanov, M. Barschkett, B. K. Basdil, J. Baxa, S. Becker, M. Beeder, L.-P. Beland, A.-H. Bello, D. B. Markovits, G. Benjamin, T. Bergeron, M. P. Blimpo, M. Binetti, C. Bonander, J. Bonneau, E. Borbáth, N. Borgen, S. T. Borgen, J. Borowsky, T. Brailey, R. Briggs, E. Brini, M. Brown, M. Brun, S. Bruns, N. Buliskeria, A. Calef, A. Cameron, P. Campa, S. Campos-Rodríguez, G. G. Cantone, F. Carpena, P. Carter, P. Castañeda Dower, O. Castek, J. Caviglia-Harris, G. C. Strand, S. Chen, S. I. Chzhen, J. Chung, J. Collins, A. Coppock, H. Cordeau, B. Couillard, J. Crechet, L. Crippa, J. Cui, C. Czymara, H. Daarstad, D. C. Dao, D. Dao, M. D. Schmandt, A. de Linde, L. De Melo, L. Deer, A. de Gendre, M. De Vera, V. Dimitrova, J. F. Dollbaum, J. M. Dollbaum, M. Donnelly, L. D. Toan Huynh, T. Dumbalska, J. Duncan, K. T. Duong, Y. Dupraz, T. Duprey, C. Dworschak, S. Ellingsrud, A. Elminejad, Y. Eissa, A. Erhart, G. Etingin-Frati, E. Fatemipour, A. Federice, J. Feld, G. Fenig, L. Fiala, M. Firouzjaeiangalougah, E. Fleisje, A. Fortier-Chouinard, J. F. Engel, N. Fréchet, R. Fortier, T. Fries, M. J. Frith, J. Gabani, T. Galipeau, S. Gallegos, A. Gangji, X. Gao, C. Garnache, A. Gáspár, R. Gauriot, E. Gavrilova, A. Ghosh, G. Gibney, G. Gibson, G. Godager, L. Goff, D. Gong, J. González, J. D. Gretton, C. Griffa, I. Grigoryeva, M. Grøtting, E. Guntermann, J. Guo, A. Gugushvili, H. Habibnia, S. Häffner, J. D. Hall, O. Hammar, A. H. Kordt, B. Hashimoto, J. S. Hartley, C. I. Hausladen, T. Havránek, H. He, M. Hepplewhite, M. Herrera-Rodriguez, F. Heuer, A. Heyes, A. T. Y. Ho, J. Holmes, A. Holzknecht, Y.-H. D. Hsu, S.-H. Hu, Y.-S. Huang, M. Huebener, C. Huber, K. P. Huynh, Z. Irsova, O. Isler, N. Jakobsson, R. Jananji, T. A. Jayalath, M. Jetter, J. John, R. J. Forshaw, F. Juan, V. Kadriu, S. Karim, E. Kelly, D. K. H. Dang, T. Khushboo, J. Kim, G. Kjellsson, A. Kjelsrud, J. Korpershoek, A. Kotsadam, L. Krashinsky, S. Kundu, A. Kustov, N. Lalayev, A. Langlois, J. Laufer, B. Lee-Whiting, A. Leibing, G. Lenz, J. Levin, P. Li, T. Li, Y. Lin, G. Lima, A. Listo, D. Liu, X. Lu, E. Lukmanova, A. Luscombe, L. R. Lusher, K. Lyu, H. Ma, N. Mäder, C. Makate, A. Malmberg, A. Maitra, M. Mandas, J. Marcus, S. Margaryan, L. Márk, D. M. Fages, A. Martignano, A. Marsh, I. Masetto, A. McCanny, E. McManus, R. McWay, L. Metson, F. Mierisch, J. M. Kinge, S. Mishra, M. Mohnen, J. Möller, R. Montambeault, S. Montpetit, L.-P. Morin, T. Morris, S. Moser, F. Y. S. Motoki, L. Muehlenbachs, A. Musulan, M. Musumeci, M. Nabin, K. Nchare, F. Neubauer, Q. M. P. Nguyen, T. Nguyen, V. Nguyen-Tien, A. Niazi,

G. Nikolaishvili, A. Nordstrom, P. Nüß, A. Odermatt, M. Olson, H. Øien, T. Ölkers, M. Oliver i Vert, E. Oral, C. Oswald, A. Ousman, Ö. Özak, S. Pandey, A. Pavlov, M. Pelli, R. Penheiro, R. Park, E. Pérez Martel, J. Ankel-Peters, T. Petrovičová, L. Phan, A. Prettyman, J. Procházka, A. Putri, J. Quandt, K. Qiu, L. Q. T. Nguyen, A. Rahman, C. H. Rea, A. Reiremo, L. Renée, J. Richardson, N. Rivers, B. Rodrigues, W. Roelofs, T. Roemer, O. Rogeberg, J. Rose, A. Roskos-Ewoldsen, P. Rosmer, B. Sabada, S. Saberian, N. Salamanca, G. Sator, D. Scates, E. Schlüter, C. Sells, S. Sen, R. Sethi, A. Shcherbiak, M. Sogaolu, M. Soosalu, E. Ø. Sørensen, M. Sovani, N. Spencer, S. Staubli, R. Stans, A. Stewart, F. Stips, K. Stockley, S. Strobel, E. Struby, J. Tang, I. Tanrisever, T. T. Yang, I. Tastan, D. Tatić, B. Tatlow, F. T. Seuyong, R. Thériault, V. Thivierge, W. Tian, F.-M. Toma, M. Totarelli, V.-A. Tran, H. Truong, N. Tsoy, K. Tuzcuoglu, D. Ubfal, L. Villalobos, J. Walterskirchen, J. T. Wang, V. Wattal, M. D. Webb, B. Weber, R. Weisser, W.-C. Weng, C. Westheide, K. White, J. Winter, T. Wochner, M. Woerman, J. Wong, R. Woodard, M. Wroński, G. C. Yang, M. Yazbeck, L. Yap, K. Yassin, H. Ye, J. Y. Yoon, C. Yurris, T. Zahra, M. Zaneva, A. Zayat, J. Zhang, Z. Zhao and Y. Zhong. Computational reproducibility: A. Brodeur, J. Haddad and P. Sun. Local organizer replication games: M. Connolly, R. Gauriot, L. Goff, C. Huber, A. Kotsadam and D. M. Fages.

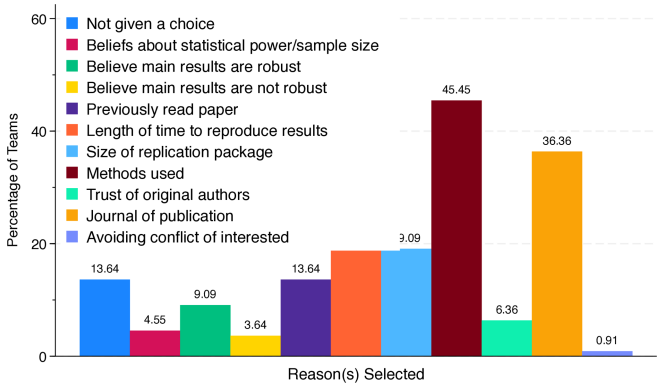
Competing interests The authors declare no competing interests.

Additional information

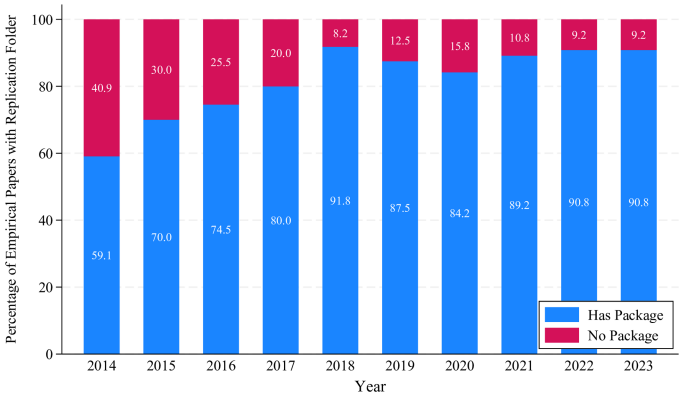
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-026-10251-x.
Peer review information *Nature* thanks Colin Camerer who co-reviewed with Anastasia Buyalskaya; T. D. Stanley; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.



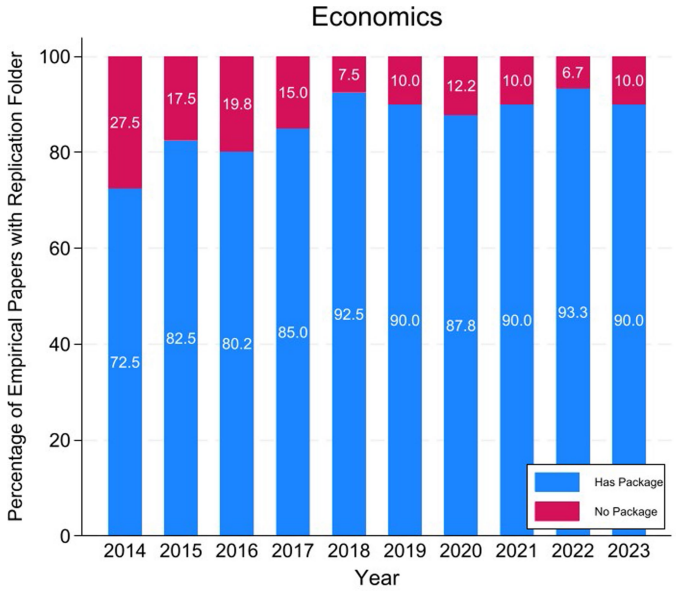
Extended Data Fig. 1 | Ten-point computationally reproducibility score. Each team assigned a reproducibility score on a scale of one to ten to the paper reproduced. See Supplementary Materials for a description of each score. Level 10 (L10) means that all necessary materials are available and produce consistent results with those presented in the paper, while level 5 (L5) means that analytic data sets and analysis code are available and they produce the same results as presented in the paper.

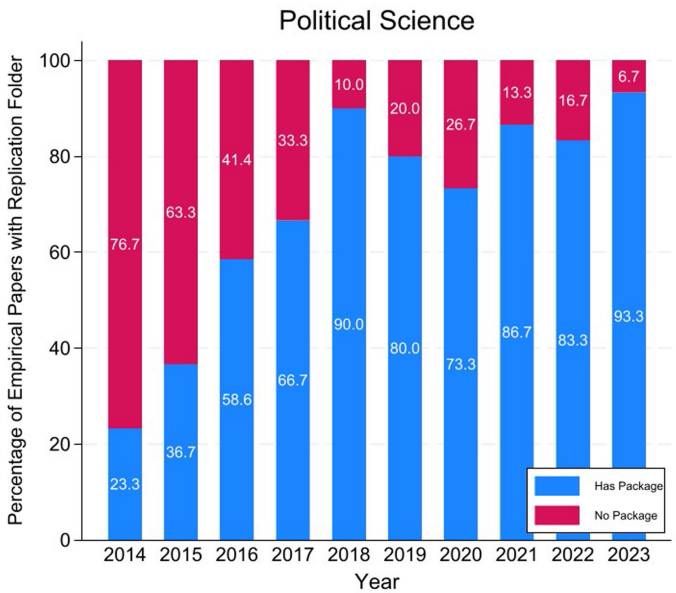


Extended Data Fig. 2 | Reasons for selecting paper? (Select all which apply). Data collected *via* survey of our reproducers after completing their reports. This figure illustrates the responses to the question: 'For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided?

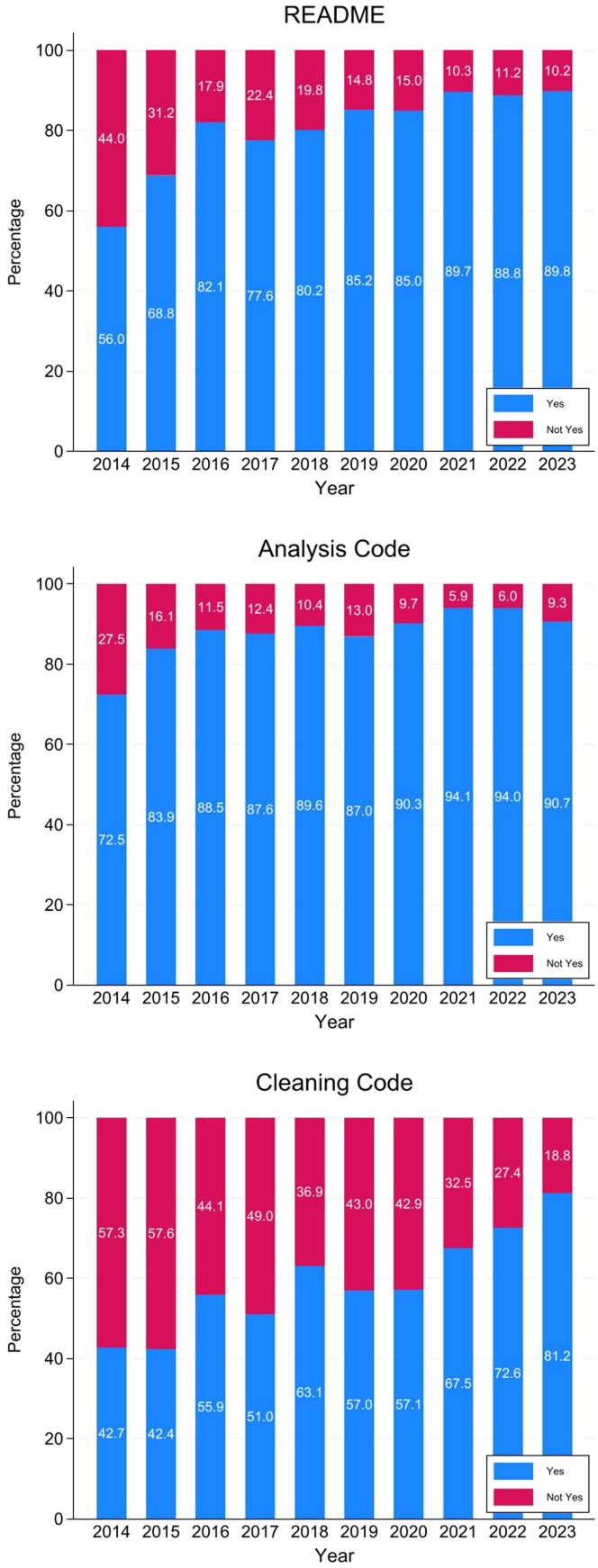


Extended Data Fig. 3 | Percentage of papers with a replication folder. The total sample is 1150 papers with 120 papers per year from 2019 to 2023 and 110 papers per year from 2018 to 2014. Each journal has 10 papers per year except *American Economic Review: Insights* which only formally became a journal in 2019 (and are omitted in earlier years). The journals sampled over correspond to those used in the manuscript's main analysis, three from political science and nine from economics. The political science journals include: *American Journal of Political Science*, *American Political Science Review*, and *Journal of Politics*. The economics journals include: *American Economic Review*, *Quarterly Journal of Economics*, *Review of Economic Studies*, *Journal of Political Economy*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Review: Insights*, *Economic Journal*.

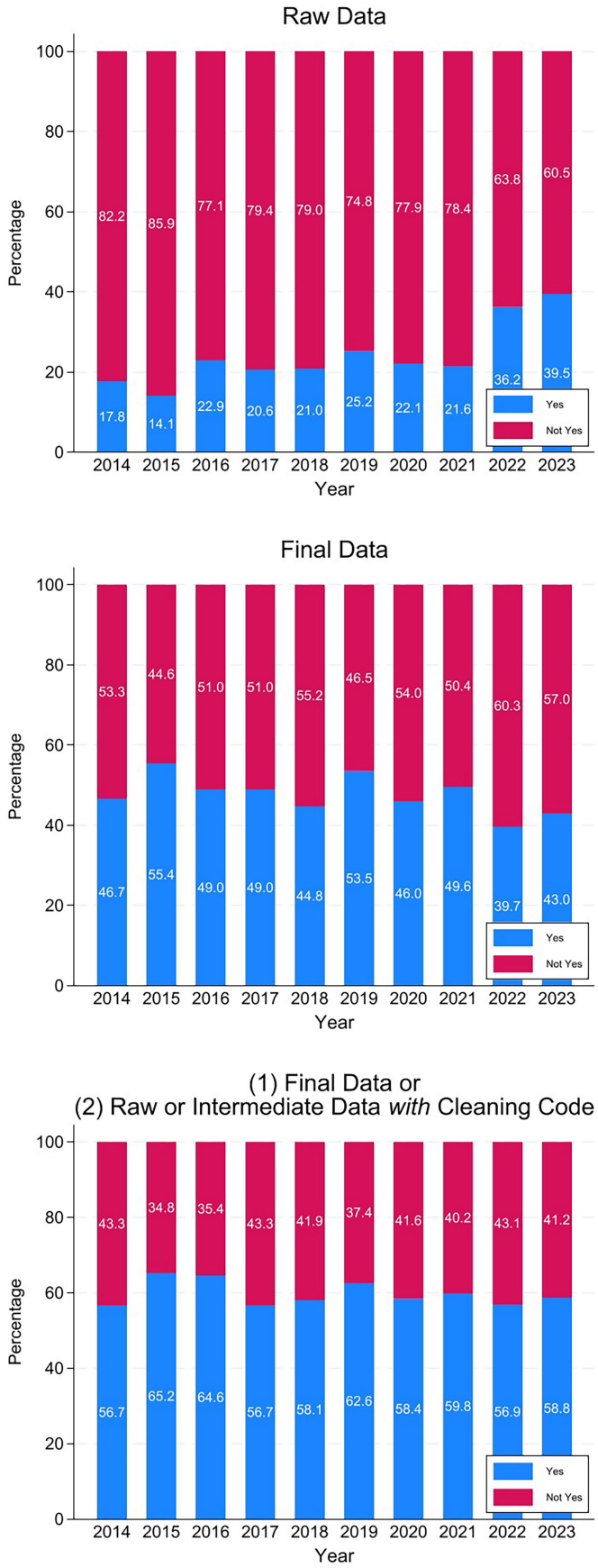




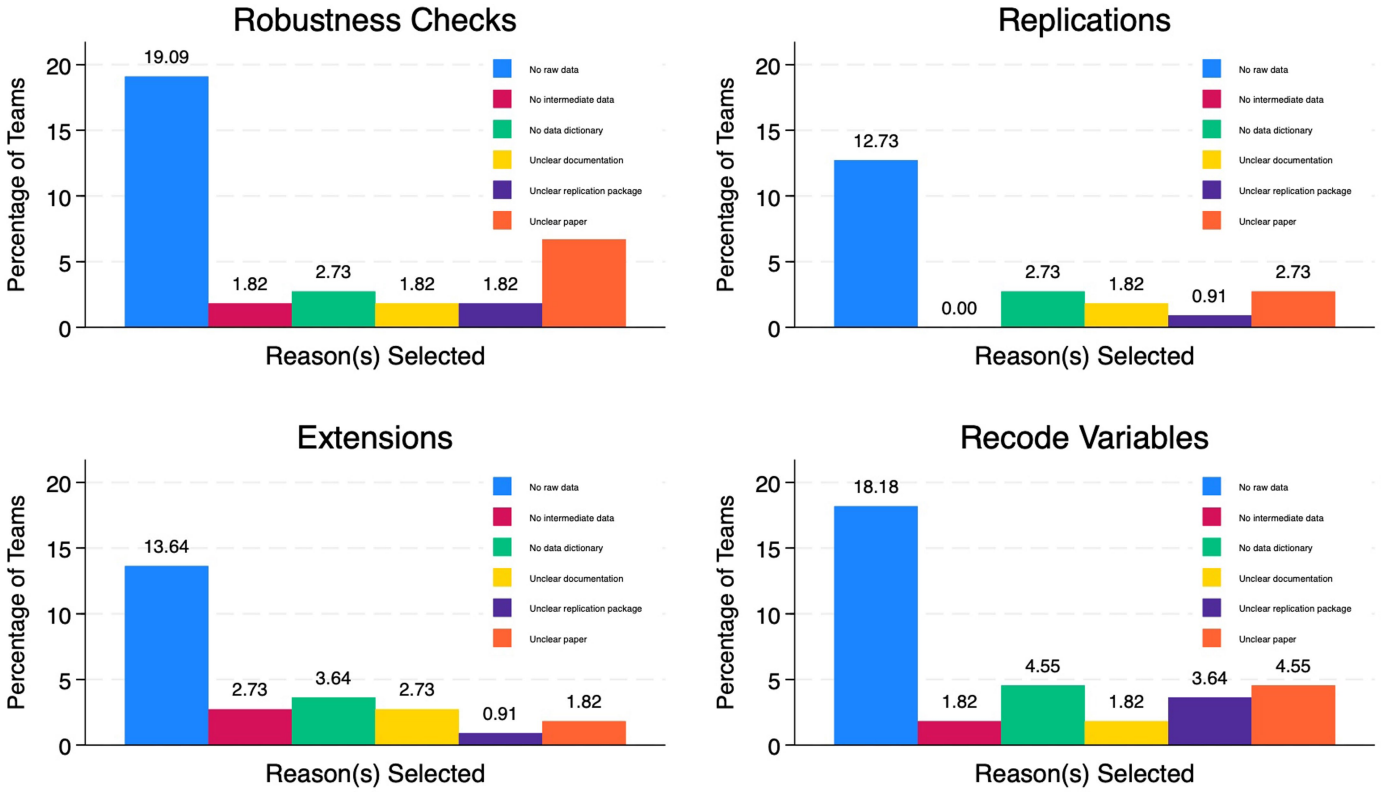
Extended Data Fig. 4 | Percentage of papers with a replication folder by discipline. Panel (**a**) is for papers published in economics journals where Panel (**b**) is for papers published in political science. The total sample is the same as Extended Data Fig. 3 is 1150 papers, where 850 papers are in the economics sample and 300 papers are in the political science sample.



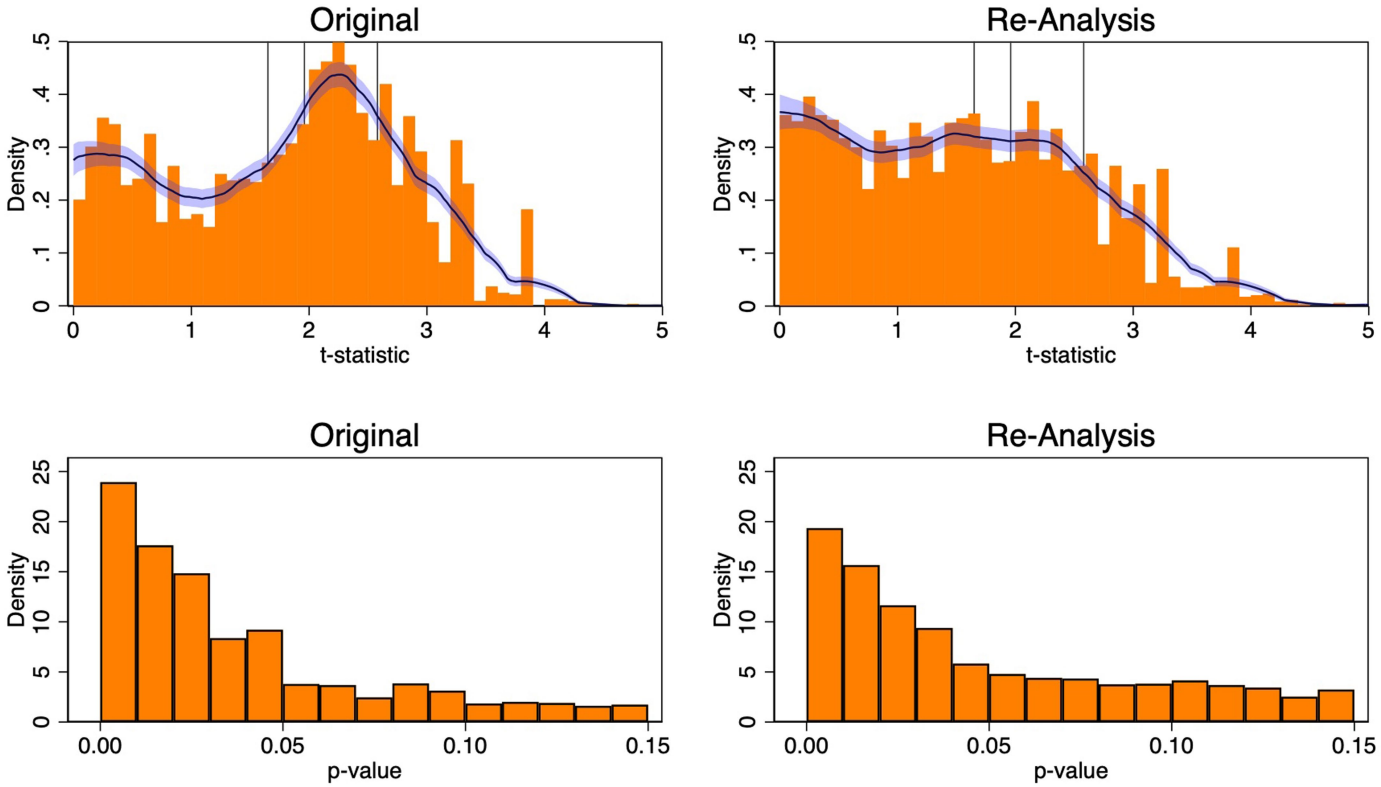
Extended Data Fig. 5 | Percentage replication folders' with contents conditional on they should have a replication folder. Each subfigure represents the proportion of the replication folders which affirmatively ('Yes') contained the variable (displayed as the title). The 'Not Yes' in the legend corresponds to those replication folders which did not affirm ('No') or had only 'Some' of the required contents. Each sample is over those observations where categories are applicable (*i.e.* not all replication packages require the same contents).



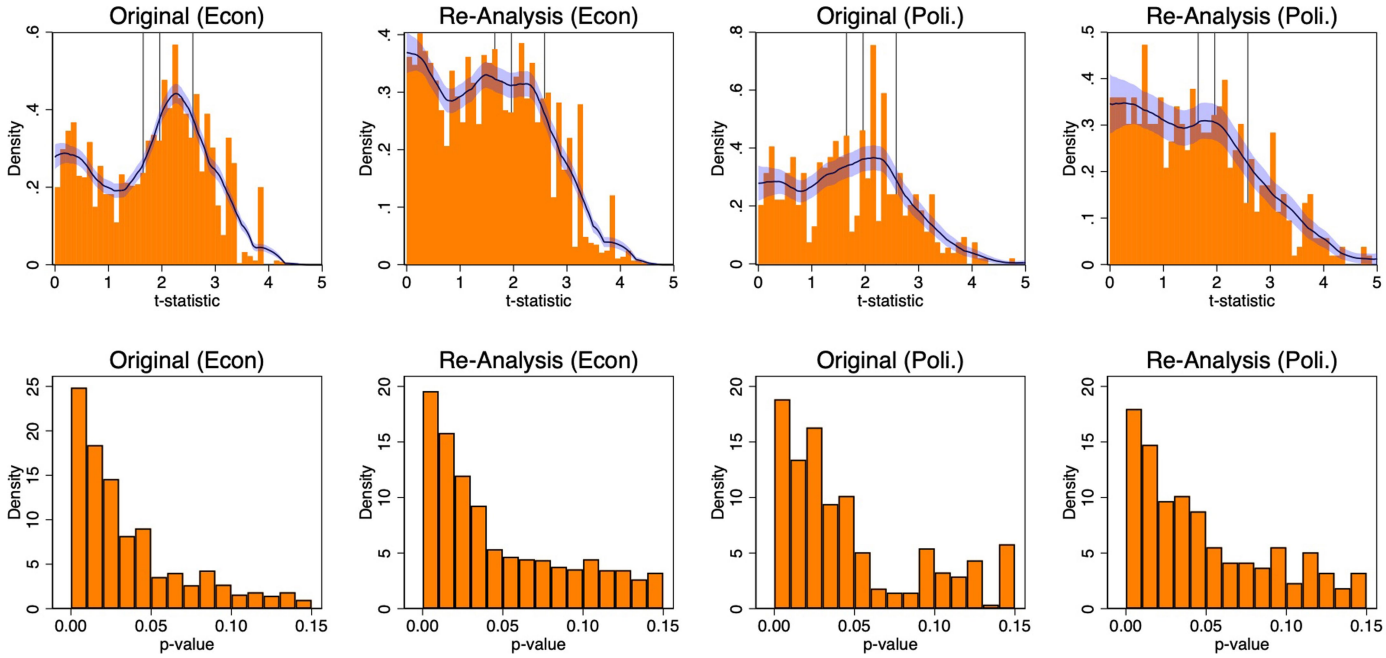
Extended Data Fig. 6 | Percentage replication folders' with contents conditional on they should have a replication folder. Each subfigure represents the proportion of the replication folders which affirmatively ('Yes') contained the variable (displayed as the title). The 'Not Yes' in the legend corresponds to those replication folders which did not affirm ('No') or had only 'Some' of the required contents. Each sample is over those observations where categories are applicable (*i.e.* not all replication packages require the same contents).



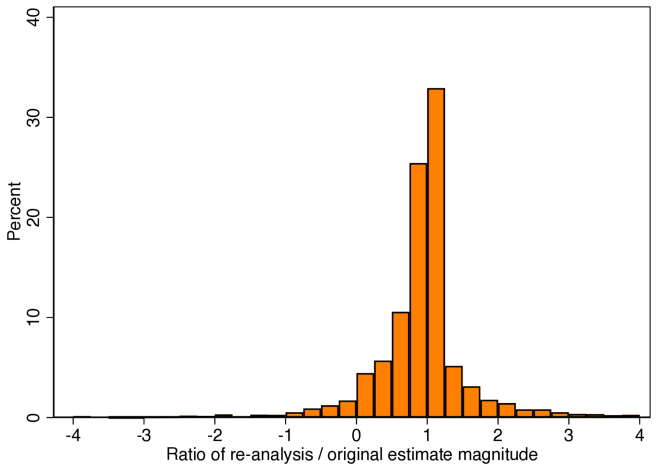
Extended Data Fig. 7 | Reasons unable to conduct robustness checks. This Figure illustrates the share of teams who were unable to perform robustness checks (top-left), replications (top-right), key variable recodes (bottom-right) or extensions (bottom-left) for various reasons represented by the different coloured bars.



Extended Data Fig. 8 | Distributions of *t*-statistics for original studies and re-analyses. The top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left panel includes all original studies in our data set. The top right panel includes all re-analysis estimates in our data set. Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel (which includes renormalization at 0). The bottom figures display histograms of test statistics for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025, among original studies and those from re-analyses, respectively.



Extended Data Fig. 9 | Distributions of *t*-statistics and *P*-values by field. We restrict the sample to articles published in the indicated field journals. Top panels display histograms of test statistics for $t \in [0, 5]$, with bins of width 0.1 respectively. Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve (which includes renormalization at 0). Bottom panels display histograms of test statistics for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025.



Extended Data Fig. 10 | Relative reproduced effect size. 48% of relative effect sizes are exactly equal to or greater than 1. This figure illustrates the ratio of re-analysis estimates and original estimates. The standardized effect sizes are normalized so that 1 equals the original effect size. A positive value indicates that the re-analysis estimate is in the same direction as in the original study. A negative value indicates that the re-analysis estimate is not in the same direction as in the original study. Outliers (3%) are excluded for visibility.

nature portfolio

Corresponding author(s): Abel Brodeur

Last updated by author(s): Jan 28, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Our codes are available here: https://zenodo.org/records/17792605 and https://osf.io/8wsqx/
Data analysis	*Provide a description of all commercial, open source and custom code used to analyse the data in this study, specifying the version used OR state that no software was used.*

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The publicly available data are available here: https://zenodo.org/records/17792605 and https://osf.io/8wsqx/. See Readme for more details.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This systematic and large-scale reproduction effort tests the reproducibility and robustness of economics and political science, contributing to a growing literature on research credibility and self-correction in science. We reproduced original analyses and conducted robustness checks of 110 articles recently published in leading economics and political science journals, all of which have mandatory data and code sharing policies. We found that over 85% of published claims were computationally reproducible. In robustness checks, our re-analyses led to 72% of statistically significant estimates to remain significant and in the same direction, and the median reproduced effect size is (nearly) the same as the originally published effect size (that is, 99% of the published effect size). Additionally, six independent research teams examined 12 pre-specified hypotheses about determinants of robustness. Research teams with more experience found lower levels of robustness, and robustness correlated with neither author characteristics nor data availability.
Research sample	110 articles. This is not a representative sample. Not all studies from our targeted journals have been reproduced or replicated. Our approach leads to an over-representation of studies using publicly available data. Another feature of our sample is that the targeted journals have a data availability policy and enforce it. This is in contrast to many top field journals in both economics and political science. Our sample should thus be viewed as very selected both in terms of impact and high data and code availability rates. In fact, approximately 45% of replication packages in our sample included raw data and complete cleaning code. An additional 13.5% provided partial cleaning code.
Sampling strategy	No sample size calculation. We explore the reasons why the 110 teams selected their paper. All teams answered the following question: ``For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided?'' 12 options were offered, including Other (please specify). Options were not mutually exclusive, so any one team could provide multiple reasons for why they selected their paper. Extended Data Figure 2 summarizes the percentage of teams who selected each category. Of note 13.6% of teams were assigned a study (ie., did not choose which study to work on), so they did not answer this question. About 45% of teams report ``Methods used'', 36% of teams selected because of the journal of publication, about 25% due to the ``Length of time to reproduce results'' and about 19% due to the ``Size of replication package''. This is in line with our provided guidelines for choosing a study. If a large portion of reproducers select papers based on the assumption that their findings are questionable, it could skew reproducibility rates downward, as there's a tendency to pick studies more prone to revealing problematic outcomes. However, in this project, only a minimal fraction of teams indicated that they chose their paper because of ex ante beliefs that main results are (not) robust/replicable (3.6%). Few teams also selected papers based on statistical power/sample size and trust of original authors. We explore if our sample is representative of all subfields within economics. We compare JEL Codes of economic papers that we reproduced relative to those of a random sample of representative journal articles published in the top 100 journals in Economics (as ranked by https://ideas.repec.org/top/top.journals.simple.html). Our sample under-represents, among other fields, C-Mathematical and Quantitative Methods, G-Financial Economics and F-International Economics.
Data collection	N/A. We did not collect data. We reproduce published articles.
Timing	The articles were published between 2022 and 2023.
Data exclusions	No data excluded.

Non-participation	No participants.
Randomization	Not applicable as we did not randomize the studies we reproduced.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*