Reproducibility and Robustness of Economics and Political Science Research

Abel Brodeur et al. (Author list and contributions are provided in the SI)[1*]

[1*]Department of Economics and Institute for Replication, University of Ottawa, 75 Laurier Avenue East, Ottawa, K1N 6N5, Ontario, Canada.

Corresponding author(s). E-mail(s): abrodeur@uottawa.ca;

Abstract

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 [1–4]. 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 [17,18]. 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.

1 Introduction

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, there has been little social science reproduction and robustness conducted at scale [10, 13, 17–23].

This project is a mega-reproduction led by the Institute for Replication (I4R), which evaluates the reproducibility and robustness of 110 published studies in economics (79) and political science (31). Our focus is on studies published in 12 prestigious journals between 2022 and 2023. While each of these journals has a data and code availability policy requiring authors to publicly share their materials upon publication, most (though not all) also appoint 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 (see Supplementary Materials 11.5).

Not all studies from our targeted journals were chosen for reproduction and robustness, 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 using publicly available data ([18]). 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 selective both in terms of impact and high data and code availability rates, and might present an optimistic upper bound on reproducibility rates. In fact, virtually all papers in our sample include replication packages with cleaned data and code to reproduce the paper's results, and about 30% also provide the raw data and cleaning code used to generate the analytical data (Extended Data Figure 1, Levels 8, 9, and 10).

While this project relates to the broader reproducibility movement in psychology, neuroscience, or biomedicine, it 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 the first evidence about the general robustness of economics and political science. Third, we concentrate on recent studies for both economics and political science. Finally, this is an ongoing initiative that aims to expand across disciplines, with the goal of mass reproducing studies and reshaping research norms at scale. This paper reports findings from the first 110 reproductions.

2 Definitions

We follow [27]'s nomenclature and define a claim **computationally reproducible** if its results can be reproduced using the original study's data and protocols. A claim is **robust** if its results are robust to alternative reasonable analytical decisions on the same data. Last, a claim is **replicable** if its results can be repeated using new data.

3 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 (Games) ([28]). Games are one-day events

open to faculty, post-docs, graduate students and other researchers. Participants are assigned to a small team of about 3–5 other researchers all working in the same subfield (*e.g.*, development economics).

Participant teams are offered a short list of (average 5) studies in their subfield of interest about three weeks before the games. They are asked to choose a paper as a team, and familiarize themselves with the data and codes publicly posted by the original authors (i.e., replication package) prior to the games. After the Game, teams submit a standardized reproduction report summarizing their results.

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

We survey the reasons why teams selected their paper (Extended Data Figure 2). While 13.6% of teams were assigned a study (*i.e.*, did not choose which study to work on), about 45% of teams report "Methods used", 36% of teams selected "because of the journal of publication" and about 25% due to 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 downward, 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 the paper is not robust is *inversely* correlated with reproducer experience ($\rho = -0.19, p < 0.000$). A few teams (5%) indicated that their choice was based on statistical power/sample size and trust of original authors.

4 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: (1) the raw and analytical data, or; (2) 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 where code failed to run or produced inconsistent results (See Supplementary Materials 11.8 and Extended Data Figure 1). Fixing paths, missing packages and software requirements were not considered failures of computational reproducibility. In those instances, we fixed paths, added missing packaged and software requirements.

Our findings suggest high rates of computationally reproducible results, but far from perfect for leading journals. Our results are in contrast with several studies documenting low computational reproducibility rates in economics [19]; [13]; [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 Supplementary Materials 11.11, data and code sharing practices have dramatically improved during this period. We found replication folders are attached to 59% of papers in 2014, while replication folder provision increases to a seemingly stable value close to 90% in 2021–2023 (Extended Data Figures 3, 4, 5 and 6). Additionally, for journals that introduced data editors during this period, much of this improvement occurred during the first year following this change.

5 Robustness

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

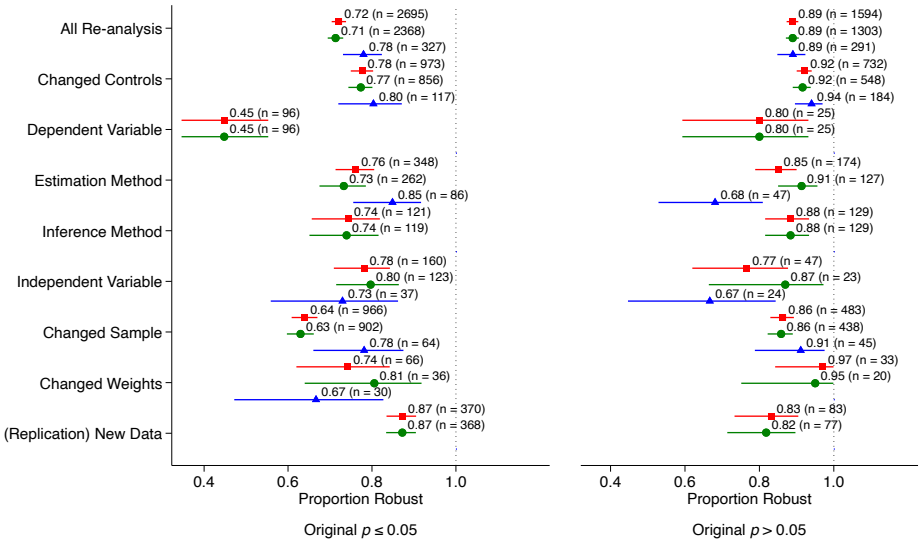
Reproducers are then free to conduct any robustness or recoding exercises. They focus on the reproducibility of the claims and have access to the replication package, allowing them to directly test the robustness of the main results. This is a crucial advantage over the traditional review process as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

However, this flexibility also brings some disadvantages. As with the journal review process with reviewers, reproducers spend different amounts of time and effort on their respective replication. Some reproducers are more experienced at coding, while others are more familiar with methods, or simply unable to implement robustness checks due to a lack of raw data (Extended Data Figure 7). This means that reproducibility efforts and type of re-analysis vary across teams. Teams worked on average 13 active days (std. dev. of 24) on the reproductions and robustness, and reports were on average 19 pages long (std. dev. of 14).

Figure 1 (top of left panel) 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.

Figure 1. Robustness Rate.

Fig. 1: Robustness Rate



Robustness rate for ... **Left panel:** ... originally statistically significant research **Second panel:** ... in economics **Third panel:** ... in political science **Right panel:** ... originally statistically *in*significant research **All panels:** Squares, circles, and triangles represent proportions, with 95% Clopper-Pearson confidence intervals presented in whiskers. Red squares represent full sample. Green circles represent economics subsample. Blue triangles represent political science subsample. Each group of three estimates represent different types of re-analysis, non-mutually exclusive. The first 8 groups do not include re-analyses that use new data (replication), while the last one does. The first estimate group contains all types of re-analysis, then all types of re-analysis in economics, then all types of re-analysis in political science. The second represents re-analyses which changed the control variables, e.g., by adding or re-defining them. The third represents re-analyses which changed the dependent variable, e.g., by employing a different standardization or binarization. The fourth represents re-analyses which changed the estimation method, e.g., by adjusting a matching procedure. The fifth represents re-analyses which changed the inference method, e.g., changed the level on which standard errors are clustered. The sixth represents re-analyses which changed the main independent variable, e.g., by taking into account treatment intensity. The seventh represents re-analyses which changed the sample, e.g., by excluding outliers. The eighth represents re-analyses which changed the weights applied, or applied weights for the first time. The last represents replicability rates for re-analyses that introduced new data, e.g., comparable outcomes from more recent survey waves.

We find large differences by re-analysis type. The re-analysis type that has the highest robustness rate (78%) is changing the independent variable measure (examples include log transformations, discretization, etc.). The re-analysis type that has the lowest robustness rate (45%) is any which included changing the dependent variable measure (e.g., categorizing the variable or log-transforming). When a replication (addition of new data, e.g., from more recent survey waves or an alternative source) is applied, the replication rate is 87%.

The average robustness rate is 71% and 78% for economics and political science, respectively, where the 6.7% difference is statistically significant (two-sample difference in proportions $z = -2.52$, $p = 0.012$, $n_1 = 2368$, $n_2 = 327$). 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

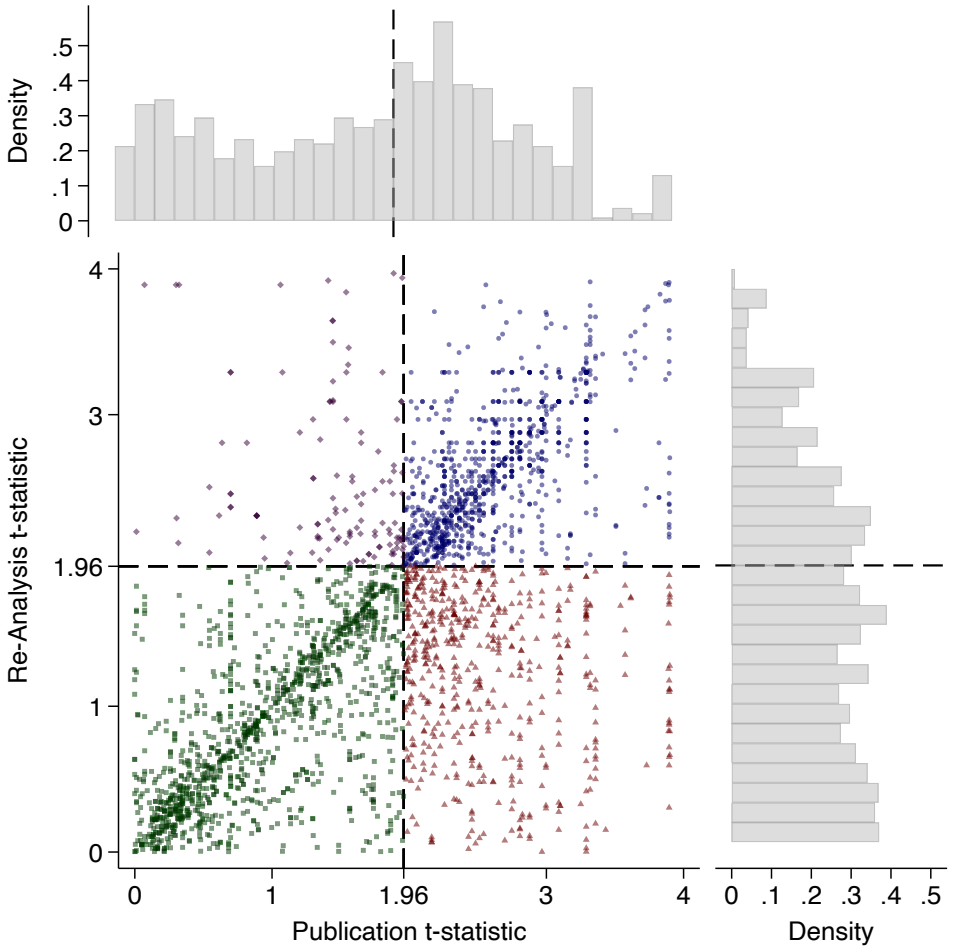
political science re-analyses). Focusing on robustness rates for originally statistically *in*significant findings, we find a robustness rate of 89%.

Supplementary Materials Appendix Table 1 shows shifts in statistical significance between all significance regions. We find that 7.44% of re-analyses find an effect with the opposite sign as the original result. In contrast, of the 62.84% of original analyses that were statistically significant, most remained significant and of the same sign (44.33%/62.84% = 70.5%). Of particular note is the 15.06% of re-analyses that find a statistically *in*significant result for originally statistically significant analyses.

We illustrate in Figure 2 the distribution of test statistics for the original point estimates and the re-analyses. We find that 53% of the originally published test statistics are statistically significant (to the right of the statistical significance threshold). In contrast, 43% of re-analyses are statistically significant (above the statistical significance threshold in the vertical axis histogram). The simple difference in proportions is statistically significant (difference of 10.4%, McNemar's $\chi^2 = 264.11$, $p < 0.001$, $n = 4750$).

Figure 2. Statistical Significance of Publication and Re-analysis.

Fig. 2: Statistical Significance of Publication and Re-analysis



Top histogram: Distribution of publication tests of significance. T-statistics over 4 truncated for exposition. The histogram's bars 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 re-analysis tests of significance. T-statistics over 4 truncated for exposition. **Scatterplot:** Each marker is a pair of test statistics, an originally published test statistic (horizontal value) and an associated re-analysis test statistic (vertical value). If the original and re-analysis test statistics were identical, this scatterplot would follow the 45 degree line. As either axis represents statistical significance, we have bifurcated each with a line at t=1.96, representing statistical significance at the 5% threshold. **Blue circles** indicate an originally statistically significant statistic that is also statistically significant under re-analysis. Represents 50% of sample. **Red triangles** indicate originally significant test statistics that are no longer statistically significant under re-analysis. Represents 14% of sample. **Green squares** indicate originally statistically insignificant test statistics that are the same under re-analysis. Represents 27% of sample. **Purple diamonds** indicate originally statistically insignificant test statistics that become statistically significant under re-analysis. Represents 3% of sample. **Not displayed** Not displayed are the 6% of test statistics that represent a sign reversal between the originally estimated effect and the effect estimated under re-analysis.

When expressed as t-statistics, the average originally published t-statistic is 1.797 whereas the average re-analysis t-statistic is 1.544. The difference between the pairs of original study estimates and re-analysis 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 ($p < 0.001$). Here, we also note the large increase in test statistic density immediately after the statistical significance threshold (Extended Data Figures 8 and 9), which offers strong evidence of publication bias in originally published research ([30, 31]). In contrast, this increase at the significance level threshold is missing from the vertical axis histogram depicting the distribution of re-analyses.

When expressed as p-values, the average originally published p-value is 0.167 whereas the average re-analysis p-value is 0.219; the difference is statistically significant (Wilcoxon signed-rank test $z = -16.007$, $p < 0.001$, $n = 4063$).

In this project, we conduct multiple re-analyses per original study, and so it is possible that much of the differences between original studies and their re-analyses are driven or characterized by large changes in a small subset of studies rather than indicative of more general shifts between original and re-analysis. 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 re-analyses 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.

6 Determinants of Robustness

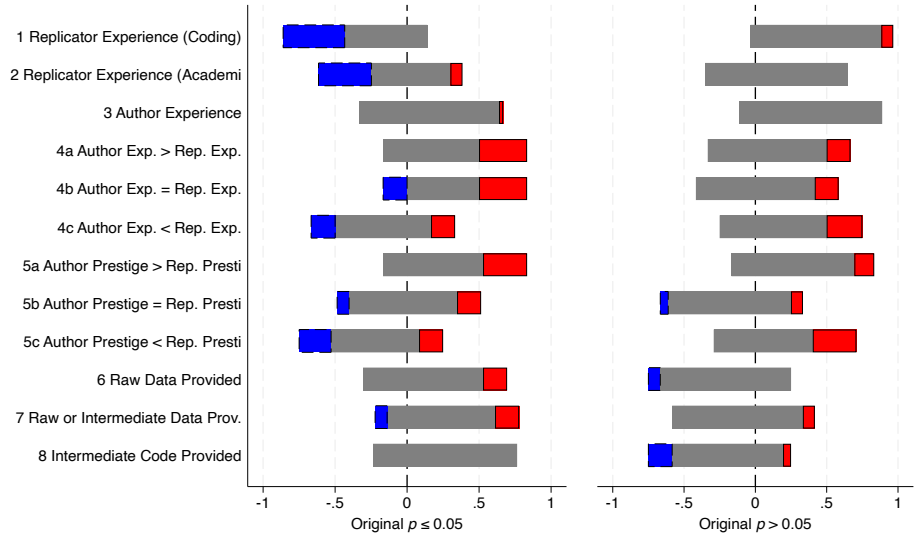
This section examines what, if any, characteristics of the authors, reproducers, or the original articles are informative of the robustness rate.

While this analysis is merely exploratory, this project applied both a pre-registration and many-analysts approach [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 10 individuals and ended up forming six many-analysts teams. Each team answered several research questions. The results are displayed in Figure 3.

Figure 3. Robustness Rate Determinants.

Fig. 3: Robustness Rate Determinants



Six independent teams answered twelve questions of the re-analysis database. Each bar represents a different question. **Left panel:** "Does reproducibility of an originally statistically significant result depend on..." **Right panel:** "Does reproducibility of an originally statistically *insignificant* result depend on..." **Both panels:** where the first bar represents "... the reproducers' experience at coding." **Blue, patterned outline** indicates the proportion of teams that indicated a negative and statistically significant relationship, in whichever manner the team defined so in their analysis. **Gray, no outline** indicates the proportion of teams that indicated a statistically insignificant relationship, where left of the zero line indicates negative and right of the zero line indicates positive. **Red, solid outline** indicates the proportion of teams that indicated a statistically significant and positive relationship. All teams equally weighted.

They began by analyzing 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; likening 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, whether authors have (4a) more, (4b) similar, or (4c) less experience than reproducers, (5a) more, (5b) similar, or (5c) less prestige (their institution, defined independently by the analysts) than reproducers, and whether (6) raw data was provided (7) raw or intermediate data was 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. One plausible interpretation of our main results therefore is that robustness in our full sample would likely have been lower if equally highly qualified replicator teams had been assigned to each paper. However, according to the results presented in the main text (Determinants of Robustness), the provision of raw or intermediate data, or the necessary cleaning codes, does not seem to affect the robustness of research.

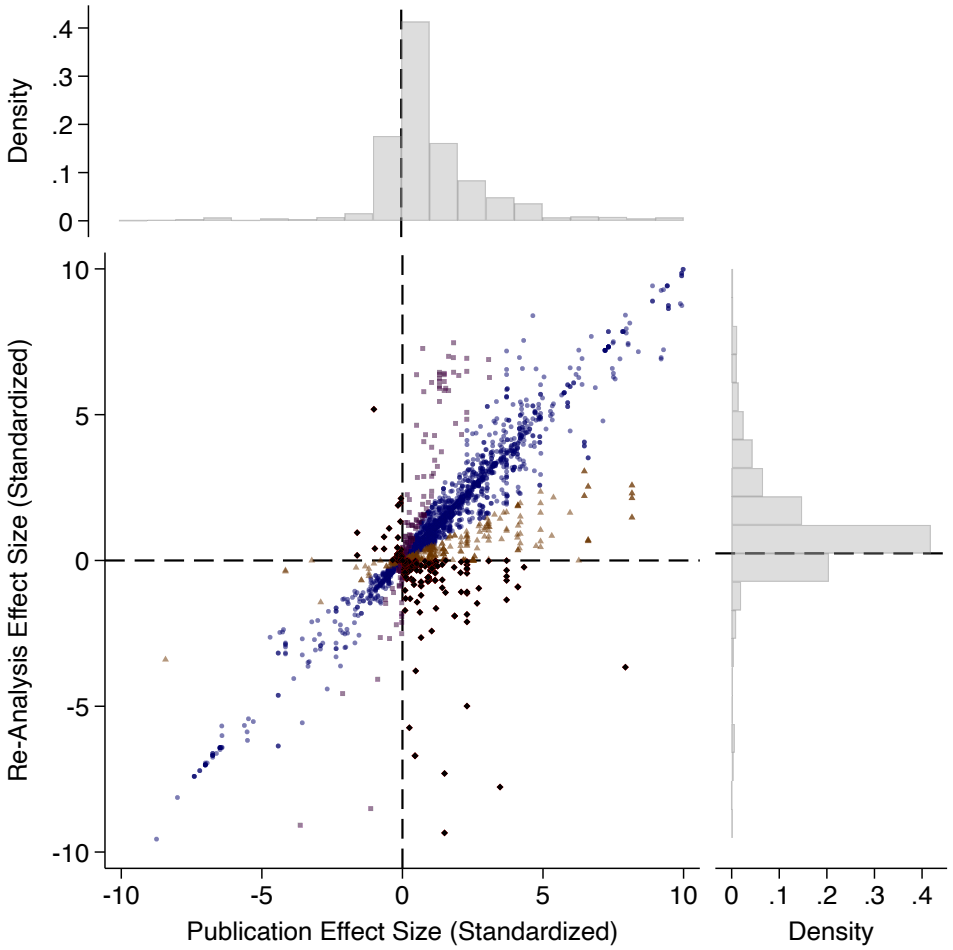
When analysts examined these same 12 hypotheses for originally statistically *in*significant results, the relationships are far more likely to be positive than negative, but (as indicated by the proportion in gray) the relationships are often not statistically significant.

7 Effect Size

Figure 4 displays publication and re-analysis effect sizes. In economics and political science, effect sizes are largely reported as non-unit-less regression coefficients, whereas in other sciences, effect sizes are often reported using more comparable measures such as Cohen's-d. Because raw effect sizes vary widely between original studies, each of the markers are standardized by the within-article average published effect size (e.g., estimated effects of 2, 4, and 6 are standardized within publication to be 0.5, 1.0, and 1.5).

Figure 4. Effect Size of Publication and Re-analysis.

Fig. 4: Effect Size of Publication and Re-analysis



Top histogram: Distribution of originally published effect size standardized by the average effect size within a published article. **Right histogram:** Distribution of re-analysis published effect size standardized by the average effect size within a published article. **Scatterplot:** Each marker is a pair of effect sizes, the originally published effect size (horizontal value) and an associated re-analysis effect size (horizontal value). If an originally estimated and re-analysis effect size were of similar magnitude (and sign), the markers would gather tightly around the 45 degree line passing through the origin **Blue circles** indicate effect sizes which are similar (between 50% to 200% of original effect size) under re-analysis. Represents 69% of sample. **Red diamonds** indicate effect size estimates which switch sign under re-analysis. Represents 6% of sample. **Orange triangles** indicate effect size estimates which are 50% or less their original magnitude under re-analysis. Represents 9% of sample. **Purple squares** indicate effect size estimates which are double or larger than their original magnitude under re-analysis. Represents 16% of sample.

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

8 Coding Errors and Recoding

We investigate the prevalence of coding errors and discrepancies between the code and article. Computational reproducibility pertains to the provided replication folder's ability to reproduce the exhibits and statistics displayed in the research (manuscripts, appendices, *etc.*). 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 (*i.e.,* missing packages or broken pathways), we identify coding errors in approximately 25% of the studies, with some studies containing multiple errors (Supplementary Materials 11.10). The prevalence of coding errors is larger for economics (26%) than political science (16%). Types of errors include: defining the dependent variable, defining the main independent variable, defining control variables, mis-specification of the estimation/model, inference or the sample. While not all of these coding errors impacted 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, mislabeling the main treatment variable for a substantial number (or all) of observations, and using different models, or estimators, than reported in the article.

It is important to note that this 25% figure likely underestimates 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 a different statistical software. Out of 23 recoding exercises, we find major differences for three studies and minor differences for 10 studies. Two of the major differences were uncovered when using a different software and looking at the authors' code. Additionally, one team who computationally reproduced the results using a different *version* of the software used by the authors uncovered noteworthy differences in the magnitude and significance of the estimates (Supplementary Materials 11.9).

9 Communication with Original Authors

I4R shares completed reproduction reports with original authors before public release ([28]). Reports are reviewed typically by A.B. or another board member mainly for

tone and structure. I4R then disseminates the report and any author response simultaneously (see SI for the full list of reports). Reproducers may revise their reports after receiving feedback from original authors.

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

Roughly 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 adjust tone. In one case, original authors conducted additional robustness checks in their non-public files at the reproducers' request.

Lastly, 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 appear in the Supplementary Materials.

10 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 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. While 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 future work will address.

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 they examined. Over 40% reported gaining a more optimistic view of the discipline, while only about 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 initiative's success and scalability 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 over 3,500 researchers, with events held every other week. These efforts show that the skilled labor 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 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 even fewer 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.

Author list

Abel Brodeur, Derek Mikola, Nikolai Cook, Lenka Fiala, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, Jacopo Gabani, Romain Gauriot, Joanne Haddad, Goncalo Lima, Jörg Ankel-Peters, Anna Dreber, Douglas Campbell, Lamis Kattan, Diego Marino Fages, Fabian Mierisch, Pu Sun, Taylor Wright, Marie Connolly, Fernando Hoces de la Guardia, Magnus Johannesson, Edward Miguel, Lars Vilhuber, Alejandro Abarca, Mahesh Acharya, Sossou Simplice Adjisse, Ahwaz Akhtar, Eduardo Alberto Ramirez Lizardi, Sabina Albrecht, Synøve Nygaard Andersen, Zubaria Andlib, Falak Arrora, Thomas Ash, Etienne Bacher, Sebastian Bachler, Félix Bacon, Manuel Bagues, Timea Balogh, Alisher Batmanov, Mara Barschkett, B. Kaan Basdil, Jaromír Baxa, Sascha Becker, Monica Beeder, Louis-Philippe Beland, Abdel-Hamid Bello, Daniel Benenson Markovits, Grant Benjamin, Thomas Bergeron, Moussa Blimpo, Marco Binetti, Carl Bonander, Joseph Bonneau, Endre Borbáth, Nicolai Topstad Borgen, Solveig Topstad Borgen, Jonathan Borowsky, Elisa Brini, Myriam Brown, Martin Brun, Stephan Bruns, Nino Buliskeria, Andrea Calef, Alistair Cameron, Pamela Campa, Santiago Campos-Rodríguez, Giulio Giacomo Cantone, Fenella Carpena, Perry Carter, Paul Castañeda Dower, Ondrej Castek, Jill Caviglia-Harris, Gabriella Chauca Strand, Shi Chen, Sya In Chzhen, Jong Chung, Jason Collins, Alexander Coppock, Hugo Cordeau, Ben Couillard, Jonathan Crechet, Lorenzo Crippa, Jeanne Cui, Christian Czymara, Haley Daarstad, Danh Chi Dao, Daniel Dao, Marco David Schmandt, Astrid de Linde, Lucas De Melo, Lachlan Deer, Micole De Vera, Velichka Dimitrova, Jan Fabian Dollbaum, Jan Matti Dollbaum, Michael Donnelly, Luu Duc Toan Huynh, Tsvetomira Dumbalska, Jamie Duncan, Kiet Tuan Duong, Thibaut Duprey, Christoph Dworschak, Sigmund Ellingsrud, Ali Elminejad, Yasmine Eissa, Andrea Erhart, Giulian Etingin-Frati, Elaheh Fatemipour, Alexa Federice, Jan Feld, Guidon Fenig, Mojtaba Firouzjaeiangalougah, Erlend Fleisje, Alexandre FortiFriter-Chouinard, Julia Francesca Engel, Nadjim Fréchet, Reid Fortier, Tilman Fries, Michael James Frith, Thomas Galipeau, Sebastián Gallegos, Areez Gangji,

Xiaoying Gao, Cloé Garnache, Attila Gáspár, Evelina Gavrilova, Arijit Ghosh, Gareth Gibney, Grant Gibson, Geir Godager, Leonard Goff, Da Gong, Javier González, Jeremy D. Gretton, Cristina Griffa, Idaliya Grigoryeva, Maja Grøtting, Eric Guntermann, Jiaqi Guo, Alexi Gugushvili, Hooman Habibnia, Sonja Häffner, Jonathan D. Hall, Olle Hammar, Amund Hanson Kordt, Barry Hashimoto, Jonathan S. Hartley, Carina I. Hausladen, Tomáš Havránek, Harry He, Matthew Hepplewhite, Mario Herrera-Rodriguez, Felix Heuer, Anthony Heyes, Anson T. Y. Ho, Jonathan Holmes, Armando Holzknecht, Yu-Hsiang Dexter Hsu, Shiang-Hung Hu, Yu-Shiuan Huang, Mathias Huebener, Christoph Huber, Kim P. Huynh, Zuzana Irsova, Ozan Isler, Niklas Jakobsson, Raphaël Jananji, Tharaka A. Jayalath, Michael Jetter, Jenny John, Rachel Joy Forshaw, Felipe Juan, Valon Kadriu, Sunny Karim, Edmund Kelly, Duy Khanh Hoang Dang, Tazia Khushboo, Jin Kim, Gustav Kjellsson, Anders Kjelsrud, Andreas Kotsadam, Jori Korpershoek, Lewis Krashinsky, Suranjana Kundu, Alexander Kustov, Nurlan Lalayev, Audrée Langlois, Jill Laufer, Blake Lee-Whiting, Andreas Leibing, Gabriel Lenz, Joel Levin, Peng Li, Tongzhe Li, Yuchen Lin, Ariel Listo, Dan Liu, Xuewen Lu, Elvina Lukmanova, Alex Luscombe, Lester R. Lusher, Ke Lyu, Hai Ma, Nicolas Mäder, Clifton Makate, Alice Malmberg, Adit Maitra, Marco Mandas, Jan Marcus, Shushanik Margaryan, Lili Márk, Andres Martignano, Abigail Marsh, Isabella Masetto, Anthony McCanny, Emma McManus, Ryan McWay, Lennard Metson, Jonas Minet Kinge, Sumit Mishra, Myra Mohnen, Jakob Möller, Rosalie Montambeault, Sébastien Montpetit, Louis-Philippe Morin, Todd Morris, Scott Moser, Fabio Motoki, Lucija Muehlenbachs, Andreea Musulan, Marco Musumeci, Munirul Nabin, Karim Nchare, Florian Neubauer, Quan M. P. Nguyen, Tuan Nguyen, Viet Nguyen-Tien, Ali Niazi, Giorgi Nikolaishvili, Ardyn Nordstrom, Patrick Nüß, Angela Odermatt, Matt Olson, Henning Øien, Tim Ölkers, Miquel Oliver i Vert, Emre Oral, Christian Oswald, Ali Ousman, Ömer Özak, Shubham Pandey, Alexandre Pavlov, Martino Pelli, Romeo Penheiro, RyuGyung Park, Eva Pérez Martel, Tereza Petrovičová, Linh Phan, Alexa Prettyman, Jakub Procházka, Aqila Putri, Julian Quandt, Kangyu Qiu, Loan Quynh Thi Nguyen, Andaleeb Rahman, Carson H. Rea, Adam Reiremo, Laëtitia Renée, Joseph Richardson, Nicholas Rivers, Bruno Rodrigues, William Roelofs, Tobias Roemer, Ole Rogeberg, Julian Rose, Andrew Roskos-Ewoldsen, Paul Rosmer, Barbara Sabada, Soodeh Saberian, Nicolas Salamanca, Georg Sator, Daniel Scates, Elmar Schlüter, Cameron Sells, Sharmi Sen, Ritika Sethi, Anna Shcherbiak, Moyosore Sogaolu, Matt Soosalu, Erik Ø. Sørensen, Manali Sovani, Noah Spencer, Stefan Staubli, Renske Stans, Anya Stewart, Felix Stips, Kieran Stockley, Stephenson Strobel, Ethan Struby, John Tang, Idil Tanrisever, Thomas Tao Yang, Ipek Tastan, Dejan Tatić, Benjamin Tatlow, Féraud Tchuisseu Seuyong, Rémi Thériault, Vincent Thivierge, Wenjie Tian, Filip-Mihai Toma, Maddalena Totarelli, Van Tran, Hung Truong, Nikita Tsoy, Kerem Tuzcuoglu, Diego Ubfal, Laura Villalobos, Julian Walterskirchen, Joseph Tao-yi Wang, Vasudha Wattal, Matthew D. Webb, Bryan Weber, Reinhard Weisser, Wei-Chien Weng, Christian Westheide, Kimberly White, Jacob Winter, Timo Wochner, Matt Woerman, Jared Wong, Ritchie Woodard, Marcin Wroński, Myra Yazbeck, Chung Yang, Luther Yap, Kareman Yassin, Hao Ye, Jin Young Yoon, Chris Yurris, Tahreen Zahra, Mirela Zaneva, Aline Zayat, Jonathan Zhang, Ziwei Zhao, Yaolang Zhong

Declarations

- Funding:
  We acknowledge support from Coefficient Giving and the Social Sciences and Humanities Research Council.
- Conflict of interest/Competing interests:
  Any views expressed herein are the authors' personal opinions and not those of Ontario Public Service. The work by Jeremy D. Gretton was not undertaken under the auspices of 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.
- Data availability: The data are available on Zenodo (link: https://zenodo.org/records/17792605; DOI: 10.5281/zenodo.17792605) and OSF (link: https://osf.io/8wsqx/; DOI: 10.17605/OSF.IO/8WSQX). See OSF for our pre-analysis plan.
- Code availability: The codes are available on Zenodo (https://zenodo.org/records/17792605) and OSF (https://osf.io/8wsqx/).
- Author contribution:
  Preparation of tables, figures, and manuscript: Abel Brodeur (University of Ottawa and Institute for Replication), Nikolai Cook (Wilfrid Laurier University), Derek Mikola (University of Ottawa and Institute for Replication), Lenka Fiala (University of Ottawa, Tilburg University and Institute for Replication)
  Conception or design of the work: Jörg Ankel-Peters (RWI - Leibniz Institute for Economic Research), Abel Brodeur (University of Ottawa and Institute for Replication), Marie Connolly (UQAM), Nikolai Cook (Wilfrid Laurier University), Anna Dreber (Stockholm School of Economics), Fernando Hoces de la Guardia (Berkeley Initiative for Transparency in the Social Sciences), Magnus Johannesson (Stockholm School of Economics), Edward Miguel (UC Berkeley), Derek Mikola (University of Ottawa and Institute for Replication), Lars Vilhuber (Cornell University)
  Analysis or interpretation of the reproducibility data: Thomas Brailey (University of Oxford), Ryan Briggs (University of Guelph), Abel Brodeur (University of Ottawa and Institute for Replication), Nikolai Cook (Wilfrid Laurier University), Alexandra de Gendre (The University of Melbourne), Yannick Dupraz (Paris Dauphine University, PSL University, LEDA, CNRS, IRD), Jacopo Gabani (World Bank & Centre for health economics, university of York), Romain Gauriot (Deakin University), Goncalo Lima (European University Institute and University of Bologna), Derek Mikola (Institute for Replication)

Analysis or interpretation of data And generating data And conception of a reproduction:
Douglas Campbell (Independent Researcher), Nikolai Cook (Wilfrid Laurier University), Joanne Haddad (Universitat Autonoma de Barcelona), Lamis Kattan (School of Foreign Service, Georgetown University Qatar), Diego Marino Fages (Durham University), Fabian Mierisch (Independent Researcher), Pu Sun (Dongbei University of Finance and Economics), Taylor Wright (Brock University), Alejandro Abarca (Texas Tech University), Mahesh Acharya (University of Calgary), Sossou Simplice Adjisse (University of Wisconsin-Madison and African School of Economics), Ahwaz Akhtar (George Washington University), Eduardo Alberto Ramirez Lizardi (University of Oslo), Sabina Albrecht (University of Queensland), Synøve Nygaard Andersen (University of Oslo), Zubaria Andlib (Lancaster University and Federal Urdu University of Arts, Science and Technology), Falak Arrora (University of Warwick), Thomas Ash (Anderson School of Management, UCLA), Etienne Bacher (Luxembourg Institute of Socio-Economic Research), Sebastian Bachler (University of Innsbruck), Félix Bacon (Laval University), Manuel Bagues (University of Warwick), Timea Balogh (UC Davis), Alisher Batmanov (UC San Diego), Mara Barschkett (University of Bonn, IZA & DIW Berlin), B. Kaan Basdil (Risk Software Technologies), Jaromír Baxa (Institute of Economic Studies, Faculty of Social Sciences, Charles University, and Institute of Information Theory and Automation AS CR), Sascha Becker (University of Warwick and Monash University), Monica Beeder (University of Southampton), Louis-Philippe Beland (Carleton University), Abdel-Hamid Bello (University of Montreal), Daniel Benenson Markovits (Columbia University), Grant Benjamin (University of Toronto), Thomas Bergeron (Université de Montréal), Moussa P. Blimpo (University of Toronto), Marco Binetti (Institute of Intercultural and International Studies, University of Bremen), Carl Bonander (Karlstad Business School, Karlstad University), Joseph Bonneau (UC Davis), Endre Borbáth (Ruprecht-Karls-Universität Heidelberg), Nicolai Topstad Borgen (Centre for Research on Equality in Education, University of Oslo), Solveig Topstad Borgen (University of Oslo), Jonathan Borowsky (University of Minnesota), Thomas Brailey (University of Oxford), Ryan Briggs (University of Guelph), Elisa Brini (University of Florence), Myriam Brown (Laval University), Martin Brun (Finnish Centre of Excellence in Tax Systems Research, Tampere University), Stephan Bruns (Hasselt University, INCHER Kassel, METRICS Stanford), Nino Buliskeria (Nazarbayev University), Andrea Calef (University College London, School of Management), Alistair Cameron (Monash University), Pamela Campa (Stockholm Institute of Transition Economics), Santiago Campos-Rodríguez (University of California, Irvine), Giulio Giacomo Cantone ("Magna Graecia" University of Catanzaro), Fenella Carpena (Oslo Business School, Oslo Metropolitan University), Perry Carter (NYU Abu Dhabi), Paul Castañeda Dower (University of Wisconsin-Madison), Ondrej Castek (Masaryk University), Jill Caviglia-Harris (Salisbury University), Gabriella Chauca Strand (Institute of Medicine, University of Gothenburg), Shi Chen (School of Economics, Zhejiang University), Sya In Chzhen (University of East Anglia), Jong Chung (Auburn University), Jason

Collins (University of Technology Sydney), Alexander Coppock (Northwestern University), Hugo Cordeau (University of Toronto), Ben Couillard (University of Toronto), Jonathan Crechet (University of Ottawa), Lorenzo Crippa (University of Strathclyde), Jeanne Cui (Beijing Normal University), Christian Czymara (Netherlands Interdisciplinary Demographic Institute), Haley Daarstad (UC Davis), Danh Chi Dao (Queen's University), Daniel Dao (University of Oxford), Marco David Schmandt (TU Berlin), Astrid de Linde (University of Oslo), Lucas De Melo (University of Nottingham, NICEP), Lachlan Deer (University of Melbourne), Alexandra de Gendre (The University of Melbourne), Micole De Vera (Banco de España), Velichka Dimitrova (Social Research Institute, University College London), Jan Fabian Dollbaum (University College Dublin), Jan Matti Dollbaum (University of Fribourg and LMU Munich), Michael Donnelly (University of Toronto), Luu Duc Toan Huynh (Queen Mary University of London), Tsvetomira Dumbalska (University of Oxford), Jamie Duncan (University of Toronto), Kiet Tuan Duong (University of York), Yannick Dupraz (Paris Dauphine University, PSL University, LEDA, CNRS, IRD), Thibaut Duprey (Bank of Canada), Christoph Dworschak (German Institute for Development Evaluation & University of York), Sigmund Ellingsrud (BI Norwegian Business School), Ali Elminejad (Nazarbayev University), Yasmine Eissa (The American University in Cairo), Andrea Erhart (University of Innsbruck), Giulian Etingin-Frati (ETH Zurich), Elaheh Fatemipour (University of Warwick), Alexa Federice (UC Davis), Jan Feld (Victoria University of Wellington), Guidon Fenig (University of Ottawa), Lenka Fiala (University of Ottawa, Tilburg University and Institute for Replication), Mojtaba Firouzjaeiangalougah (Masaryk University), Erlend Fleisje (Oslo Economics), Alexandre Fortier-Chouinard (Université Laval), Julia Francesca Engel (Kiel University), Nadjim Fréchet (Concordia University), Reid Fortier (VisualAIM), Tilman Fries (LMU Munich), Michael James Frith (University of Edinburgh), Jacopo Gabani (World Bank & Centre for health economics, university of York), Thomas Galipeau (University of Toronto), Sebastián Gallegos (UAI Business School), Areez Gangji (Independent Researcher), Xiaoying Gao (University of York), Cloé Garnache (Oslo Metropolitan University), Attila Gáspár (ELTE Centre for Economic and Regional Studies; Central European University), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Arijit Ghosh (RWI - Leibniz Institute for Economic Research), Garreth Gibney (University of Galway), Grant Gibson (Canadian Research Data Centre Network and McMaster University), Geir Godager (University of Oslo), Leonard Goff (University of Calgary), Da Gong (State University of New York, Geneseo), Javier González (Southern Methodist University), Jeremy D. Gretton (Ontario Public Service's Behavioural Insights Unit), Cristina Griffa (University of Chile), Idaliya Grigoryeva (UC San Diego), Maja Grøtting (The Norwegian Institute of Public Health), Eric Guntermann (UC Berkeley), Jiaqi Guo (University of Birmingham), Alexi Gugushvili (University of Oslo), Hooman Habibnia (WU Vienna University of Economics and Business), Sonja Häffner (Peace Research Institute Oslo), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus University and Institute for Futures Studies), Amund Hanson Kordt (University of Oslo), Barry Hashimoto (Independent), Jonathan S. Hartley (Stanford University), Carina I. Hausladen

(University of Konstanz), Tomáš Havránek (Institute of Economic Studies, Faculty of Social Sciences, Charles University; and Faculty of International Relations, Prague University of Economics and Business), Harry He (University of California, San Diego), Matthew Hepplewhite (University of Oxford), Mario Herrera-Rodriguez (CREST-Ecole Polytechnique; Programa Estado de la Nacion), Felix Heuer (RWI – Leibniz Institute for Economic Research), Anthony Heyes (University of Birmingham), Anson T. Y. Ho (Toronto Metropolitan University), Jonathan Holmes (University of Ottawa), Armando Holzknecht (University of Innsbruck), Yu-Hsiang Dexter Hsu (University of California, Davis), Shiang-Hung Hu (California Institute of Technology), Yu-Shiuan Huang (National Chengchi University), Mathias Huebener (Federal Institute for Population Research (BiB)), Christoph Huber (Aalto University), Kim P. Huynh (Indiana University, Department of Economics; and Université d'Orléans and the Laboratoire d'Économie d'Orléans), Zuzana Irsova (Institute of Economic Studies, Faculty of Social Sciences, Charles University, and Anglo-American University, Prague), Ozan Isler (The University of Queensland), Niklas Jakobsson (Karlstad University & FBK-IRVAPP), Raphaël Jananji (Université de Montréal), Tharaka A. Jayalath (Global Water Security Center), Michael Jetter (University of Western Australia), Jenny John (University of Ottawa), Rachel Joy Forshaw (Heriot-Watt University), Felipe Juan (Howard University), Valon Kadriu (University of Kassel and INCHER), Sunny Karim (Carleton University), Edmund Kelly (University of Oxford), Duy Khanh Hoang Dang (University College London), Tazia Khushboo (University of Calgary), Jin Kim (Chinese University of Hong Kong), Gustav Kjellsson (Centre for Health Governance & HEPER, School of Public Health & Communicy Medicine, University of Gothenburg), Anders Kjelsrud (Oslo Metropolitan University), Jori Korpershoek (Erasmus University Rotterdam), Andreas Kotsadam (Ragnar Frisch Centre for Economic Research), Lewis Krashinsky (Princeton University), Suranjana Kundu (World Inequality Lab, Paris School of Economics), Alexander Kustov (University of Notre Dame), Nurlan Lalayev (University of Warwick), Audrée Langlois (Université Laval), Jill Laufer (UC Center Sacramento (UC Davis)), Blake Lee-Whiting (University of Western Ontario), Andreas Leibing (Dresden University of Technology), Gabriel Lenz (UC Berkeley), Joel Levin (UC San Diego), Peng Li (University of Bath), Tongzhe Li (University of Guelph), Yuchen Lin (University of Warwick), Goncalo Lima (European University Institute and University of Bologna), Ariel Listo (University of Maryland), Dan Liu (Australian National University), Xuewen Lu (University of Calgary), Elvina Lukmanova (New Economic School), Alex Luscombe (Government of Canada), Lester R. Lusher (University of Pittsburgh), Ke Lyu (University of Nevada, Reno), Hai Ma (McGill University), Nicolas Mäder (Knauss School of Business, University of San Diego), Clifton Makate (Norwegian University of Life Sciences and Norwegian Geotechnical Institute), Alice Malmberg (UC Davis), Adit Maitra (The University of Melbourne), Marco Mandas (University of Cagliari), Jan Marcus (Freie Universität Berlin), Shushanik Margaryan (University of Potsdam), Lili Márk (Central European University), Diego Marino Fages (Durham University), Andres Martignano (University of Nottingham), Abigail Marsh (Finance Canada), Isabella Masetto

(London School of Economics and Political Science), Anthony McCanny (University of Toronto), Emma McManus (Health Organisation, Policy and Economics, The University of Manchester), Ryan McWay (University of Minnesota), Lennard Metson (London School of Economics and Political Science), Fabian Mierisch (Independent Researcher), Jonas Minet Kinge (University of Oslo), Sumit Mishra (Krea University), Myra Mohnen (University of Ottawa), Jakob Möller (WU Vienna University of Economics and Business), Rosalie Montambeault (Université Laval), Sébastien Montpetit (University of Warwick), Louis-Philippe Morin (University of Ottawa), Todd Morris (University of Queensland), Scott Moser (University of Nottingham, School of Politics and International Relations), Fabio Motoki (University of Texas Rio Grande Valley), Lucija Muehlenbachs (University of Calgary and Resources for the Future), Andreea Musulan (University of Montreal, IVADO, Mila), Marco Musumeci (University of Padova), Munirul Nabin (Deakin University), Karim Nchare (Vanderbilt University), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Quan M. P. Nguyen (University of Sussex), Tuan Nguyen (Hasselt University), Viet Nguyen-Tien (London School of Economics), Ali Niazi (University of Calgary), Giorgi Nikolaishvili (Wake Forest University), Ardyn Nordstrom (Carleton University), Patrick Nüß (IWH Halle), Angela Odermatt (University of Oxford), Matt Olson (University of Pennsylvania Wharton), Henning Øien (Department of Health Management and Health Economics, University of Oslo), Tim Ölkers (Humboldt University zu Berlin), Miquel Oliver i Vert (Universitat de Girona), Emre Oral (University of Mannheim), Christian Oswald (University of the Bundeswehr Munich), Ali Ousman (McGill University), Ömer Özak (Department of Economics, Southern Methodist University, IZA and GLO), Shubham Pandey (Institute of Psychology, Osnabrück University), Alexandre Pavlov (Université de Montréal), Martino Pelli (Asian Development Bank), Romeo Penheiro (University of Houston), RyuGyung Park (Government Department at William & Mary), Eva Pérez Martel (Universitat Autonoma de Barcelona), Jörg Ankel-Peters (RWI - Leibniz Institute for Economic Research), Tereza Petrovičová (UCSD), Linh Phan (UC Davis), Alexa Prettyman (Towson University), Jakub Procházka (Masaryk University), Aqila Putri (University of Maryland), Julian Quandt (WU Vienna University of Economics and Business), Kangyu Qiu (McMaster University), Loan Quynh Thi Nguyen (National Economics University), Andaleeb Rahman (Cornell University), Carson H. Rea (Emory University), Adam Reiremo (Norwegian School of Economics), Laëtitia Renée (Université de Montréal), Joseph Richardson (Lancaster University), Nicholas Rivers (University of Ottawa), Bruno Rodrigues (Ministry of Research and Higher Education, Luxembourg), William Roelofs (University of Toronto), Tobias Roemer (University of Oxford), Ole Rogeberg (Ragnar Frisch Centre for Economic Research), Julian Rose (RWI - Leibniz Institute for Economic Research), Andrew Roskos-Ewoldsen (UC Davis), Paul Rosmer (Humboldt University of Berlin & Berlin School of Economics), Barbara Sabada (Bank of Canada), Soodeh Saberian (University of Manitoba), Nicolas Salamanca (The University of Melbourne), Georg Sator (University of Nottingham & Institute for Advanced Studies Vienna), Daniel Scates (UC Davis), Elmar Schlüter (Justus Liebig University, Giessen), Cameron Sells (Indepenent Researcher), Sharmi Sen (Monash University),

Ritika Sethi (University of Chicago), Anna Shcherbiak (WU Vienna University of Economics and Business), Moyosore Sogaolu (GATE, Rotman, University of Toronto), Matt Soosalu (Carleton University), Erik Ø. Sørensen (NHH Norwegian School of Economics), Manali Sovani (Tufts University), Noah Spencer (University of Toronto), Stefan Staubli (University of Calgary), Renske Stans (The Netherlands Court of Audit), Anya Stewart (UC Davis), Felix Stips (Institute for Employment Research (IAB)), Kieran Stockley (University of Nottingham), Stephenson Strobel (McMaster University), Ethan Struby (Carleton College, Boston College, and Minnesota Supercomputing Institute), John Tang (Utrecht University), Idil Tanrisever (University of California, Irvine), Thomas Tao Yang (Australian National University), Ipek Tastan (University of Calgary), Dejan Tatić (WU Vienna University of Economics and Business), Benjamin Tatlow (University of Nottingham), Féraud Tchuisseu Seuyong (Université de Montréal), Rémi Thériault (New York University), Vincent Thivierge (University of Ottawa), Wenjie Tian (University of Ottawa), Filip-Mihai Toma (Bucharest University of Economic Studies), Maddalena Totarelli (Ifo Institute & Ludwig Maximilian University of Munich), Van-Anh Tran (Monash University), Hung Truong (University of Ottawa), Nikita Tsoy (INSAIT, Sofia University), Kerem Tuzcuoglu (Amazon), Diego Ubfal (World Bank), Laura Villalobos (Salisbury University), Julian Walterskirchen (University of Gothenburg), Joseph Tao-yi Wang (Department of Economics and Taiwan Social Resilience Research Center, National Taiwan University), Vasudha Wattal (The University of Manchester), Matthew D. Webb (Carleton University), Bryan Weber (College of Staten Island - CUNY), Reinhard Weisser (University of the West of England), Wei-Chien Weng (University of California, Davis), Christian Westheide (Stockholm Business School, Stockholm University & Leibniz Institute for Financial Research SAFE), Kimberly White (Ludwig Maximilian University of Munich), Jacob Winter (University of Toronto), Timo Wochner (ETH Zurich & KOF Institute), Matt Woerman (Colorado State University), Jared Wong (Yale University), Ritchie Woodard (University of East Anglia), Marcin Wroński (SGH Warsaw School of Economics), Gustav Chung Yang (Harvard University), Myra Yazbeck (University of Ottawa), Luther Yap (National University of Singapore), Kareman Yassin (Hitotsubashi University), Hao Ye (University of Pennsylvania / Community for Rigor), Jin Young Yoon (Queen's University), Chris Yurris (McGill University), Tahreen Zahra (Carleton University), Mirela Zaneva (University of Oxford), Aline Zayat (University of Ottawa), Jonathan Zhang (Duke University and Sanford School of Public Policy), Ziwei Zhao (University of Lausanne and Swiss Finance Institute), Yaolang Zhong (University of Warwick)

Computational reproducibility:

Abel Brodeur (University of Ottawa and Institute for Replication), Joanne Haddad (Universitat Autonoma de Barcelona), Pu Sun (Dongbei University of Finance and Economics)

Local organizer Replication Games:

Marie Connolly (UQAM), Romain Gauriot (Deakin University), Leonard Goff (University of Calgary), Christoph Huber (Aalto University), Andreas Kotsadam

(Ragnar Frisch Centre for Economic Research), Diego Marino Fages (Durham University)

Figure Legends

Figure 1: Robustness Rate. Legend: Robustness rate for ... **Left panel:** ... originally statistically significant research **Second panel:** ... in economics **Third panel:** ... in political science **Right panel:** ... originally statistically *in*significant research **All panels:** Squares, circles, and triangles represent proportions, with 95% Clopper-Pearson confidence intervals presented in whiskers. Red squares represent full sample. Green circles represent economics subsample. Blue triangles represent political science subsample. Each group of three estimates represent different types of re-analysis, non-mutually exclusive. The first 8 groups do not include re-analyses that use new data (replication), while the last one does. The first estimate group contains all types of re-analysis, then all types of re-analysis in economics, then all types of re-analysis in political science. The second represents re-analyses which changed the control variables, e.g., by adding or re-defining them. The third represents re-analyses which changed the dependent variable, e.g., by employing a different standardization or binarization. The fourth represents re-analyses which changed the estimation method, e.g., by adjusting a matching procedure. The fifth represents re-analyses which changed the inference method, e.g., changed the level on which standard errors are clustered. The sixth represents re-analyses which changed the main independent variable, e.g., by taking into account treatment intensity. The seventh represents re-analyses which changed the sample, e.g., by excluding outliers. The eighth represents re-analyses which changed the weights applied, or applied weights for the first time. The last represents replicability rates for re-analyses that introduced new data, e.g., comparable outcomes from more recent survey waves.

Figure 2. Statistical Significance of Publication and Re-analysis. Legend: **Top histogram:** Distribution of publication tests of significance. T-statistics over 4 truncated for exposition. The histogram's bars 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 re-analysis tests of significance. T-statistics over 4 truncated for exposition. **Scatterplot:** Each marker is a pair of test statistics, an originally published test statistic (horizontal value) and an associated re-analysis test statistic (vertical value). If the original and re-analysis test statistics were identical, this scatterplot would follow the 45 degree line. As either axis represents statistical significance, we have bifurcated each with a line at t=1.96, representing statistical significance at the 5% threshold. **Blue circles** indicate an originally statistically significant statistic that is also statistically significant under re-analysis. Represents 50% of sample. **Red triangles** indicate originally significant test statistics that are no longer statistically significant under re-analysis. Represents 14% of sample. **Green squares** indicate originally statistically insignificant test statistics that are the same under re-analysis. Represents 27% of sample. **Purple diamonds** indicate originally statistically insignificant test statistics that become

statistically significant under re-analysis. Represents 3% of sample. **Not displayed** Not displayed are the 6% of test statistics that represent a sign reversal between the originally estimated effect and the effect estimated under re-analysis.

Figure 3. Robustness Rate Determinants. Legend: Six independent teams answered twelve questions of the re-analysis database. Each bar represents a different question. **Left panel:** "Does reproducibility of an originally statistically significant result depend on..." **Right panel:** "Does reproducibility of an originally statistically *insignificant* result depend on..." **Both panels:** where the first bar represents "... the reproducers' experience at coding." **Blue, patterned outline** indicates the proportion of teams that indicated a negative and statistically significant relationship, in whichever manner the team defined so in their analysis. **Gray, no outline** indicates the proportion of teams that indicated a statistically insignificant relationship, where left of the zero line indicates negative and right of the zero line indicates positive. **Red, solid outline** indicates the proportion of teams that indicated a statistically significant and positive relationship. All teams equally weighted.

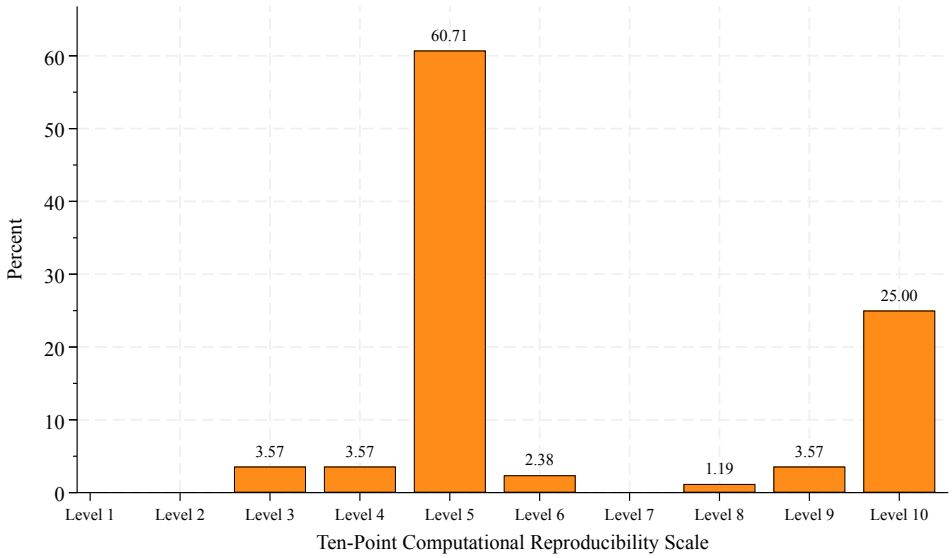
Figure 4. Effect Size of Publication and Re-analysis. Legend: **Top histogram:** Distribution of originally published effect size standardized by the average effect size within a published article. **Right histogram:** Distribution of re-analysis published effect size standardized by the average effect size within a published article. **Scatterplot:** Each marker is a pair of effect sizes, the originally published effect size (horizontal value) and an associated re-analysis effect size (horizontal value). If an originally estimated and re-analysis effect size were of similar magnitude (and sign), the markers would gather tightly around the 45 degree line passing through the origin **Blue circles** indicate effect sizes which are similar (between 50% to 200% of original effect size) under re-analysis. Represents 69% of sample. **Red diamonds** indicate effect size estimates which switch sign under re-analysis. Represents 6% of sample. **Orange triangles** indicate effect size estimates which are 50% or less their original magnitude under re-analysis. Represents 9% of sample. **Purple squares** indicate effect size estimates which are double or larger than their original magnitude under re-analysis. Represents 16% of sample.

Extended Data Figure Legends

Extended Data Figure 1: 10-Point Computationally Reproducibility Score. Legend: 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 Figure 2: Reasons Select Paper? (Select all which apply). Legend: Data collected *via* survey of our reproducers after completing their reports. This

Fig. 5: 10-Point Computationally Reproducibility Score



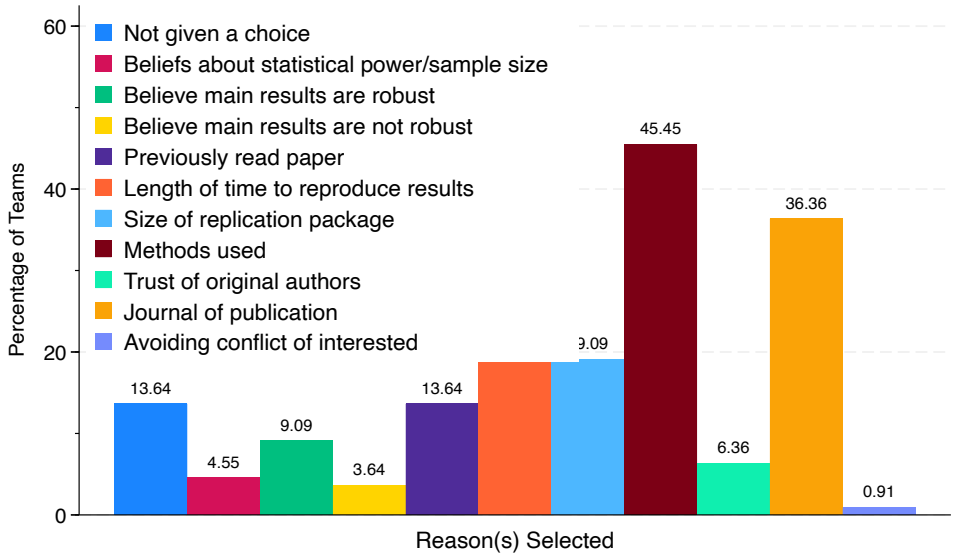
Notes: Each team assigned a reproducibility score on a scale of one to ten to the paper reproduced. See Supplementary Materials for a description of each score. Level 10 (L10) means that all necessary materials are available and produce consistent results with those presented in the paper, while level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

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 Figure 3: Percentage of Papers with a Replication Folder. Legend: 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 Figure 4: Percentage of Papers with a Replication Folder by Discipline. Legend: Panel (a) is for papers published in economics journals where

Fig. 6: For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided? (Select all which apply)



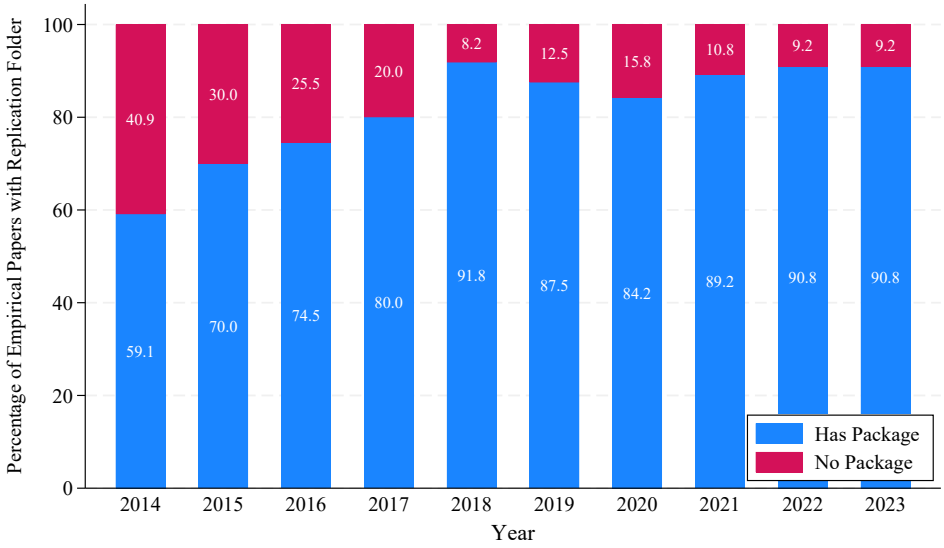
Notes: Data collected *via* survey of our reproducers after completing their reports. This figure illustrates the responses to the question: "For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided?"

Panel (b) is for papers published in political science. The total sample is the same as Extended Data Figure 3 is 1150 papers, where 850 papers are in the economics sample and 300 papers are in the political science sample.

Extended Data Figure 5: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder. Legend: 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 Figure 6: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder. Legend: 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).

Fig. 7: Percentage of Papers with a Replication Folder



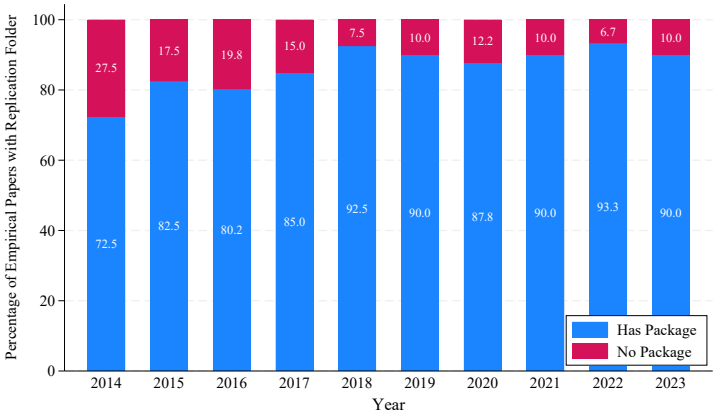
The total sample is 1150 papers with 120 papers per year from 2019 to 2023 and 110 papers per year from 2018 to 2014. Each journal has 10 papers per year except *American Economic Review: Insights* which only formally became a journal in 2019 (and are omitted in earlier years). The journals sampled over 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 Figure 7: Reasons Unable Conduct Robustness Checks. Legend: 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.

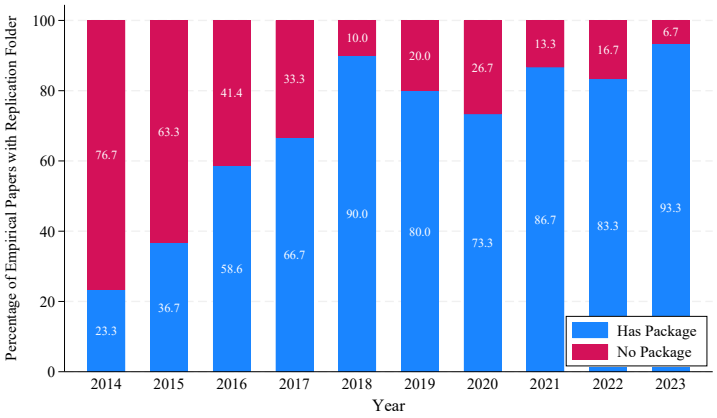
Extended Data Figure 8: Distributions of t-Statistics for Original Studies and Re-Analyses. Legend: The top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left panel includes all original studies in our data set. The top right panel includes all re-analysis 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.

Fig. 8: Percentage of Papers with a Replication Folder by Discipline

(a) Economics



(b) Political Science

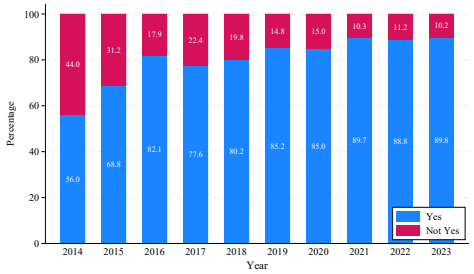


Panel (a) is for papers published in economics journals where Panel (b) is for papers published in political science. The total sample is the same as Figure 7 is 1150 papers, where 850 papers are in the economics sample and 300 papers are in the political science sample.

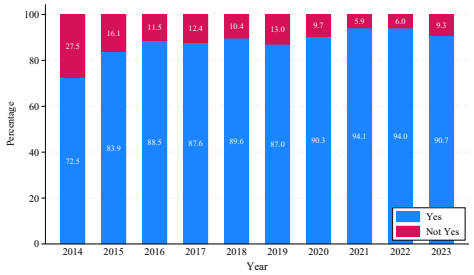
Extended Data Figure 9: Distributions of t-Statistics and p-values by Field. Legend: We restrict the sample to articles published in the indicated field. journals. Top panels display histograms of test statistics for $t \in [0, 5]$, with bins of width 0.1 respectively. Vertical reference lines are displayed 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.

Fig. 9: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder

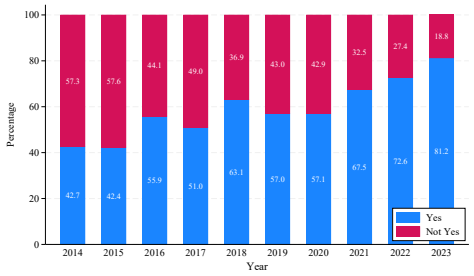
(a) README



(b) Analysis Code



(c) Cleaning Code

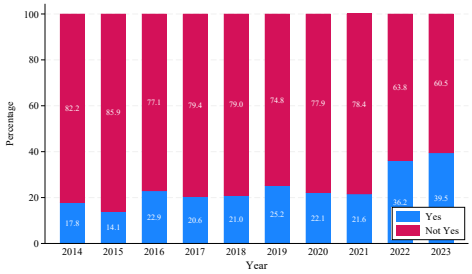


Each subfigure represents the proportion of the replication folders which affirmatively ("Yes") contained the variable (displayed as the title). The "Not Yes" in the legend corresponds to those replication folders which did not affirm ("No") or had only "Some" of the required contents. Each sample is over those observations where categories are applicable (*i.e.* not all replication packages require the same contents).

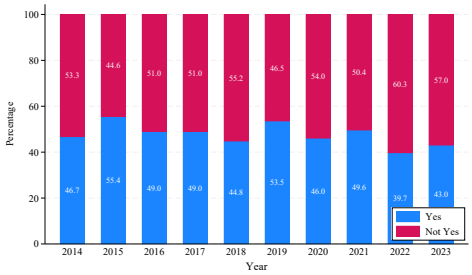
Extended Data Figure 10: Relative Reproduced Effect Size. Legend: 48% of relative effect sizes are exactly equal to or greater than 1. This figure illustrates the ratio of re-analysis estimates and original estimates. The standardized effect sizes are normalized so that 1 equals the original effect size. A positive value indicates that the

Fig. 10: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder

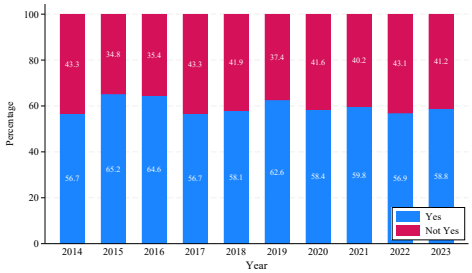
(a) Raw Data



(b) Final Data



(c) Final Data + Raw or Intermediate Data *with* Cleaning Code

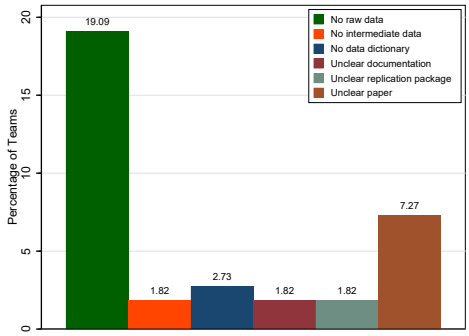


Each subfigure represents the proportion of the replication folders which affirmatively ("Yes") contained the variable (displayed as the title). The "Not Yes" in the legend corresponds to those replication folders which did not affirm ("No") or had only "Some" of the required contents. Each sample is over those observations where categories are applicable (*i.e.* not all replication packages require the same contents).

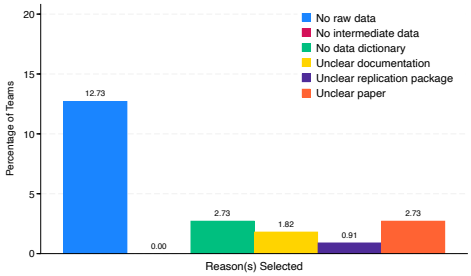
re-analysis estimate is in the same direction as in the original study. A negative value indicates that the re-analysis estimate is not in the same direction as in the original

Fig. 11: For which of the following reasons were you unable to conduct robustness checks, recoding exercises, extensions, or a replication using new data, prior to communications with the original authors? (Select all which apply)

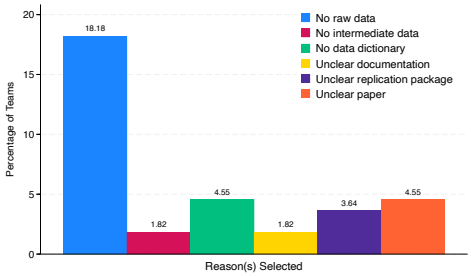
(a) Robustness Checks



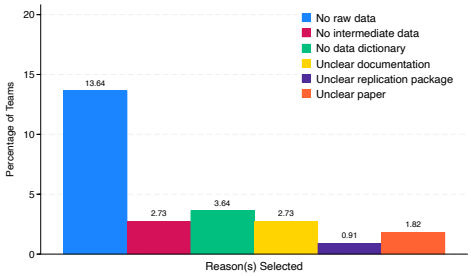
(b) Replications



(c) Recoding key variables



(d) Extensions using Original Authors' Data

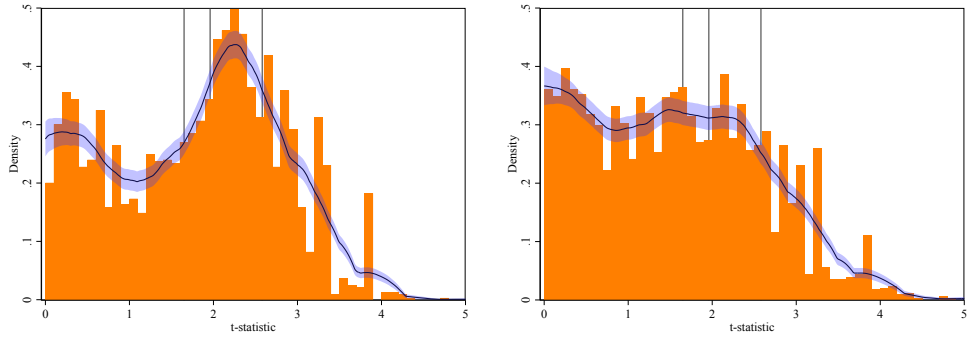


Notes: This Figure illustrates the share of teams who were unable to perform robustness checks (top-left), replications (top-right), key variable recodes (bottom-right) or extensions (bottom-left) for various reasons represented by the different coloured bars.

Fig. 12: Distributions of t-Statistics for Original Studies and Re-Analyses

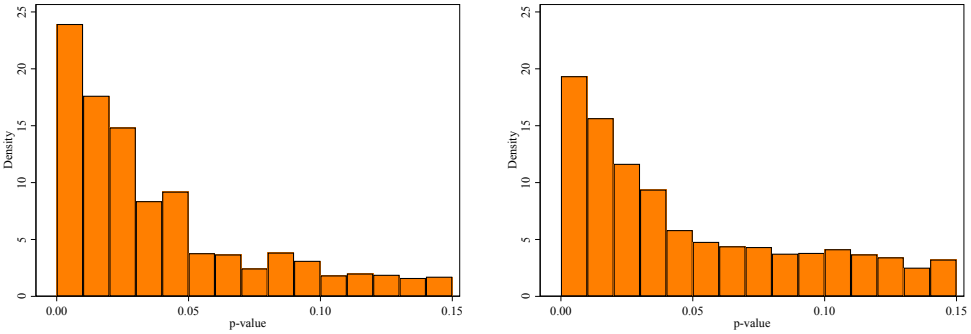
(a) Original Studies - t-statistics

(b) Re-Analysis Studies - t-statistics



(c) Original Studies - p-values

(d) Re-Analysis Studies - p-values



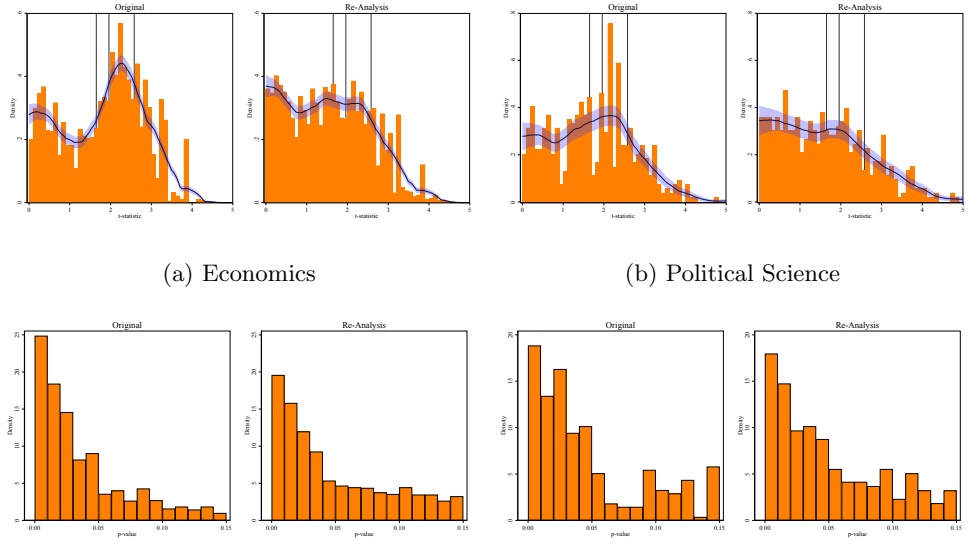
Notes: The top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left panel includes all original studies in our data set. The top right panel includes all re-analysis estimates in our data set. Vertical reference lines are displayed at conventional two-tailed 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.

study. Outliers (3%) are excluded for visibility.

References

1. Vazire, S. Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology* **3,** 1 (2017).

Fig. 13: Distributions of t-Statistics and p-values by Field



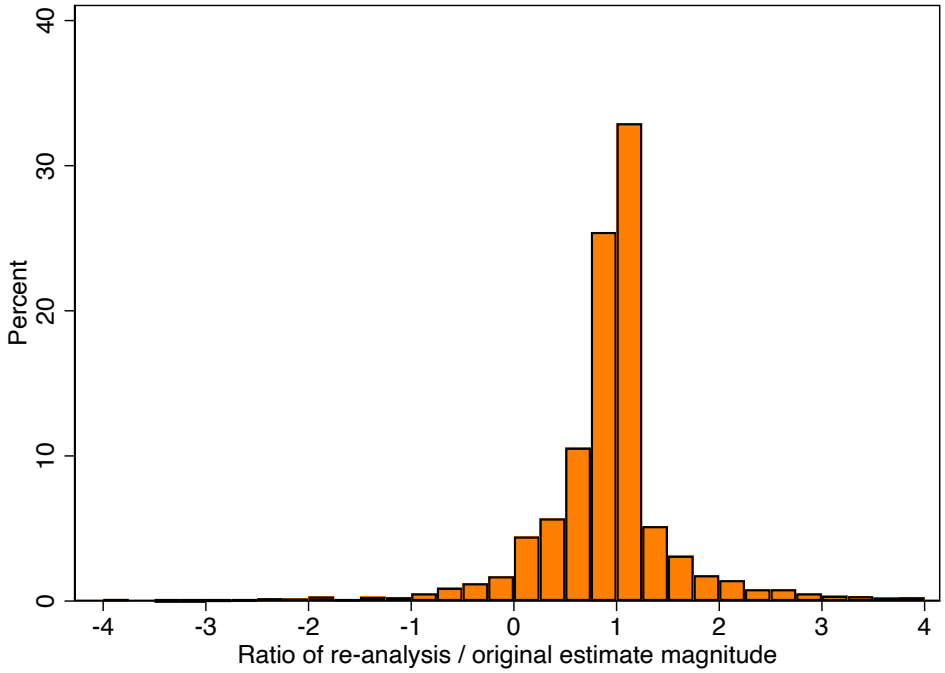
(a) Economics (b) Political Science

(c) Economics (d) Political Science

Notes: 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.

2. Donoho, D. L., Maleki, A., Rahman, I. U., Shahram, M. & Stodden, V. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* **11,** 8–18 (2008).
3. King, G. Replication, Replication. *PS: Political Science & Politics* **28,** 444–452 (1995).
4. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. What does research reproducibility mean? *Science Translational Medicine* **8,** 341ps12–341ps12 (2016).
5. Marcoci, A. *et al.* Predicting the replicability of social and behavioural science claims in COVID-19 preprints. *Nature human behaviour* **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. *Journal of Computational Neuroscience* **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 Medicine* **4,** e28 (2007).
8. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science* https://www.nap.edu/catalog/25303 (National Academies Press, 2019).

Fig. 14: Relative Reproduced Effect Size



Notes: 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.

9. Peterson, D. & Panofsky, A. Self-Correction in Science: The Diagnostic and Integrative Motives for Replication. *Social Studies of Science* **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. *Proceedings of the National Academy of Sciences* **112,** 15267–15268 (2015).
12. Freese, J. & Peterson, D. Replication in Social Science. *Annual Review of Sociology* **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. *Economic Journal* **127** (2017).
15. Munafò, M. R. *et al.* A Manifesto for Reproducible Science. *Nature Human Behaviour* **1,** 1–9 (2017).

16. Nosek, B. A. *et al.* Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology* **73,** 719–748 (2022).
17. Askarov, Z., Doucouliagos, A., Doucouliagos, H. & Stanley, T. The Significance of Data-sharing Policy. *Journal of the European Economic Association* **21,** 1191–1226 (2023).
18. Brodeur, A., Cook, N. & Neisser, C. P-Hacking, Data Type and Data-Sharing Policy. *Economic Journal* **134,** 985–1018 (2024).
19. Chang, A. C. & Li, P. Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review* **11,** 185–206 (2022).
20. Christensen, G. & Miguel, E. Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature* **56,** 920–80 (2018).
21. Dafoe, A. Science Deserves Better: the Imperative to Share Complete Replication Files. *PS: Political Science & Politics* **47,** 60–66 (2014).
22. McCullough, B., McGeary, K. A. & Harrison, T. D. Do Economics Journal Archives Promote Replicable Research? *Canadian Journal of Economics* **41,** 1406–1420 (Nov. 2008).
23. Pérignon, C. *et al. Computational Reproducibility in Finance: Evidence from 1,000 Tests* HEC Paris Paper. 2023.
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. *Nature Human Behaviour* **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. *Economic Inquiry* (2023).
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. *Nature Human Behaviour* **4,** 1208–1214 (2020).
30. Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics* **8,** 1–32 (Jan. 2016).
31. Brodeur, A., Cook, N. & Heyes, A. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* **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. *Proceedings of the National Academy of Sciences* **119,** e2203150119 (2022).

34. Huntington-Klein, N. *et al.* The Influence of Hidden Researcher Decisions in Applied Microeconomics. *Economic Inquiry* **59,** 944–960 (2021).
35. Menkveld, A. J. *et al.* Non-Standard Errors. *Journal of Finance* (Forthcoming).
36. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* **1,** 337–356 (2018).
37. Fišar, M. *et al.* Reproducibility in Management Science. *Management Science* (2023).
38. Ankel-Peters, J., Fiala, N. & Neubauer, F. Do Economists Replicate? *Journal of Economic Behavior & Organization* **212,** 219–232 (2023).
39. Vilhuber, L., Turrito, J. & Welch, K. Report by the AEA Data Editor. *AEA Papers and Proceedings* **110,** 764–75. https://www.aeaweb.org/articles?id=10.1257/pandp.110.764 (May 2020).

11 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, 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.

I4R's Board.

First, I4R has a board of editors who recommend potential reproducers. All board members are nominated by the lead author, A.B. He 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). Games are one-day meet-ups open to faculty, post-docs, graduate students and other researchers. Participants join a small team of about 3–5 researchers all working in the same subfield (*e.g.*, development economics).

11.1 Types of Re-Analyses

We group re-analyses into eight groups: (i) alternative control variables, (ii) change the sample, (iii) change (coding of) the dependent variable, (iv) change (coding of) the main independent variable, (v) change estimation method, (vi) change inference method, (vii) change weighting scheme and (viii) replication using new data.

11.2 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 which are plots of coefficients (i.e., event studies) we encouraged reproducers to give 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.B. and D.M.) tried to subjectively describe if 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 percent—we believe to have display the same result as the original paper and can be reasonably compared.

11.3 Non Comparable Re-Analyses

As mentioned earlier, a direct comparison is not possible between the original analysis and the reproducers' analysis for about 15% of re-analyses. In applied microeconomics and politics papers, this may be due 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 which enter a model, possibly using accepted values in the field or estimated from an alternative dataset.

82 articles have at least one non-comparable estimate. Only a small proportion (10 re-analyses) were not directly comparable for all reported re-analysis estimates. For not directly comparable re-analyses, we report the proportion that reproducers indicated were of the same statistical significance as the original and same sign. 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 re-analysis 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 re-analysis was of the same (non)significance (92% for the 10% level).

11.4 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 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.

11.5 Journal Policy

The *American Journal of Political Science* does not have a data editor. Instead, the computational reproducibility is carried out by the staff at the Odum Institute for Research in Social Science, at the University of North Carolina, Chapel Hill. The journals which 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 internally.

Data editors make sure that the replication packages include the data and codes, and that the documentation (e.g., Readme) 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 data set(s). Their teams also run the codes and make sure that the output is similar to what is reported in the article. They do not look for coding errors nor run robustness checks.

11.6 Many-Analysts Approach

Our approach and research questions, which we detail below, were pre-registered. Our pre-analysis plan was pre-registered here: https://osf.io/8wsqx/. The pre-analysis plan was pre-registered prior to sharing the Meta Database with analysts. See the SI for 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?"

11.6.1 Data for Analysts

Analysts were not given access to raw data (database, team leader surveys, individual surveys). Rather, they were given access to intermediate/analytical data which was cleaned and merged in a manner which would be consistent for their analysis. Giving researchers a downstream dataset allowed A.B. and D.M. to make restrictions on what the analysts could do. The clearest example of this would be defining dependent variables which 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, of which questions from the team leader surveys and individual surveys were added. Much of the information from the individual surveys were 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 November 6th, 2023 and February 12th, 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, *etc.* 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.

11.6.2 Team Construction

We asked a subset of coauthors on this paper (reproducers) if they would like to help analyze our database. We informed them that we would "have different teams independently working together at answering the same research questions (e.g., 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 September 21st and October 8th 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 September 21st and October 8th. 10 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; (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 timezones. We had two teams of three, one team of two, and two individuals. A.B. 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 PI ultimately provided each volunteer with a payment of $3,000 CAD, this compensation was not disclosed or anticipated at the time they agreed to participate.

11.7 Database: Sample Composition

The database described above provides 6,583 re-analyzed test statistics from 103 reproduction reports. (Seven reports did not include robustness checks.) The other test statistics are estimates obtained by re-coding the analysis.

Supplementary Materials Appendix Table 11 provides summary statistics for the full sample and by journal. In total, 83 reproduction reports were completed through Games in comparison to 27 through the editorial board stream. 79 reproduction reports are for the field of economics against 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 re-analysis is statistically significant at 5% level and same sign. For this dependent variable, we only keep original estimates statistically significant at the 5% level.
Second Dependent Variable: dummy variable indicating whether the re-analysis is statistically significant at 10% level and same sign. For this dependent variable, we only keep original estimates statistically significant at the 10% level.
Third Dependent Variable: dummy variable indicating whether the re-analysis 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 re-analysis 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 re-analyzed test statistics per article is about 60. The standard deviation 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), while other teams have 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 five different robustness checks for each outcome variable and specification, then this would lead to 30 re-analyzed 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 Materials Appendix 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 (*i.e.*, years since completing their Ph.D.) 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.

While 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.

11.8 Computational Reproducibility

We rely on the Social Science Reproduction Platform (SSRP)'s 10-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 (Visit https://bitss.github.io/ACRE/assessment.html#score for 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, while 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 data sets and analysis code are available, but the code fails to run or produces results inconsistent with the paper (not CRA). Possible improvements include: debugging the analysis code or obtaining raw data.
Level 5 (L5): Analytic data sets 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 is missing. Possible improvements include: adding raw data.
Level 7 (L7): Cleaning code is available and complete, and raw data is 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 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 Figure 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 (L10) means that all necessary materials are available and produce consistent results with those presented in the paper. Level 5 (L5) means that analytic data sets and analysis code are available, and they produce the same results as presented in the paper. In other words, L5 indicates that the reproducers successfully (computationally) reproduced the numerical results using

the analytical data, but the raw data were not provided, while L10 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 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 [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.

11.9 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 (e.g., different types of Random Number Generations, rounding rules and numerical precision). We categorized recoding exercises done by reproducers into three categories: (i) identical numerical results, (ii) minor differences, and (iii) 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. In contrast, major differences lead to major differences in one or multiple claims.

11.10 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 not impacting the claims or conclusions of the article.

11.11 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 labor 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 a) was not empirical, b) was not a standard article (we excluded comments, replies and corrections, retraction notices, and editor notes, even if they

were empirical in nature), c) was a duplicate that had already been considered (e.g., issue number one, article number five, was drawn twice in a row), or d) 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 five years earlier, we did not collect for this journal, leading to 10 fewer papers per year.

Methods references

13. Gertler, P., Galiani, S. & Romero, M. How to Make Replication the Norm. *Nature* **554,** 417–9 (2018).
19. Chang, A. C. & Li, P. Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review* **11,** 185–206 (2022).
37. Fišar, M. *et al.* Reproducibility in Management Science. *Management Science* (2023).
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).

SI: Reproducibility and Robustness of Economics and Political Science Research

Abel Brodeur et al. (Author list and contributions are provided in the SI)[1*]

[1*]Department of Economics and Institute for Replication, University of Ottawa, 75 Laurier Avenue East, Ottawa, K1N 6N5, Ontario, Canada.

Corresponding author(s). E-mail(s): abrodeur@uottawa.ca;

Methods and Additional Results

Replication Games

So far, teams have been as small as one individual or as large as seven.The locations of Games are chosen based on (i) local interests, (ii) geography, (iii) possibility to have the Games as part of a major conference, and (iv) EDI considerations.

I4R groups graduate students with faculty members and senior researchers, ensuring a mix of junior and more senior economists in each team. A virtual meeting with the organizers before the Games allows each team to ask questions and discuss a game plan. During the Games, A.B., D.M. or one of I4R's co-directors, provide live assistance to the teams.

Participants are offered a short list of (about 5) studies in their field of interest about three weeks before the Games. They are asked to choose a paper as a team, read it and familiarize themselves with the replication package prior to the Games.

Teams are asked to develop a game plan for the Games; each team member should know what they are supposed to do during the Games. A virtual meeting with the organizers before the Games allows each team to ask questions and discuss a game plan. During the games, A.B., D.M. or one of I4R's co-directors, provide live assistance to each team. Teams then have to write a (templated - https://osf.io/8dkxc/) report summarizing their work and results in the following months. Of note, virtually all teams kept working on their reproduction after the Games and some even started the re-analysis prior to the Games.

Participants are offered the possibility to virtually attend Games. In our sample of completed reports, about 68% of participants attended the games in-person, while 32% virtually attended the events. Most teams are fully virtual or in-person, with only

a small share of teams having a mix of virtual and in-person participants. Mixed teams are typically due to a variety of reasons (*e.g.*, canceled flight for one participant), or late registrations.

We asked a subset of games participants the following question: "Why did you choose to participate in the Replication Games?" We offered seven potential options, with an empty box to provide additional reasons. We find that a majority of respondents chose the responses "Learn about academic replications and reproductions", "Expand your network", and "Contribute to Open Science". Other popular responses include "Improve your ability to program and code" and "Improve your ability to conduct research".

Teams have on average worked 13 active days on their reproduction (std. dev. of 24). Supplementary Materials Appendix Figure 1 shows the distribution of days across reports, trimmed at over 100 days.

About half the teams worked from 5 to 20 days on their reproduction report. Most of the remaining teams worked between 25 to 85 active days. A very small fraction worked less than 5 days. This is due to the reproducers not being able to conduct robustness checks. In contrast, about 8% of teams worked more than 100 days. This is typically due to uncovering major coding errors or issues with the original study and having to engage in multiple rounds of back and forth with the original authors. There is also the potential for people to have spent many days on their paper even if the number of hours were low. Reports are on average 19 pages long, with a standard deviation of 14.

In terms of retention for the Replication Games, over 90% of registered participants ended up participating in the event. Furthermore, within one year of completing the first two replication Games (October and November 2022), 85% of teams had completed a report.

The goal for all reproducers is clearly stated; testing whether the main claims are reproducible and robust. I4R emphasizes to reproducers that the goal is NOT to show that the results are not reproducible. The goal is instead to test if the results are reproducible to recoding and/or robustness checks. This is key as some reproducers might engage in reverse specification searching (i.e., selective reporting of insignificant results). Moreover, we ask reproducers from I4R's Board stream to provide a pre-re-analysis plan. The game plan acts as a pre-re-analysis plan for the second stream.

In practice, some teams in both streams did not write a pre-re-analysis plan and virtually all teams that did write one ended up deviating from it. The latter is because it is very unclear from only reading the original paper what is the range of re-analyses that is feasible. Reproducers had to carefully look at the replication package provided by the authors to gauge whether specific robustness checks were implementable given data availability. Our re-analyses should thus all be considered as not pre-registered.

Supplementary Materials Appendix Table 2 provides summary statistics.

Types of Re-Analyses

One of our main objectives is to document the relative importance of several robustness checks and re-analyses in impacting the magnitude and significance of the original

point estimates. We group the robustness checks and coding exercises conducted by the reproducers into eight groups.

Alternative control variables: Removing, adding or changing control variables. In our sample, there are 1,939 new re-analyses involving alternative controls.

Change the sample: Decreasing or increasing the sample size. In our sample, there are 1,774 new re-analyses involving changing the sample size. Reproducers may change the sample by adding/removing years, geographical units or individuals. For instance, a team could check if the results are robust to adding/removing a state to/from the analytical sample.

Change (coding of) the dependent variable: The reproducers may change the coding of the dependent variable. In our sample, there are 285 new re-analyses involving changing the dependent variable. Examples include using an alternative standardization of the outcome variable, alternative calculation of factor loadings for an index in a different software, using an indicator variable for an outcome above a certain threshold, and using a composite index of several indicators as the dependent variable.

Change (coding of) the main independent variable: The reproducers may change the coding of the main independent variable. In our sample, there are 264 new re-analyses involving changing the main independent variable. Examples include using a continuous variable instead of a dummy or a factor variable for treatment, broadening the definition of treated based on physical proximity, using a different type of TV program in television exposure, and allowing for non-linear effects in institutional exposure.

Change estimation method: This category involves any changes to the estimation method. In our sample, there are 605 new re-analyses involving changing the estimation method. Examples include using non-linear models and changing the variables used for matching.

Change inference method: This category involves changing the inference method. In our sample, there are 542 new re-analyses involving changing the inference method. Examples include bootstrapping the standard errors and clustering at a different level.

Change weighting scheme: This category involves changing the weighting scheme. In our sample, there are 126 new re-analyses involving changing the weighting scheme. Examples include removing a weighting scheme used by the authors.

Replication using new data: Replication using new data involve both collecting new data or using data from another data source. In our sample, there are 469 new re-analyses involving using new data. Replicators have used new data for the dependent, independent or control variables.

In practice, many reproducer teams performed multiple robustness checks *simultaneously* in a single robustness exercise, or, combined two independent robustness checks into a new, third robustness check. We tracked all the changes reproducers made when comparing to an original estimate and coded accordingly. In our sample, about 809 re-analyses fall into at least two categories of simultaneous robustness checks.

Supplementary Materials Appendix Table 3 provides a decomposition of reports and test statistics by type of re-analyses. The most popular re-analyses involve using alternative control variables and changing the sample. In contrast, only 14 reports had any robustness check which changed the weighting scheme and only 15 reproduction reports had any robustness checks which used new data.

The types of re-analyses are quite similar for economics and political science. Using alternative control variables, changing the sample and changing the estimation method/model are among the most popular re-analyses for both fields. One noticeable difference is that reproducers are more likely to change the method of inference for economic articles than in political science.

Database

In what follows, we describe our database. The database is mainly built from three sources of raw data: (1) reports; (2) surveys of individual reproducers; and (3) surveys of teams of reproducers. We also collected information from publicly available *curricula vitae* of all original authors and reproducers.

Reports: Two of the lead authors (A.B. and D.M.) and research assistants read reports and copied test statistics into an Excel file. We also coded and grouped robustness and replicability exercises, and information on computational reproducibility and coding errors. The work being entered by RAs was checked by A.B. or D.M. for completeness and accuracy. If any part of any entry was unclear, they were checked again and discussed.

Only a subset of results was considered suitable for our research. We follow the following criteria. We exclude extensions of the original authors' research, effects by heterogeneity, or mediation analyses. These analyses correspond to situations where there are no "original" estimates to which we can reasonably compare the reproducers' estimates. Most often, reproducers included tables and figures which were the output of a computational reproduction using the original authors' replication package. These are always left out for the re-analyses. After being checked, reproducers would then be contacted with their subset of the database and asked to confirm our transcribing of their reports into the database.

Coding errors and discrepancies are also excluded from the re-analyses. We discuss coding errors and discrepancies between original authors' values in their published paper compared to what their replication package produces in the Methods Section.

We report some additional information in our database. We collect information on the journal, year of publication, number of Google Scholar citations at the time of entry into the database, the research field, the position of the test in the original article and the number of original authors and reproducers. We also collect information from *curricula vitae* of all the original authors and reproducers. We obtained information on their academic affiliation at the time of publication, their position at the main institution and year the PhD was earned. In addition, we gather for each author and reproducer the following information (at the time of completing the replication): the total number of Google Scholar citations and whether they had published in a Top-5 economic journal, a leading political science journal, and/or one

of the other economic journals we are reproducing/replicating. The Top-5 economic journals are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics* and the *Review of Economic Studies*. The leading political science journals considered here are the *American Journal of Political Science*, *American Political Science Review* and *Journal of Politics*.

Surveys: We asked all reproducers to fill out an individual survey. We also asked one author per reproduction report to fill out a team survey. Both surveys gave additional information on the academic and programming experience of reproducers, how long their report took to create and the completeness of the original authors' replication package, and whether they improved it. Teams were invited to answer the surveys following the completion of transcribing their report.

The team survey provides additional information on data availability, computational reproducibility, the reasons the paper to be reproduced/replicated was chosen, how long it took to run the code provided in the replication package, reasons they were unable to conduct specific robustness exercises, *etc.* We also asked whether there was any communication with the original authors for clarifications and how it improved the quality of the report.

The individual survey also provides us information about whether the reproducers participated in the Games, whether they virtually attended, why they participated in the Games, and their general experience, and how it improved their networking and coding skills. We conclude the individual survey with subjective questions such as "How does the quality of the replication package affect your view of the discipline as a whole?"

Communication with Original Authors

Once a report is completed, A.B. reviews it if it falls within his expertise. Otherwise, someone else on I4R's board reviews the report. This review involves checking the tone and structure of the report. A.B. then shares the report with the original authors. A.B. emailed all the original authors unless there were more than 5 authors. A reminder was sent a few months later if the original authors did not respond to the initial email. If the authors did not respond to the reminder, the report was released after 6 months.

Reproducers may change their report after receiving the original authors' response, allowing them to include their feedback. This is especially important if a re-analysis was judged unreasonable. I4R then allows the original authors to change their response as well. Of note, the reproducers may remain anonymous. In practice, about 11% of reproducers have decided to remain anonymous.

Original authors have been incredibly fast at providing a response, perhaps since papers being reproduced have just been published. See [1] (pages 133-244) for a link to authors' responses.

In some instances, original authors requested to see the reproducers' replication package, which we provided. See Supplementary Materials Appendix Table 4 for a breakdown by discipline.

How often do reproducers and original authors agree? This is a key question as reproducers have freedom to conduct any recoding or sensitivity analysis. This freedom might lead to disagreement on the validity of some re-analyses. We document (dis)agreements in multiple ways. First, authors' final responses (i.e., post-mediation) were coded as whether there remained disagreements between authors and reproducers. The coding was done by A.B. and three ambiguous cases were discussed at length with D.M.

Overall, we find that there are remaining disagreements for only 23% of articles in our sample. This percentage goes up to over 75% if we restrict the sample to articles for which the original authors wrote a formal response, suggesting that the majority of formal responses we obtained include some sources of disagreements. Disagreements are mostly due to the validity of the re-analyses. There were no remaining disagreements on the presence of coding errors, but authors and reproducers sometimes disagreed on their importance. Disagreements on the scope of the re-analyses and definition of reproducibility were quite rare, and there were also disagreements involving the tone or interpretation of the re-analyses/errors.

We observed a general lack of adversariality between original authors and reproducers [2]. The broad lack of adversariality is potentially due to the high rates of reproducibility and replicability, but also perhaps on the institutionalization of replications and the fact that discussion between original authors and reproducers is mediated by the I4R. Moreover, original authors may feel less targeted by our reproducers as our aim is to mass-reproduce and replicate studies published in leading economic and political science outlets.

We asked reproducers whether their team or I4R contacted, or attempted to contact, the original authors for clarifications. About 40% responded "yes". About 10% reached out because the replication package was unclear, while 17% needed help to computationally reproduce the original authors' results. Another 17% were unable to access the original authors' data. Other reasons include verifying coding errors, clarifications about the design model parameters or other coding decisions.

Study Selection

As a benchmark, A.B. investigated whether studies published at the *Journal of Development Economics* (JDE) using publicly available data complied with the journal's mandatory data sharing policy. He manually checked the presence of a replication package on JDE's website for all articles published in four volumes in 2022. Out of 75 studies, 47 did not provide a replication package or mentioned that data and codes will be made available upon request. The remaining 28 studies can be categorized as follows: 13 report relying on confidential data; 14 provided a link to a replication package; and one provided only Stata codes and information on how to obtain the data. He then contacted (through I4R's email) all authors who did not provide a replication package. Seven ended up providing a package. Some authors mentioned that they did not know that the policy existed. A few mentioned that they shared the replication materials with JDE and were surprised that it was not posted.

We explore the reasons why 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 (*i.e.*, 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.

Supplementary Materials Appendix Table 5 explores 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 IDEAS/RePec). This comparison benchmark comes from [3]. A comparison of the two samples suggest that some subfields are under-represented. Our sample under-represents, among other fields, C-Mathematical and Quantitative Methods, G-Financial Economics and F-International Economics.

Many Analysts: Additional Results

Each row in Supplementary Materials Appendix Table 6 represents one of the eight research questions. The four columns represent four broad categories regarding research teams' coefficient estimate(s) to the research question: (1) negative and statistically significant, (2) negative and not-statistically significant, (3) positive and not statistically significant and (4) positive and statistically significant. The left-to-right order of the column categories corresponds to where the associated analyst t-statistic would fall on the real number line. While the dependent variable (which does not change in this table) is the same for each team, each team chooses their own primary independent variable. Each cell represents the proportion of analyst-estimated relationships by category. The cells are team-weighted so that if a many-analyst team presents three estimates and another team presents a single estimate, the first team's estimates enter the proportion as 1/3 each.

The cell in the first row and first column tells us that 42.8% of results from the many-analysts find a negative and statistically significant relationship between the coding experience of a reproducer and the reproducibility rate for estimates that were originally statistically significant at the 5% level (i.e., lower reproducibility rate for more experienced reproducers).

From the second column, it becomes clear that, if there is a relationship between reproducers experience coding and the reproducibility rate, it seems to be almost

definitively negative with a combined proportion of 86% of results returned as negative and statistically significant or negative and not statistically significant at the 5% level. Only 14% of estimates find a positive relationship between the reproducers' experience coding and the reproducibility rate - of which none of the estimated positive relationships estimated were statistically significant. (The associated row in Supplementary Materials Appendix Table 7 , which looks at the replication for the 10% threshold finds the same pattern.) This result potentially suggests that reproducers with more experience coding are better suited to detecting and correcting less-than-robust estimations - possibly because of having greater expertise with the methods used.

Supplementary Materials Appendix Table 6 presents results where the dependent variable takes a value of one if an originally 5% statistically significant result was reproduced also at the 5% level. Supplementary Materials Appendix Table 7 has the same structure, but uses the 10% threshold. Supplementary Materials Appendix Table 8 then examines whether an originally *not* 5% statistically significant result was reproduced, while Supplementary Materials Appendix Table 9 continues this with the 10% threshold.

For the second research question - whether the reproduction robustness rate depends on the reproducers' academic experience, a somewhat similar albeit less starkly negative result is found with some proportion moving into the positive and statistically significant category. That said, the ratio of negative-and-significant results to positive-and-significant results remains above 4 to 1. The associated row in Supplementary Materials Appendix Table 7, which looks at robustness for the 10% threshold finds the same pattern, although with 75% of many-analysts results being negatively signed.

For the third research question - whether the replication rate depends on the author's experience seems to be centered on the null. Combined, the negative and not statistically significant and the positive and not statistically significant cells contain 97.2% of results. The null hypothesis dominates in Supplementary Materials Appendix Tables 7,8, and 9 (which examine reproducibility rates for originally statistically significant at the 10% level, not statistically significant at the 5% level, and not statistically significant at the 10% level, respectively) as well.

For the fourth research question, (which has three sub-questions depending on the relative hierarchy of reproducer and original author experience) there seems to be a positive relationship when authors have more or the same level of experience as the reproducer (research question 4a and 4b). This relationship, however, weakens to a likely null when authors have comparatively less experience than their reproducers. Supplementary Materials Appendix Tables 7, 8 and 9, find similar patterns.

For the fifth research question, which has the same comparative structure as the fourth while focusing now on the relative prestige of the authors and reproducers, the same (albeit weaker) pattern is found. When authors have more prestige than their reproducers, there is a very positive relationship with replication rate. When original authors and reproducers have similar prestige levels, this relationship becomes much more likely to be a null (since the middle two columns so outsize the outer two columns). When the authors have less prestige than the reproducers, then the

relationship seems to be negative: 22% finding a negative and statistically significant relationship. In Appendix Table 7, we see the same pattern. When examining replication rate of originally statistically insignificant results, the null hypothesis dominates.

The null hypothesis seems to dominate for the final three research questions, with statistical significance not being achieved in either direction for more than one-sixth of the teams' analyses. This means that the replication rate does not seem to have a relationship with whether the authors provided raw data (research question 6), both raw and intermediate data (research question 7) or cleaning codes (research question 8). See Supplementary Materials Appendix Tables 7, 8, and 9 as well.

In Supplementary Materials Appendix Table 10, we reproduce the analyses in Supplementary Materials Appendix Tables 6, 7, 8 and 9 while only including estimates if the analyst team indicated that, in their opinion, the estimated effect size was economically meaningful. Results are broadly consistent with those described above without the restriction.

The results may reflect that our focus is on journals with a data and code availability policy. The provision (or not) of raw data, intermediate data, or cleaning codes, may thus be due to data type rather than selective data/code provision by original authors. Our results are consistent with [4] who document no relationship between the presence of a data and code availability policy and the incidence of p-hacking, including for research leveraging harder-to-access (e.g., administrative) data. They also document a statistically insignificant relationship between voluntary provision of data by authors on their homepages and selective reporting.

Replication Packages and Expectations

In an assessment of reproducers' expectations regarding the quality of replication packages, we ask reproducers the following question in the individual survey: "Which of the following best describes how the replication package aligned with your expectations". We find that more than half of reproducers report that the replication package aligned reasonably with expectations, and an additional 26% of reproducers indicated that the replication packages exceeded their initial expectations. Fewer than 10% report that the replication package was worse than expected, possibly indicating that for this small proportion of reproducers, the provided materials did not meet the anticipated quality standards or may have lacked certain elements critical for an effective reproduction process. Overall, we find it encouraging that most reproducers found that the provided materials exceeded or aligned well with their initial expectations.

Re-Analyses, P-Hacking, and Publication Bias

t- and p-Curves: We present both t-statistics and p-curves in Supplementary Materials Appendix Figure 2. The top left panel provides the distribution of t-statistics from the *originally* published estimates. We restrict the visualization to $t \in [0, 5]$, present bins of width 0.1, and present an Epanechnikov kernel (with standard errors in blue, along with renormalization at zero) which softens valleys and peaks. We provide reference lines at the conventional two-tailed significance levels. Roughly 60%, 50%, and

25% of test statistics are significant at the 10%, 5% and 1% levels, respectively. We note especially that the distribution exhibits a peak (global maximum) just above the two-star statistical significance threshold of $t = 1.96$ and a valley before the one-star statistical significance threshold between $t = 1.0$ and $t = 1.65$. We take this as our first piece of evidence that the original studies in our sample suffer from (marginal) p-hacking and publication bias. The bottom left panel provides the equivalent p-curve for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025. We have removed $p < 0.0025$ (for a two-tailed test this is roughly $t = 3$) for illustrative purposes only, as inclusion of that mass in the left-most bar of the p-curve leads the resolution of the remaining bars to be quite low. We note that, much like the peak after $t = 1.96$ and the valley just before, the p-curve exhibits a too-tall bar just to the left of the $p = 0.05$ threshold. Whether interpreted through the t-curve or p-curve, we consider this to be our first piece of evidence that the sample of original studies suffers from some form of p-hacking and publication bias.

We present t-and-p-curves using data from [5] in the right panels to serve as a benchmark with which to compare the original studies. The top right panel presents the distribution of t-statistics associated with hypothesis tests from articles published in 25 leading economics journals in 2015 and 2018. These articles rely on one of four popular identification methods (i.e., difference-in-differences, instrumental variable, randomized controlled trials, and regression discontinuity design). Overall, the distribution from our original studies sample is similar to that in [5], although with visually markedly more bunching around the 5% significance threshold.

This could be due to at least three reasons. First, the extent of p-hacking and publication bias might be larger in our sample. Second, reproducers might focus on the most central claim(s) in original studies, while [5] focus on all claims. Arguably, the central claim(s) could be more p-hacked or suffer from more publication bias. Third, reproducers might choose to reproduce studies finding an effect or focus on replicating claims that reject the null hypothesis.

Extended Data Figure 8 directly compares the distribution of test statistics for original studies and our re-analyses. Just as in Supplementary Materials Appendix Figure 2, the top panels present t-distributions while the bottom panels present p-curves, and the left panels present the original studies while the right panels now present statistical significance for the re-analyses. (See Supplementary Materials Appendix Figure 3 for the weighted distributions. For the re-analyses, we use the inverse of the number of test statistics presented in the reproduction report to weigh observations.) We use this visual analysis to test whether re-analyses are less likely to reject the null hypothesis than their original counterparts. If they are, we would expect to see less of a peak (global maximum) just beyond the 5% statistical significance threshold and a shift in the mass of test statistics leftward to the statistical insignificance region, i.e., if re-analyses 're-distribute' the mass of test statistics without (or with less of) the distorting effects of publication bias or p-hacking.

Our findings are striking. Moving from left to right in the top panels - from the original to the re-analysis test statistics - there is a large shift in the mass of test statistics from the *just* statistically significant at the 5% level region to the statistically

insignificant and 10% significance regions ($[0.10 > p > 0.05]$). We note this following the global maximum has shifted in mass into where the valley was, and noting also the much greater mass where $t = 0$. This visual result suggests that re-analyses decrease the statistical significance of many originally published test statistics. This is confirmed by a Kolmogorov–Smirnov test which rejects the null of equality of distributions ($p < 0.000$). A similar result emerges from visual inspection of the bottom panels which display the same statistical significance distributions using p-values. An over-abundance of just statistically significant results here is reflected in a particularly large bar just to the left of $p = 0.05$. Under the assumption of no p-hacking and publication bias, the p-curve should be non-increasing - this particularly large bar is too large. We note that, in the same manner as the t-statistics no longer displaying a marked peak once they have been re-analyzed, the p-curve resulting from re-analysis is much better characterized as non-increasing (particularly at the statistical significance thresholds).

The top panels of Extended Data Figure 9 reproduce the top panel of Extended Data Figure 8 for economics and political science while the bottom panels of Extended Data Figure 9 reproduce the bottom panel of Extended Data Figure 8. A reduction in the peak of t-statistics or a reduction of the p-value bar just to left of $p = 0.05$ can be seen for both economics and political science.

Supplementary Materials Appendix Figure 4 extends the visual analysis by offering a direct comparison of the statistical significance of an original estimate and its corresponding re-analysis. Depicted is a histogram of $(p_{\text{replication}} - p_{\text{original}})$ with bars of width 0.05. Interpretation of this difference-statistic is as follows. If the original estimate and its re-analysis have very similar p-values, then the difference-statistic will be close to zero. If the re-analysis p-value is high (indicating statistical insignificance) while the original p-value is low (indicating statistical significance), then this difference-statistic will add to the right tail of the distribution. Notably, this is what we see—a large proportion of re-analyses find similar p-values as the original (represented by both tall bars just above and just below zero), while we also see that the right tail (which indicates re-analyses finding a lower statistical significance on average) being much thicker than the left tail (which indicates an original study finding a lower statistical significance than its re-analysis). This trend is robust to weights and is present in economics as well as in political science (second through fourth panels of Supplementary Materials Appendix Figure 4.

So far, we have not distinguished between re-analyses that find an effect in the same versus opposite direction as the original estimate. This is potentially problematic if a large fraction of re-analyses finds a significant effect in the opposite direction. In Supplementary Materials Appendix Figure 5 we make this distinction. Whenever the re-analysis estimates an effect that is in the opposite direction, we assign the t-statistic (top panels) or p-value (bottom panels) a negative value. We see that the statistical significance of an original estimate with a re-analysis with an oppositely-signed effect are often statistically significant. There is also still positive t-statistics, highlighting the mass peak's disappearance when moving from original to re-analysis.

Overall, our graphical analysis suggests that re-analyses can lead to both increases and decreases in statistical significance, although the average effect is a reduction. In

all cases, there appears to be a downward shift of an over-abundance of just marginally significant test statistics at the 5% level to the less and not statistically significant regions.

Supplementary Materials Appendix Table 1 explicitly presents the change in statistical significance from the original to a re-analysis at the test-statistic level. See the main text for a description of this table.

Formal Tests for P-Hacking and Publication Bias: We next formally document how re-analyses display a markedly different presence of p-hacking and publication bias. We first rely on caliper tests ([6]) which analyze test statistics within a narrow range slightly above and below a statistical significance threshold. The rationale behind this approach is rooted in the assumption that in the absence of manipulation, be it due to publication bias or p-hacking, we would anticipate a comparable frequency of test statistics falling just below a significance threshold and those falling just above it.

We estimate probit models where the dependent variable is a dummy variable that takes the value one if a test statistic is statistically significant at the 5%-level, and zero otherwise:

$$Pr(Significant_{pr} = 1) = \Phi(\alpha + \lambda Reanalysis_{r}) \tag{1}$$

where $Significant_{pr}$ is a dummy variable for whether p-value p in report r is statistically significant at the 10%, 5% or 1%-level. We rely on probit models throughout and present the average marginal effects and associated standard errors clustered at the report-level. The variable of interest is $Reanalysis_{r}$, which represents a dummy variable that takes a value of one if the p-value is associated with a re-analysis, and zero if it is associated with the original publication.

The estimates are reported in Supplementary Materials Appendix Table 13 for the 5% significance threshold. In column 1, we restrict the sample to [0.05±0.04]. The other columns repeat the specification in column 1 but with narrower bandwidths. We find that re-analysis test statistics are about 10 percentage points less likely to be statistically significant than an originally published test statistic. See Supplementary Materials Appendix Table 14 for the 10% threshold. The point estimates for the 10% level are similar, albeit less statistically significant.

We then rely on an application of [7]. The results are presented in Supplementary Materials Appendix Table 15. The columns μ, τ, and df represent the model's estimated parameters (using an underlying t-distribution and symmetric sign probabilities). The fourth column $[0, 1.645]$ presents the relative publication probability for a t-statistic in the $[0, 1.645]$ interval compared to one in the reference interval of $(2.576, \infty)$.

We find that a not statistically significant 'original analysis' test statistic is 17.16% as likely as a very statistically significant test statistic to be observed (published). Similarly, for the $(1.645, 1.96]$ interval, the original analyses offer only a 38.29% relative publication probability. These findings suggest that original articles in our sample suffer from severe publication bias. As a comparison, we estimate that the same relative 'publication' probability for our re-analyses. This comparison serves only as a

benchmark since re-analyses are not submitted for publication and thus do not suffer from publication bias. Nonetheless, we see this comparison as insightful. We find that the relative 'publication' probability for a re-analysis jumps to 27.31% from 17.16%. This trend continues for the $(1.645, 1.96]$ interval, where we observe a 64.30% relative publication probability in a re-analysis versus 38.29%. For the relative publication probability of test statistics significant at the 5% level, the original analyses offer an almost equal probability of 107.40%, whereas the re-analysis is now slightly lower than the original at 89.94%.

The second and third panels offer a similar analysis for the economics and political science subsamples, respectively. The economics subsample behaves similarly to that of the full sample. The political science subsample behaves similarly, with the exception of the not statistically significant interval where the original analysis is more likely to have not statistically significant result published.

We adopt diverse methodologies introduced by [5] and [8] as our foundation. Our initial focus is on randomization tests, as designed by [5] to affirm the visually apparent discontinuities near conventional statistical thresholds. We assess whether the concentration of test statistics just above versus just below these thresholds significantly differs between the original studies and the re-analyses.

We operate under the assumption that the underlying distribution of p-values (for any research method) is continuous and infinitely differentiable. Any observed discontinuity in p-values is inferred to result from p-hacking or publication bias.

It is pertinent to note that publication bias is likely to operate predominantly in a single direction (towards significance), as an excess of successes is more indicative of bias than a scarcity. Hence, one-sided p-values are considered for our tests. The outcomes are detailed in Supplementary Materials Appendix Table 16 for the 5% threshold. In the first panel we use observations where $(0.01 < p < 0.09)$. The lower panels use smaller windows. In the first panel, 78.3% of the original analysis p-values within this window are significant. A test for whether this proportion is statistically greater than 0.50 yields a p-value of 0.000. Similarly, we obtain very small p-values for the smaller windows, confirming the presence of p-hacking or publication bias in the sample of original studies.

We further test for the presence of p-hacking and publication bias by employing the methodology and code by [8], and conducting six distinct tests to assess p-hacking and publication bias: Binomial, Fisher's, Discontinuity, CS1, CS2B, and LCM. The outcomes are detailed in Supplementary Materials Appendix Figure 6. This figure presents p-curves and test statistics for the battery of p-hacking tests for the full sample in the first panel, for the economics subsample in the second, and the political science subsample in the third.

In the absence of p-hacking and publication bias, the p-curve should be non-increasing; a spike just to the left of the 0.05 threshold is indicative of p-hacking. This spike is present in the full sample, though larger in the political science subsample than the economics subsample.

Tests based on non-increasingness include the Binomial Test and Fisher's test. Only for the political science subsample is there sufficient evidence to reject the null that the density (PDF) of p-values is non-increasing. In the absence of p-hacking, the

PDF is continuous. Again, only for the political science subsample is there sufficient evidence to reject the null that the density (PDF) of p-values is continuous.

Under general assumptions, p-curves are completely monotonic (the CS1 test) and are upper bounded in PDF and its derivatives (CS2B test). Here the trend reverses, in that only the full sample and the economics subsample offer sufficient evidence to reject the null of monotonicity and violations of the upper bound and derivatives of the PDF.

Last, a consequence of hypothesizing the non-increasingness of the PDF is that the PDF is also concave. The LCM test (Least Concave Majorant) assesses concavity of the CDF of p-values. Again, only the full sample and the economics subsample offer sufficient evidence to reject the null of concavity.

Overall, we take this mixed evidence to indicate the presence of p-hacking in both the economics and political science subsamples, as well as the full sample.

Recoding: Additional Results

Supplementary Materials Appendix Table 12 shows our results. Out of 23 recoding exercises, we find major differences for three studies and minor differences for 10 studies. Two of the major differences were uncovered when using a different software and looking at the authors' code.

Additionally, one team that computationally reproduced the results using a different *version* of the software used by the authors uncovered noteworthy differences in the magnitude and significance of the estimates. About half the main claims were no longer reproducible (i.e., same sign and statistically insignificant or different sign) due to a change in the defaults used by base R when generating random numbers starting in version 3.6.0. This is the only instance where using a different version of the software led to major differences in the size and significance of the estimates.

These results suggest that most teams who recoded using a different software language or without looking at the authors' code could obtain similar or very similar results.

Time Trends in Data and Code Availability: Additional Results

Replication Folder Availability: Extended Data Figure 3 displays the percentage of papers which have a replication package over the sample of 1150 papers which *should* have a replication package according to the AEA 2020 Definition. We see a general increase in the trend of replication folders being provided between 2014 and 2020. We found replication folders are attached to 59.1% and 70.0% of papers in 2014 and 2015, respectively. Replication folder provision then increases to a seemingly stable value close to 90% in 2021, 2022 and 2023.

Extended Data Figure 4 breaks down the previous figure's sample into those journals sampled from economics and political science. While the increasing trend within both samples exists also in the subsamples, political science starts from a much lower inclusion of replication packages. Political science papers have percentages of papers with replication folders equal to 23.3%, 36.7%, 58.6% and 66.7% in years 2014, 2015,

2016, and 2017, respectively. In comparison, economics papers have percentages equal to 72.5%, 82.5%, 80.2%, and 85.0% in years 2014, 2015, 2016 and 2017, respectively.

Contents of Replication Folders: The data presented in Extended Figures 5 through 6 are subsamples of varying sizes, reflecting the variation in what is required for each paper's replication folder. Each figure represents the percentage which possess the associated variable (README, cleaning code, raw data, *etc.*). We display the percentage of a binary variable equal to "Yes" if the package contained "All" of the field, and "Not Yes" if the variable had only "Some" or "None" of that variable. We are generally finding that replication folders are improving in their contents over time, especially regarding the inclusion of READMEs and cleaning code. In earlier years, replication folders were more likely to include analysis code and final data. Part of the reason is that providing raw data and cleaning code makes redundant the inclusion of final data (since one can generate the final data from the package). Extended Data Figure 6 uses an alternative measure for "final data" which adds replication folders that have complete ("Yes") raw/intermediate data *and* have complete ("Yes") cleaning code to those replication folders which have "final data" explicitly included. This alternative definition yields "final" data inclusion to be around 60% for the whole sample with a range 56.7% as a minimum in 2014 and 2017 and a maximum of 65.2% in 2015.

Comparison Between Journals with a Data Editor and Journals Without a Data Editor in 2023: One question raised is the importance of data editors in improving the quality of replication folders. We split the sample presented earlier in this section into those journals which did not have a data editor in 2023 and those which did. Recall the journals which had a data editor in 2023 include: *American Journal of Political Science*, *Journal of Politics*, *American Economic Review*, *Review of Economic Studies*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Review: Insights*, and *Economic Journal*. Journals that did not have a data editor in 2023 include: *American Political Science Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*.

We continue to use the binary variables presented in the previous section and calculate a simple t-test to understand the difference in means which we present in Supplementary Materials Appendix Table 18. In general, journals with data editors are more likely to have replication folders (difference about 19%), are more likely to contain READMEs (difference about 16%), and contain complete code (differences in cleaning and analysis code being about 26% and 17%, respectively). The provision of data is more nuanced. Journals with data editors were more likely to provide raw data than those without data editors (difference about 19%) but less likely to provide final or intermediate data (differences equal to -20% and -17%, respectively. Again, we believe this is likely due to raw data being sufficient for producing final data. When using our alternative definition of "final" data inclusion that takes into account raw (or intermediate data) with complete cleaning code, the difference between journals with data editors and those without data editors reduces to about -6%.

Additional Discussion

We aim for high-quality reproduction reports and believe our process contributes positively to the scientific community for at least four reasons. First, original authors are allowed to respond and may point out flaws in the reproducers' work. In practice, original authors and reproducers do not disagree on the completeness of the replication package (e.g., whether raw data is provided) nor on the presence of major coding errors. Disagreements are almost always about the validity of robustness and replicability. Second, A.B. or a co-director at I4R checks the tone of both the original authors' response and reproducers' report. Third, while reproducers may make mistakes, so do reviewers and editors. Our reproducers have the advantage of having access to the replication package. They may identify coding errors and uncover coding decisions which may not be discussed in the main body of the article. For example, multiple studies in our sample do not mention the use of a weighting scheme for their main analysis. This coding decision is obvious to a reproducer, but not to an editor or reviewer. Relatedly, our teams of reproducers spent on average 13 active days working on their reproducibility and replicability. This may compare favorably to a typical referee report, which is not prepared with peers and may involve subjectivity about the contribution of the paper to the literature. Fourth, reproducers learn throughout the process and benefit from this experience. This, in itself, is a positive contribution.

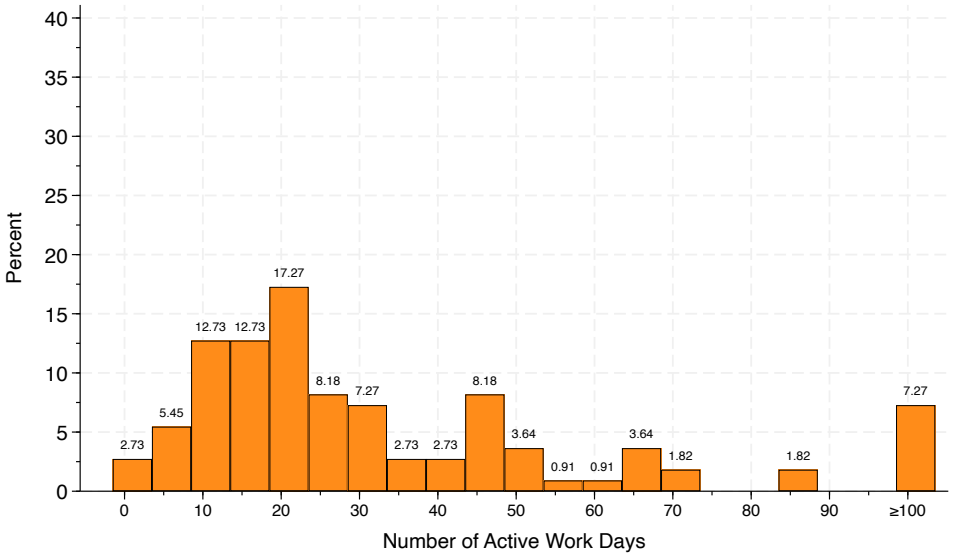
Barriers to Reproducibility and Robustness: We ask the following question in the team survey: "For which of the following reasons were you unable to conduct robustness checks, recoding exercises, extensions, or a replication using new data, prior to communications with the original authors? (Select all which apply)". Extensive Data Figure 6 provides a summary of the responses for these four categories. Out of 110 teams, 64 did not respond to the question. This suggests that the majority of teams felt their replication packages contained enough to create a reproduction report for I4R. That said, the lack of raw data restricted most what reproducers could do when analyzing a paper across all four categories. Raw data inhibited 19% of teams when trying to do robustness checks and 18% of teams wanting to recode key variables. 12% of teams also believed the lack of raw data inhibited their ability to perform a reproduction and 13% of teams believed it inhibited their ability to perform extensions. ([9] also provide evidence that non-reproducibility for the journal *Management Science* is due to non-availability/accessibility of data.) The remaining reasons for potential hurdles reproducers could have faced (like no intermediate data, no data dictionary, unclear documentation, and/or unclear replication package) did not affect most teams. About 7% of teams felt the original paper was unclear to the point of not being able to perform robustness checks. We thus see a lack of raw data provided in a replication package as a significant barrier to reproducibility and replicability, even in our selected sample of journals which have data and code availability policies.

On the Benefits for Reproducers: We document several benefits of conducting reproductions and replications. We ask the following question in the individual survey: "Please indicate the degree to which your experience with I4R has contributed to your improvement in the following areas." We offer six choices: (i) Networking, (ii) coding skills, (iii) capacity to write a good replication package, (iv) learning difference between reproduction and replication, (v) further ability as a researcher and (vi)

communicate issues with a paper to others. Supplementary Materials Appendix Table 17 provides a breakdown of the responses. We find that about 70% of reproducers responded that their experience with I4R contributed either a lot or moderately to their: (1) capacity to write a good replication package and (2) learning the difference between reproduction and replication. Reproducers further said their experience with I4R contributed at least moderately to furthering their ability as a researcher (about 53%) and their ability to communicate issues with a paper to others (about 60%).

Figures

Fig. 1: Histogram of Number of Active Work Days

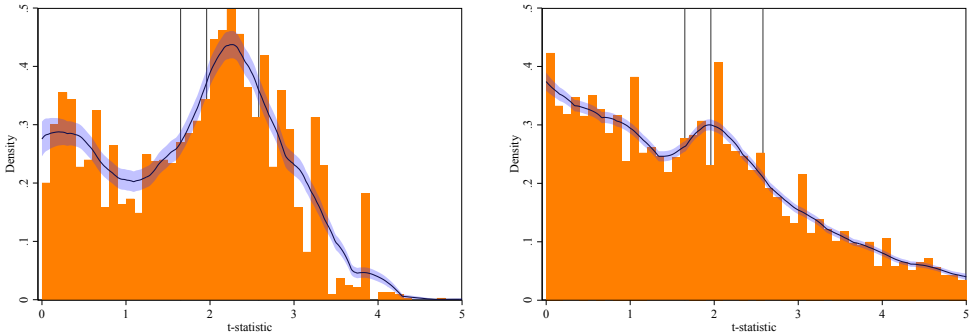


Legend: Data collected *via* survey of our reproducers after completing their reports. This figure illustrates the number of active days each team worked on their report.

Fig. 2: Distributions of t-Statistics and p-Values for Original Studies and [5]

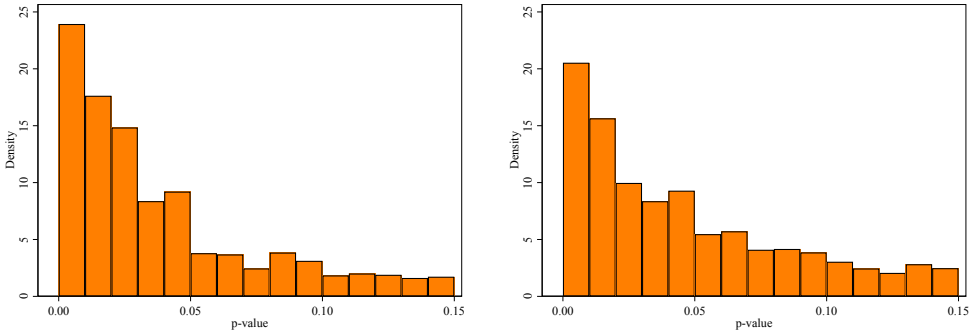
(a) Original Studies - t-statistics

(b) [5] Top-Ranked 25 Journals



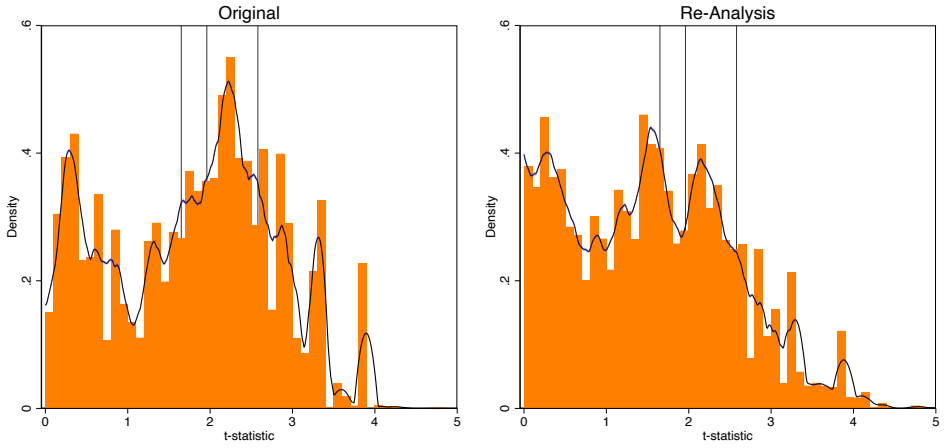
(c) Original Studies - p-values

(d) [5] Top-Ranked 25 Journals

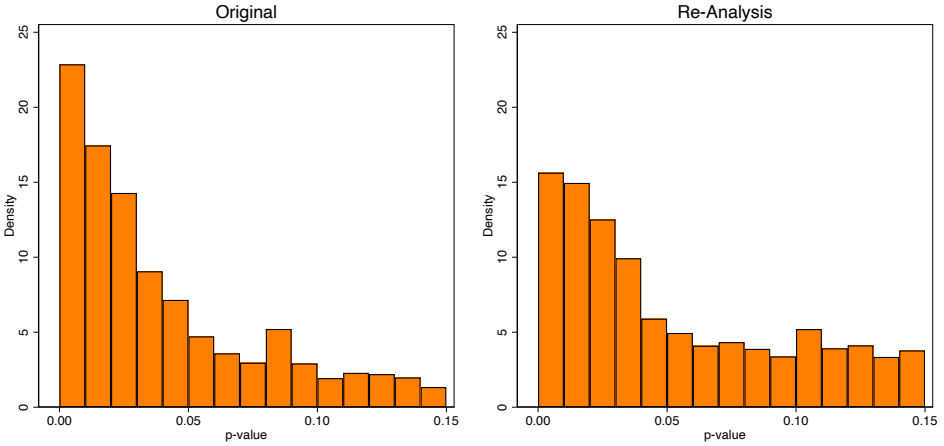


Legend: The top figures display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left figure includes all original studies in our data set. As a comparison, the top right figure plots the corresponding histogram of z-statistics from the top-ranked 25 economics journals published in 2015 and 2018 (from [5]). Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve (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 [5], respectively.

Fig. 3: Weighted Distributions of Statistics for Original Studies and Re-Analyses



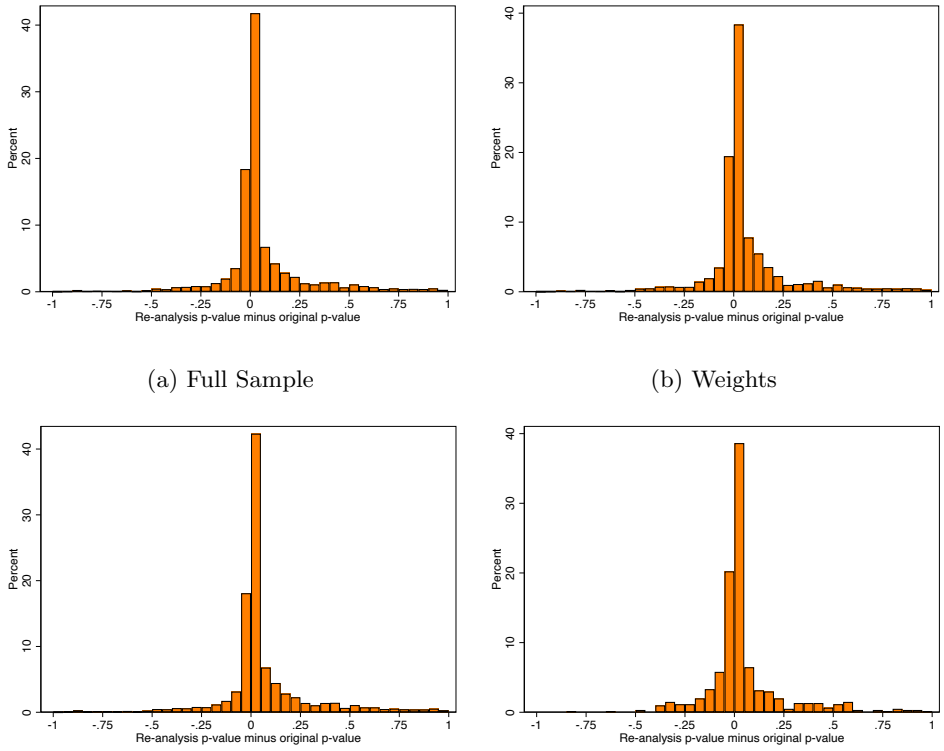
(a) t-statistic



(b) p-value

Legend: Top figures display histograms of test statistics for $t \in [0, 5]$, with bins of width 0.1, among original studies and re-analyses, respectively. Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve (which includes renormalization at 0). We use the inverse of the number of tests presented in the same article to weight observations. 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 re-analyses, respectively. We use the inverse of the number of tests presented in the same article to weight observations.

Fig. 4: Distribution of $p_{\text{re-analysis}} - p_{\text{original}}$ by Weights and Fields



(a) Full Sample

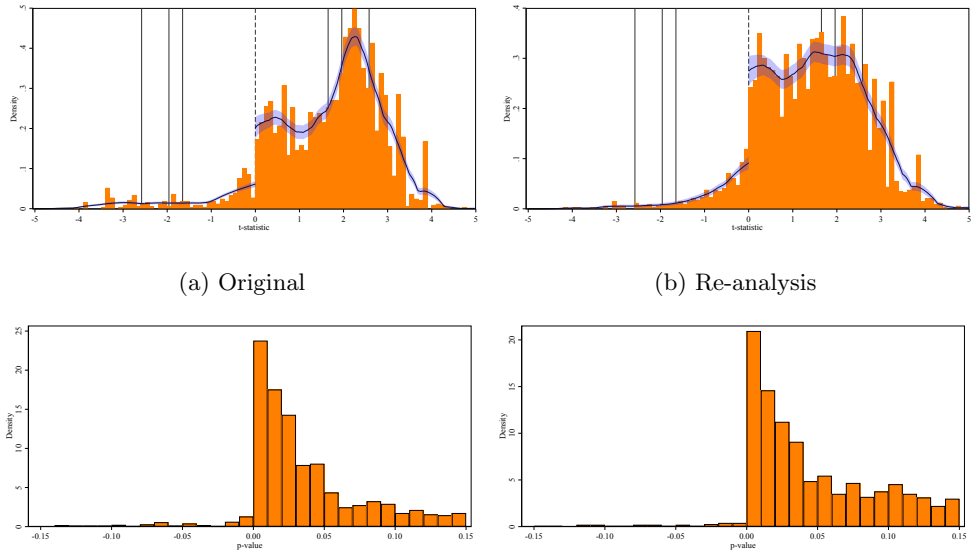
(b) Weights

(c) Economics

(d) Political Science

Legend: Second panel: We use the inverse of the number of test statistics in each reproduction report to weight observations. Third and fourth panel: The sample is restricted to original articles published in the indicated field. All panels: This figure presents the distribution of $(p_{\text{re-analysis}} - p_{\text{original}})$

Fig. 5: t and p-curves where negative represents a sign change from original to reproducer

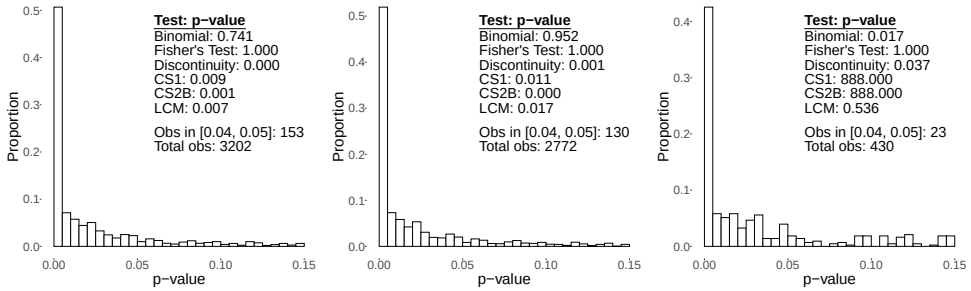


(a) Original (b) Re-analysis

(c) Original (d) Re-analysis

Legend: Top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. We have added a dashed reference line at $t = 0$, demarcating the areas where the reproducers' and original estimates agree in sign. For both sides of the zero line, vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve (which includes renormalization at 0), separately estimated for the positive and negative masses. Bottom panels display a histogram of test statistics for $p \in [0.00, 0.15]$, with bins of width 0.01. The left panels display statistics associated with originally published estimates. The right panels display statistics associated with reproducers' estimates. If the reproducer's estimated effect was of the opposite sign than the originally published estimate, we set the sign of the associated statistic to be negative.

Fig. 6: Applying [8]'s Tests



Notes: *Legend*: This figure present p-curves and results for the battery of p-hacking tests proposed in [8] for the full sample in the first panel, for the economics subsample in the second, and the political science subsample in the third. An error code of "888.00" represents an inability for that test to be calculated.

Tables

Table 1: Shifts in Statistical Significance Regions

	Re-Analysis Significance Level			
Original Significance Level	Sign Change	Not Sig.	Sig. at 5%	Total
Not Significant	4.99	28.47	3.71	37.16
Significant 5%	2.45	15.06	45.33	62.84
Total	7.44	43.53	49.03	100.00

Legend: This table illustrates shifts across significance and insignificance regions. Each row focuses on an initial level of statistical significance. Each column reports the share of re-analyses that ended up in each statistical significance region.

Table 2: Summary Statistics: Original Authors and Reproducers

	Mean (1)	Standard Deviation (2)	Minimum (3)	Maximum (4)
Test Statistics per Report	59.84	72.67	0	421
Year	2022.13	0.33	2022	2023
Economic Articles	0.72	0.45	0	1
Proportion of Economics Papers in Top 5	0.43	0.50	0	1
GS Citations (As of Report Completed)	43.98	71.39	0	573
Original Authors				
Number Original Authors	2.63	1.23	1	6
Share Graduate Student	0.06	0.18	0	1
Avg. Experience (Years since PhD)	11.21	6.34	0	31.50
Avg. GS Citations	4269.05	8882.00	31	55633.5
Replicators				
Number Replicators	3.25	1.22	1	7
Share Published Top 5 Econ/Targeted Poli Sci	0.15	0.36	0	1
Share Pub. Targeted Journals	0.30	0.46	0	1
Share Pub. Top 5/Targeted Poli Sci (Past 5 Years)	0.14	0.34	0	1
Share Pub. Targeted Journals (Past 5 Years)	0.26	0.44	0	1
Share Team Graduate Student	0.49	0.34	0	1
Avg. Experience (Years since PhD)	3.12	3.10	0	13.50
Avg. GS Citations	478.49	1016.67	0	6095.33
Comfortable programming in Stata	0.74	0.44	0	1
Comfortable programming in R	0.64	0.48	0	1
Comfortable programming in MATLAB	0.14	0.34	0	1

Legend: Each observation is an article. We do not weight test statistics. The Top 5 journals in economics are the American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies. The 3 leading political science journals in our sample are the American Journal of Political Science, American Political Science Review and Journal of Politics. Panels two and three focus on the original authors and reproducers, respectively. Average experience is the mean of years since PhD. GS citations in the top panel refers to the number of Google Scholar citations for the original article as of the completion of the reproduction report. Average GS citations in the bottom panels refers to the number of Google Scholar citations at the time the report is completed.

Table 3: Summary Statistics by Types of Re-Analyses

	# Articles (1)	# Articles RGs (2)	# Articles Editor (3)	# Tests (4)
All Re-Analyses	103	81	22	6583
All Simultaneous Robustness Checks	51	41	10	809
Full Sample				
By Re-Analyses: Change in				
Control variables	58	45	13	1939
Sample	75	57	18	1774
Dependent Variable	23	18	5	285
Main Independent Variable	20	19	1	264
Estimation Method	33	28	5	605
Inference Method	23	19	4	542
Weighting Scheme	14	10	4	126
Use New Data	15	13	2	469
Economics				
By Re-Analyses: Change in				
Control variables	45	36	9	1612
Sample	55	47	8	1647
Dependent Variable	19	17	2	279
Main Independent Variable	15	15	0	195
Estimation Method	22	21	1	433
Inference Method	19	15	4	507
Weighting Scheme	9	8	1	80
Use New Data	13	11	2	461
Political Science				
By Re-Analyses: Change in				
Control variables	13	9	4	327
Sample	20	10	10	127
Dependent Variable	4	1	3	6
Main Independent Variable	5	4	1	69
Estimation Method	11	7	4	172
Inference Method	4	4	0	35
Weighting Scheme	5	2	3	46
Use New Data	2	2	0	8

Legend: This table shows the number of articles and test statistics for all re-analyses (top panel), by types of re-analyses (2nd panel), by types of re-analyses for economic articles (3rd panel) and by types of re-analyses for political science articles (bottom panel), respectively. The second and third columns show the number of reports created *via* replication games and editor stream, respectively.

Table 4: Communication with Original Authors

	# Authors Contacted (1)	% Responded (2)	% Short Note (3)	% Feedback (4)	% Formal Response (5)
Economics	75	93%	11%	61%	28%
Political Science	31	97%	14%	53%	33%
Total	106	94%	11%	59%	30%

Legend: This table provides information about original authors' responses. The second column shows that 94% of original authors that A.B. reached out to responded to his email. The remaining columns restrict the sample to those that responded.

Table 5: JEL Codes in our Sample

	Our Sample (All)		Representative Sample	
Top 10 JEL Codes in our Sample	Rank	%	Rank	%
D: Microeconomics	1	54.4	1	15.2
J: Labor and Demographic Economics	2	33.8	5	8.4
O: Economic Dev., Innov., Tech. Change, and Growth	3	33.8	6	7.9
I: Health, Education, and Welfare	4	29.4	10	6.3
H: Public Economics	5	17.6	9	6.3
N: Economic History	6	17.6	15	1.4
C: Mathematical and Quantitative Methods	7	16.2	2	15.1
E: Macroeconomics and Monetary Economics	8	13.2	4	10.7
L: Industrial Organization	9	13.2	11	5.6
G: Financial Economics	10	5.8	3	13.9
Q: Ag. and NR Econ & Envr. and Ecological Econ	11	7.4	7	7.7
P: Pol. Econ. and Comp. Economic Systems	12	5.8	17	0.8
Z: Other Special Topics	13	8.3	16	1
M: Bus. Admin and Bus. Econ & Mktg & Accg & Personnel Econ	14	3.3	13	1.8
R: Urban, Rural, Regional, Real Estate, and Trans. Economics	15	5.8	12	2.9
F: International Economics	16	2.5	8	7.6
K: Law and Economics	17	8.3	14	1.4
A: Gen. Econ & Teaching	18	NA	18	0.4
B: History of Econ Thought, Methodol., Heterodox Approaches	19	NA	19	0.4
Y: Miscellaneous Categories	20	NA	20	0.2

Legend: This table compares the JEL Codes in our sample and in a representative sample of economics papers ([3]). The JEL Codes are only available for some of the economic journals.

Table 6: Many-Analysts' Robustness Rate and Reproducer Characteristics For Published Results Originally Statistically Significant at the 5% Level

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	42.78	43.33	13.89	0.00	100.00
2	36.75	24.79	30.13	8.33	100.00
3	0.00	33.33	63.89	2.78	100.00
4a	0.00	16.67	50.00	33.33	100.00
4b	16.67	0.00	50.00	33.33	100.00
4c	16.67	50.00	16.67	16.67	100.00
5a	0.00	16.67	52.78	30.56	100.00
5b	8.33	40.28	34.72	16.67	100.00
5c	22.22	52.78	8.33	16.67	100.00
6	0.00	30.56	52.78	16.67	100.00
7	8.33	13.89	61.11	16.67	100.00
8	0.00	23.61	76.39	0.00	100.00

Legend: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between reproduction robustness rate and selected author/reproducer characteristics. This table restricts the analysis to originally published estimates that were statistically significant at the 5% level. The **columns** represent one of four categories that a many-analyst could classify their analysis; if a many-analyst found a relationship that was statistically significant (at the 5% level) and positive, it was included in the column 'Pos. & Sig.' The **rows** represent eight pre-registered research questions (two have three sub-questions). They are: 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? 4a- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (i) Are reproducibility/replicability rate higher when authors' experience is high, and replicators' experience is low (in comparison to similar levels)? 4b- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (ii) Are reproducibility/replicability rate higher when authors' experience and replicators' experience is similar (in comparison to dissimilar levels)? 4c- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (iii) Are reproducibility/replicability rate higher when authors' experience is low, and replicators' experience is high (in comparison to similar levels)? 5a- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (i) Are reproducibility/replicability rate higher when authors' have high prestige, and replicators' experience have low prestige (in comparison to similar levels)? 5b- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (ii) Are reproducibility/replicability rate higher when authors' and replicators' prestige is similar (in comparison to dissimilar levels)? 5c- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (iii) 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? For example, the **top row can be interpreted** as no many-analysts find a positive and statistically significant relationship between replicators' experience coding and replication rate. 13.89% of many-analyst teams find a positive but not statistically significant relationship. 42.78% find a negative and statistically significant relationship, and 43.33% of many-analyst teams find a negative and not statistically significant relationship. Most many-analysts provide more than one estimate per research question; this table weights many-analysts equally.

Table 7: Many-Analysts' Robustness Rate and Reproducer Characteristics For Published Results Originally Statistically Significant at the 10% Level

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	28.33	68.89	2.78	0.00	100.00
2	37.96	37.04	16.67	8.33	100.00
3	0.00	47.22	50.00	2.78	100.00
4a	0.00	8.33	33.33	58.33	100.00
4b	16.67	8.33	41.67	33.33	100.00
4c	8.33	58.33	0.00	33.33	100.00
5a	5.56	19.44	25.00	50.00	100.00
5b	16.67	36.11	30.56	16.67	100.00
5c	13.89	69.44	0.00	16.67	100.00
6	0.00	16.67	66.67	16.67	100.00
7	8.33	0.00	55.56	36.11	100.00
8	0.00	16.67	75.00	8.33	100.00

Legend: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between reproduction robustness rate and selected author/reproducer characteristics. This table restricts the analysis to originally published estimates that were statistically significant at the 10% level. The **columns** represent one of four categories that a many-analyst could classify their analysis; if a many-analyst found a relationship that was statistically significant (at the 5% level) and positive, it was included in the column 'Pos. & Sig.' The **rows** represent eight pre-registered research questions (two have three sub-questions). They are: 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? 4a- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (i) Are reproducibility/replicability rate higher when authors' experience is high, and replicators' experience is low (in comparison to similar levels)? 4b- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (ii) Are reproducibility/replicability rate higher when authors' experience and replicators' experience is similar (in comparison to dissimilar levels)? 4c- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (iii) Are reproducibility/replicability rate higher when authors' experience is low, and replicators' experience is high (in comparison to similar levels)? 5a- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (i) Are reproducibility/replicability rate higher when authors' have high prestige, and replicators' experience have low prestige (in comparison to similar levels)? 5b- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (ii) Are reproducibility/replicability rate higher when authors' and replicators' prestige is similar (in comparison to dissimilar levels)? 5c- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (iii) 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? Most many-analysts provide more than one estimate per research question; this table weights many-analysts equally.

Table 8: Many-Analysts' Robustness Rate and Reproducer Characteristics For Published Results Originally **Not** Statistically Significant at the 5% Level

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	3.33	88.33	8.33	100.00
2	0.00	35.19	64.81	0.00	100.00
3	0.00	11.11	88.89	0.00	100.00
4a	0.00	33.33	50.00	16.67	100.00
4b	0.00	41.67	41.67	16.67	100.00
4c	0.00	25.00	50.00	25.00	100.00
5a	0.00	16.67	69.44	13.89	100.00
5b	5.56	61.11	25.00	8.33	100.00
5c	0.00	29.17	40.28	30.56	100.00
6	8.33	66.67	25.00	0.00	100.00
7	0.00	58.33	33.33	8.33	100.00
8	16.67	58.33	19.44	5.56	100.00

Legend: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between reproduction robustness rate and selected author/reproducer characteristics. This table restricts the analysis to originally published estimates that were not statistically significant at the 5% level. The **columns** represent one of four categories that a many-analyst could classify their analysis; if a many-analyst found a relationship that was statistically significant (at the 5% level) and positive, it was included in the column 'Pos. & Sig.' The **rows** represent eight pre-registered research questions (two have three sub-questions). They are: 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? 4a- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (i) Are reproducibility/replicability rate higher when authors' experience is high, and replicators' experience is low (in comparison to similar levels)? 4b- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (ii) Are reproducibility/replicability rate higher when authors' experience and replicators' experience is similar (in comparison to dissimilar levels)? 4c- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (iii) Are reproducibility/replicability rate higher when authors' experience is low, and replicators' experience is high (in comparison to similar levels)? 5a- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (i) Are reproducibility/replicability rate higher when authors' have high prestige, and replicators' experience have low prestige (in comparison to similar levels)? 5b- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (ii) Are reproducibility/replicability rate higher when authors' and replicators' prestige is similar (in comparison to dissimilar levels)? 5c- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (iii) 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? Most many-analysts provide more than one estimate per research question; this table weights many-analysts equally.

Table 9: Many-Analysts' Robustness Rate and Reproducer Characteristics For Published Results Originally **Not** Statistically Significant at the 10% Level

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	11.67	71.67	16.67	100.00
2	0.00	35.19	64.81	0.00	100.00
3	0.00	36.11	63.89	0.00	100.00
4a	0.00	16.67	75.00	8.33	100.00
4b	0.00	38.89	52.78	8.33	100.00
4c	0.00	16.67	66.67	16.67	100.00
5a	0.00	45.83	29.17	25.00	100.00
5b	0.00	66.67	25.00	8.33	100.00
5c	0.00	37.50	37.50	25.00	100.00
6	0.00	83.33	16.67	0.00	100.00
7	0.00	61.11	30.56	8.33	100.00
8	16.67	58.33	16.67	8.33	100.00

Legend: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between reproduction robustness rate and selected author/reproducer characteristics. This table restricts the analysis to originally published estimates that were not statistically significant at the 10% level. The **columns** represent one of four categories that a many-analyst could classify their analysis; if a many-analyst found a relationship that was statistically significant (at the 5% level) and positive, it was included in the column 'Pos. & Sig.' The **rows** represent eight pre-registered research questions (two have three sub-questions). They are: 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? 4a- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (i) Are reproducibility/replicability rate higher when authors' experience is high, and replicators' experience is low (in comparison to similar levels)? 4b- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (ii) Are reproducibility/replicability rate higher when authors' experience and replicators' experience is similar (in comparison to dissimilar levels)? 4c- Does reproducibility/replicability rate depend on the interaction of the authors' experience and replicators' experience? In particular, (iii) Are reproducibility/replicability rate higher when authors' experience is low, and replicators' experience is high (in comparison to similar levels)? 5a- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (i) Are reproducibility/replicability rate higher when authors' have high prestige, and replicators' experience have low prestige (in comparison to similar levels)? 5b- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (ii) Are reproducibility/replicability rate higher when authors' and replicators' prestige is similar (in comparison to dissimilar levels)? 5c- Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige? In particular, (iii) 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? Most many-analysts provide more than one estimate per research question; this table weights many-analysts equally.

Table 10: Many-Analysts' Robustness Rate and Reproducer Characteristics - Only if Analyst Indicated the Effect Size was Meaningful

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	54.17	45.83	0.00	0.00	100.00
2	47.33	28.67	14.00	10.00	100.00
3	0.00	27.78	38.89	33.33	100.00
4a	0.00	0.00	50.00	50.00	100.00
4b	20.00	0.00	40.00	40.00	100.00
4c	16.67	50.00	16.67	16.67	100.00
5a	0.00	16.67	52.78	30.56	100.00
5b	12.50	25.00	37.50	25.00	100.00
5c	33.33	41.67	0.00	25.00	100.00
6	0.00	30.00	50.00	20.00	100.00
7	20.00	6.67	53.33	20.00	100.00
8	0.00	34.00	66.00	0.00	100.00

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	50.00	50.00	0.00	0.00	100.00
2	55.00	25.00	10.00	10.00	100.00
3	0.00	41.67	25.00	33.33	100.00
4a	0.00	0.00	12.50	87.50	100.00
4b	25.00	0.00	25.00	50.00	100.00
4c	8.33	58.33	0.00	33.33	100.00
5a	6.67	13.33	20.00	60.00	100.00
5b	25.00	25.00	25.00	25.00	100.00
5c	16.67	63.33	0.00	20.00	100.00
6	0.00	20.00	60.00	20.00	100.00
7	20.00	0.00	26.67	53.33	100.00
8	0.00	37.50	50.00	12.50	100.00

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	0.00	83.33	16.67	100.00
2	0.00	100.00	0.00	0.00	100.00
3	0.00	41.67	58.33	0.00	100.00
4a	0.00	33.33	33.33	33.33	100.00
4b	0.00	33.33	33.33	33.33	100.00
4c	0.00	33.33	33.33	33.33	100.00
5a	0.00	0.00	72.22	27.78	100.00
5b	11.11	72.22	0.00	16.67	100.00
5c	0.00	37.50	16.67	45.83	100.00
6	12.50	75.00	12.50	0.00	100.00
7	0.00	50.00	33.33	16.67	100.00
8	50.00	33.33	5.56	11.11	100.00

	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	0.00	83.33	16.67	100.00
2	0.00	100.00	0.00	0.00	100.00
3	0.00	75.00	25.00	0.00	100.00
4a	0.00	0.00	83.33	16.67	100.00
4b	0.00	50.00	33.33	16.67	100.00
4c	0.00	12.50	75.00	12.50	100.00
5a	0.00	12.50	50.00	37.50	100.00
5b	0.00	83.33	0.00	16.67	100.00
5c	0.00	37.50	25.00	37.50	100.00
6	0.00	87.50	12.50	0.00	100.00
7	0.00	38.89	44.44	16.67	100.00
8	33.33	50.00	0.00	16.67	100.00

Legend: This table presents the same analysis as in Tables 6, 7, 8, and 9 while only including analyst results that were indicated by the analysis that "in your opinion, is the estimated effect size economically meaningful?" The first panel corresponds to Table 6. The second panel corresponds to Table 7. The third panel corresponds to Table 8. The fourth panel corresponds to Table 9. The rows correspond to the same research questions, and the columns represent the same effect sign and statistical significance categories. The cells remain weighted in the same manner.

Table 11: Summary Statistics by Journal

Discipline and Journal	# Articles Total (1)	# Articles RGs (2)	# Articles Editor (3)	# Tests (4)	Data Editor (5)
Economics	79	67	12	5,494	
American Economic Review	17	12	5	1,392	Yes
American Economic Review: Insights	2	0	2	149	Yes
American Economic J.: Applied Economics	9	6	3	260	Yes
American Economic J.: Economic Policy	11	11	0	811	Yes
American Economic J.: Macroeconomics	3	3	0	25	Yes
Economic Journal	20	18	2	1,262	Yes
Journal of Political Economy	8	8	0	1,283	No
Quarterly Journal of Economics	4	4	0	101	No
Review of Economic Studies	5	5	0	211	Yes
Political Science	31	16	15	1,089	
American Journal of Political Science	13	6	7	539	External
American Political Science Review	6	3	3	214	No
Journal of Politics	12	7	5	336	Yes
Total	110	83	27	6,583	

Legend: This table provides an overview of test statistics and articles reproduced and/or replicated by journal. Columns 1 and 4 indicate the number of article and test statistics per journal, respectively. Columns 3 and 4 report the number of articles per stream, where RGs is an acronym for Replication Games. Column 5 indicates if the journal has a data editor.

Table 12: Recoding Using Same or Different Softwares

	Identical (1)	Minor Differences (2)	Major Differences (3)	Total (4)
Same Software (Without Looking)	2	2	1	5
Different Software (Without Looking)	1	1	0	2
Different Software (Looking)	8	7	2	17
Total	10	10	3	23

Legend: This table illustrates the number of reports recoding the analysis (i) in the same software without looking at the authors' code/programs, (ii) using a different software language without looking at the authors' code/programs or (iii) using a different software language looking at the authors' code/programs.

Table 13: Caliper Tests, Significance at 5% Level

	Significant at 5% Level			
	(1)	(2)	(3)	(4)
Re-Analysis=1	-0.080***	-0.094**	-0.073	-0.136**
	(0.029)	(0.045)	(0.051)	(0.068)
Observations	2,027	1,353	801	420
Threshold	0.05	0.05	0.05	0.05
Window	0.04	0.03	0.02	0.01

Legend: The dependent variable takes a value of one if $p \leq 0.05$. The variable Re-Analysis takes a value of one if the p-value is associated with a re-analysis, and zero if it is associated with the original publication. For example, in column 1 a Re-Analysis p-value is 8.9% less likely to be statistically significant than an original publication p-value at the 5% level in the small window of $0.01 \leq p \leq 0.09$. Standard errors clustered at the paper level. Estimated using probit, with marginal effects presented.

Table 14: Caliper Tests, Significance at 10% Level

	Significant at 10% Level			
	(1)	(2)	(3)	(4)
Re-Analysis=1	-0.085	-0.097	-0.134*	-0.169*
	(0.053)	(0.063)	(0.072)	(0.091)
Observations	812	634	445	212
Threshold	0.10	0.10	0.10	0.10
Window	0.04	0.03	0.02	0.01

Legend: The dependent variable takes a value of one if $p \leq 0.10$. The variable Re-Analysis takes a value of one if the p-value is associated with a re-analysis, and zero if it is associated with the original publication. Standard errors clustered at the paper level. Estimated using probit, with marginal effects presented.

Table 15: Applying [7]

	μ	τ	df	$[0, 1.645]$	$(1.645, 1.96]$	$(1.96, 2.576]$
Original Analysis	0.0006	0.0024	1.2705	0.1716	0.3829	1.0740
Re-Analysis	0.0001	0.0000	1.2836	0.2731	0.6430	0.8994
Original Economics	0.0002	0.0011	1.1969	0.1522	0.3910	1.0556
Re-Analysis Economics	0.0000	0.0000	1.1942	0.2705	0.6107	0.9020
Original Political Science	0.0155	0.0254	2.1907	0.3078	0.3496	1.1846
Re-Analysis Political Science	0.0069	0.0155	2.4069	0.2653	0.6693	0.7916

Legend: An application of [7]. The columns μ, τ, and df represent the model's estimated parameters (using an underlying t-distribution and symmetric sign probabilities). The fourth column $[0, 1.645]$ presents the relative publication probability for a t-statistic in the $[0, 1.645]$ interval compared to one in the reference interval of $(2.576, \infty)$.

Table 16: Randomization Tests, Significance at 5% Level

	Original Analysis
Proportion Significant in .05 ± .04	0.783
One-Sided p-value against 0.50	0.000
Number of Tests in .05 ± .04	2027.000
Proportion Significant in .05 ± .03	0.751
One-Sided p-value against 0.50	0.000
Number of Tests in .05 ± .03	1353.000
Proportion Significant in .05 ± .02	0.689
One-Sided p-value against 0.50	0.000
Number of Tests in .05 ± .02	801.000
Proportion Significant in .05 ± .01	0.690
One-Sided p-value against 0.50	0.000
Number of Tests in .05 ± .01	420.000

Legend: Following [5], in this table we present the results of binomial proportion tests where a success is defined as a statistically significant observation at the 5% level. In the first panel we use observations where $(0.01 < p < 0.09)$. The lower panels use smaller windows. We test if the proportion is statistically greater than 0.50. The associated p-values are then reported. We also include the number of observations in the third row. We do not weight articles.

Table 17: Please indicate the degree to which your experience with I4R has contributed to your improvement in the following areas (select all which apply):

	Nothing	A Little	Moderately	A Lot	Don't Know	Not Applicable
Networking	10.40	46.82	27.17	10.69	2.89	2.02
Coding Skills	19.08	40.17	26.88	10.98	1.73	1.16
Capacity to write a good replication package	5.19	21.90	46.97	23.63	1.15	1.15
Learning difference between reproduction and replication	6.65	19.36	36.71	33.53	3.47	0.29
Further ability as a researcher	5.20	39.02	38.15	17.05	0.29	0.29
Communicate issues with a paper to others	3.75	28.82	41.50	23.05	0.58	2.31

Legend: This table provides information on reproducers' feelings about how I4R contributed to their improvement in various areas. Each row represents a different category. Values are percentages and all rows in a category sum to 100. All values are unweighted.

Table 18: Mean Differences in Replication Package Contents in 2023 by those with and without a Data Editor

	Mean			
	Has Data Editor	No Data Editor	Difference	P-value
Is link to replication folder on website?	0.956	0.767	0.189	0.002
Does replication package contain a README?	0.922	0.767	0.156	0.021
Does replication package contain cleaning code?	0.822	0.567	0.256	0.004
Does replication package contain analysis code?	0.933	0.767	0.167	0.011
Does replication package contain raw data?	0.422	0.233	0.189	0.065
Does replication package contain intermediate data?	0.200	0.400	-0.200	0.029
Does replication package contain final data?	0.367	0.533	-0.167	0.110
(i) Final Data or (ii) cleaning code + raw/intermediate data?	0.544	0.600	-0.056	0.599

Legend: Columns (1) and (2) display the mean of the binary statistic labelled for each row. Column (3) shows the difference between column (1) and column (2). Column (4) shows the p-value of the applicable t-test. The binary variables are equal to one if the contents of the package were believed to be there and equal to zero if none or only some of the contents were included in the replication package. Samples vary across all rows as the statistics omits observations where it did not apply for whatever reason (*e.g.* simulations may have no data and no cleaning code). Journals which *had* a data editor in 2023 include: *American Journal of Political Science*, *Journal of Politics*, *American Economic Review*, *Review of Economic Studies*, *American Economic Journal: Macroeconomics*, *American Economic Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *American Economic Review: Insights*, and *Economic Journal*. Journals that did *not have* a data editor in 2023 include: *American Political Science Review*, *Journal of Political Economy*, and *Quarterly Journal of Economics*.

List of Articles and Reproduction

Reproduction Report

Title Original Study: Antinormative Messaging, Group Cues, and the Nuclear Ban Treaty

doi: https://doi.org/10.1086/714924, Journal of Politics

Abstract: Herzog, Baron, and Gibbons (2022) explore the effects of exposure to official elite rhetoric and group cues on public support against the international nuclear weapons prohibition norm. The authors find that elite cues, in particular security and institutional cues, increase individuals' opposition to the Treaty on the Prohibition of Nuclear Weapons (TPNW). However, elite cues do not seem to have an effect on changing individuals' broader attitudes towards nuclear weapons, as measured by individuals' existing opposition to nuclear arms. We replicate and expand the authors' methods and results to test the robustness of the effects found in the study. First, we reproduce the main finding using the authors' original data and method. We do not find any coding errors that undermine the authors' analysis

or conclusions. Second, we test the robustness of the results by (1) using a different operationalization of party identity, and (2) calculating additional subgroup analysis for gender. We find no significant differences between our replicated and the original results, however females' support for the TPNW is more responsive to security cues, while males' support is more responsive to institutions cues.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/97.htm

Replication Package: https://osf.io/xbvzg/

Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/98.htm

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=5bc4beb0bd5aef9d7a5ba5284fc6?persistentId=doi%3A10.7910%2FDVN%2FGLT4FX&version=&q=&fileTypeGroupFacet=%22Text%22&fileAccess=&fileSortField=size

Reproduction Report

Title Original Study: Ascriptive Characteristics and Perceptions of Impropriety in the Rule of Law: Race, Gender, and Public Assessments of Whether Judges Can Be Impartial

doi: https://doi.org/10.1111/ajps.12599, American Journal of Political Science

Abstract: Ono & Zilis (2022) investigated the effects of ascriptive characteristics of US American judges, such as race, gender, and ethnicity, on citizens' perceptions of the judges' professional impropriety and bias in their rulings. They conducted two studies, comparing citizens' perceptions of different ascriptive characteristics and judgments about the judges' biases and the need for recusal from cases. They found that political and ideological predispositions shape perceptions of judicial impropriety. In this comment, we recode the analysis using a different software and conduct robustness checks. We were able to reproduce the main results.

Link to Full Report: https://osf.io/yf48r/

Replication Package: https://osf.io/yf48r/

Link to Original Authors' Response: No response.

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZHOL6Y

Reproduction Report

Title Original Study: Assortative Matching at the Top of the Distribution: Evidence from the World's Most Exclusive Marriage Market
doi: https://doi.org/10.1257/app.20180463, American Economic Journal: Applied Economics
Report's Abstract: Goni (2022) relies on a novel data on peerage marriages in Britain to examine the impact of matching technology on marital sorting. He relies on the London Season interruption (1861 – 1863) as a natural experiment that raised search costs and reduced market segregation. In his preferred specification, he exploits exogenous variation in womens' probability to marry during the interruption for their age in 1861 and finds that the interruption increased the probability of marrying a commoner; reduced the probability of marrying an heir, increased the difference in spouses' family landholdings (in absolute value); decreased the difference in spouses' family landholdings (husband – wife); and increased the likelihood of never getting married (See Table 2, columns 1 to 6, respectively). First, we reproduce the papers' main findings and find no coding errors. Second, we test the robustness of the results to (1) the use of additional fixed effects and (2) sample restrictions. Finally, we examine the heterogeneous effects of this interruption by age and year. We find that original estimates are robust and are not significantly affected using these alternative specifications.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/47.htm
Link to Replicators' Package: https://osf.io/pqsem/
Original Author's Response: "I now reviewed the report carefully and with interest, and I am glad to see that the authors succeeded in replicating all results and found no coding errors. I hope the replication package was clear and easy to work with. I am also happy to see that they performed several additional robustness checks and heterogeneity analysis, and that these show that the original estimates "are robust and are not significantly affected using these alternative specifications" (p. 1). Given this, and the replicators' conclusion that "the study's main findings demonstrate robustness and reliability" (p. 7), I think that there is nothing substantial for me to write in a response in the form of a discussion paper. This is because both the replication exercise and the additional analysis found no major issues in the original work to respond to. I would also like to thank the authors for the fairness and professionalism of their report, and also for the time and effort they put in producing it, from which I ultimately benefit — as it adds to the credibility of my original paper — as well as the profession as a whole benefits — as making replication exercises more common is important for economics.

Please let me know if I can be of any further assistance regarding this Reproduction Report in the future. I am at your or the authors' disposal, in case I can be of help in clarifying anything in the replication package or in the analysis of the original paper. As I stated above, I believe that increasing replication rates is important for our field, as it is making original datasets publicly available — even when, as in the case of my paper, the data collection is an important part of my contribution, and in this situation, many do not grant public access to the original data."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/140921/version/V1/view

Reproduction Report

Title Original Study: Black Workers in White Places: Daytime Racial Diversity and White Public Opinion
doi: https://doi.org/10.1086/716289, Journal of Politics
Report's Abstract: In this replication study, we revisit the main empirical claims of Hamel and Wilcox-Archuleta's (HW) 2022 study on the impact of daytime racial diversity on White Americans' voting behavior and racial attitudes. HW introduce a novel zip code level measure of racial diversity that accounts for the influx of Black workers during daytime, showing that conventional purely residential based measures often underestimate the true degree of experienced racial diversity. Using survey data from the CCES, their findings suggest a negative correlation between racial flux and White Americans' Democratic voting tendencies and a positive correlation with racial resentment and opposition to affirmative action, all while controlling for the residential share of Blacks in the zip code. We assess the replicability of these findings by: (1) replicating the main results using the provided replication code, (2) reconstructing the racial flux measure and survey from raw data, (3) conducting multiverse analyses, and (4) replicating the analysis using an alternative data source. Our replication validates the robustness and accuracy of HW's initial conclusions, emphasizing the role of daytime racial diversity in shaping White Americans' political and racial attitudes.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/61.htm
Link to Replicators' Package: https://osf.io/ue4pm/
Original Authors' Response: "We enjoyed reading the replication, and don't see a need to write a response.

Thank you for doing this important work."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=da88db7b3367419a2d1a87a9e687?persistentId=doi%3A10.7910%2FDVN%2FFMOR6K&version=&q=&fileTypeGroupFacet=&fileAccess=&fileSortField=type

Reproduction Report

Title Original Study: Brahmin Left Versus Merchant Right: Changing Political Cleavages in 21 Western Democracies, 1948–2020
doi: https://doi.org/10.1093/qje/qjab036, Quarterly Journal of Economics
Report's Abstract: Gethin, Martínez-Toledano and Piketty (2022) analyze the long-run evolution of political cleavages using a new database on socioeconomic determinants of voting from approximately 300 elections in 21 Western democracies between 1948 and 2020. They find that, in the 1950s and 1960s, voting for the "left" was associated with lower-educated and low-income voters. After that, voting for the "left" has gradually become associated with higher-educated voters, while high income voters have continued to vote for the "right". In the 2010s, there is a disconnection between the effects of income and education on voting. In this replication, we first conduct a computational reproduction, using the replication package provided by the authors. Second, we do a robustness replication testing to what extent the original results are robust to i) restricting the sample to "core" left and right parties, ii) analyzing the top 80% versus bottom 20%, iii) weighting by population, iv) dropping control variables, and v) using country fixed effects. The main results of the paper are found to be largely replicable and robust.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/19.htm
Link to Replicators' Package: https://osf.io/2hpeq/
Original Authors' Response: "Thank you for your mail and for your interesting report! We are happy to see that you were able to easily replicate our results and that our main conclusions were found to be largely robust. In this context, we do not think that an answer from our side would be particularly useful: we are happy with the report as it is.

Thank you for the very valuable work that your institute is producing in testing the replicability and robustness of published studies!"
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XUSWG6

Reproduction Report

Title Original Study: Bubbles, Crashes, and Economic Growth: Theory and Evidence

doi: https://doi.org/10.1257/mac.20220015, American Economic Journal: Macroeconomics

Report's Abstract: Guerron-Quintana, Hirano, and Jinnai (2023) explore the short-, medium-, and long-run effects of financial bubbles on economic growth by way of a macroeconomic general equilibrium framework. In their model, a key theoretical result is that, in net terms, the "crowding in" of capital investment during a bubble ushers the economy onto a higher balanced growth path post-bubble than it was on pre-bubble (Figure 10), thus (seemingly) suggesting that economic bubbles are growth-enhancing. In turn, the main result of the paper is that this positive view of bubbles is a fallacy so long as the latter are recurrent, namely because a counterfactual economy in which bubbles never occur in the first place grows at a significantly faster pace (Figure 10). The reason for this is that the expectation of future bubbles stifles capital investment and, as such, reduces economic growth in the long run.

We successfully reproduce the paper's main figures using the original code provided in the replication package. Given the hard-coded nature of all empirical data used in the paper, most of our efforts are devoted to reproducing the employed empirical data itself and, in turn, conducting a direct replication with our own measures. Using various specifications of the HP filter, we are successful in qualitatively, but not quantitatively reproducing the paper's main time series (stock-market-to-GDP ratio). Nevertheless, even without updating the model's parameterization, the paper's main empirical findings (i.e. Figures 8-10) are largely robust to our own measure. In turn, we are successful in quantitatively reproducing the second key time series (credit-to-GDP ratio), albeit only with a highly unusual specification of the HP filter's smoothing parameter (10^{10} instead of 1600 for quarterly data). We find that, unlike in the case of the stock-to-GDP ratio, the paper's (auxiliary) findings are not robust to our own credit-to-GDP series

Link to Full Report: https://osf.io/d76tn/

Link to Replicators' Package: https://osf.io/d76tn/

Original Authors' Response: Provided a short response and answered a question. Did not provide a final response as of November 2025.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/173441/version/V1/view

Reproduction Report

Title Original Study: Campaign Contributions and Roll-Call Voting in the U.S. House of Representatives: The Case of the Sugar Industry
doi: https://doi.org/10.1017/S0003055422000466, American Political Science Review
Report's Abstract: In their study, Grier et al. (2023) explore the causal relationship between campaign contributions and roll-call voting. Their analysis focuses on the influence of campaign contributions on two specific anti-sugar votes conducted in 2013 and 2018. The authors identify a substantial increase in inflationadjusted sugar contributions from the sugar industry to incumbent politicians between these two voting events. The aim of our research is to replicate and validate the authors' main models. In addition to cross-platform replication, we conduct several robustness checks to further examine the reliability of their findings. These include (1) clustering the standard errors, (2) utilizing an Ordinary Least Squares (OLS) model instead of the authors' logistic regression, and (3) altering the dependent variable to represent the change in the vote from 2013 to 2018. Our results largely confirm the authors' findings and reveal additional insights regarding the money buys vote hypothesis.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/57.htm
Link to Replicators' Package: https://osf.io/4hjb9/
Original Authors' Final Response: "We thank the Institute for Replication for their diligent work replicating and performing some extensions to our 2023 APSR paper. Replication is an important and often undervalued work in the scientific process. Of course we are quite pleased to see that our results do replicate and that the extensions performed largely support the results and ideas we advanced in our paper. Keep up the good work!"
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2IFZR9

Reproduction Report

Title Original Study: Can Information Reduce Ethnic Discrimination? Evidence from Airbnb

doi: https://doi.org/10.1257/app.20190188, American Economic Journal: Applied Economics

Report's Abstract: Laouénan & Rathelot (2022) investigate the mechanism underlying ethnic discrimination using self-collected panel data from Airbnb between 2014 and 2017. They find that hosts from minority groups charge 3.2% less than those from the majority group within the same neighbourhood. Using a theoretical framework, they estimate that the ethnic price gap vanishes as more information (reviews) become available conditional on observables. The point estimates for their main results are statistically significant at the 1% level. This finding suggests that ethnic discrimination is due to statistical discrimination rather than taste-based discrimination. First, we reproduce the original article's main findings using R, whereby the authors of the original article use STATA. We can reproduce the main findings in R except for a few marginal discrepancies at the second or third decimal place. Second, we extend two robustness analyses reported in the original article. These robustness analyses impose restrictions on the sample and these restrictions are not justified in the article. Once these restrictions are not imposed, the picture becomes more complex and the robustness analysis warrants more discussion. However, only a small fraction of the observations causes some ambiguity and there might be good reasons to impose restrictions. Transparently presenting the robustness analyses with and without restrictions, motivating the restrictions and discussing its implications for the main findings would have been desirable. Generally, the original article does a great job with regard to reproducibility by providing data, code and documentation that ease the reproduction of a complex analysis. We conclude that our reproduction and replication support the main findings of the original article.

Link to Full Report: https://osf.io/zn98a/

Link to Replicators' Package: https://github.com/TuanNguyen04/Replication_Airbnb

Original Authors' Response: The authors provided initial feedback which the replicators took it into account.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/120078/version/V1/view

Reproduction Report

Title Original Study: Can Technology Solve the Principal-Agent Problem? Evidence from China's War on Air Pollution

doi: https://doi.org/10.1257/aeri.20200373, American Economic Review: Insights

Report's Abstract: Greenstone et al. examine the effect of the introduction of automatic air pollution monitoring on the reporting of local air pollution in China. Using 654 regression discontinuity designs (RDDs) based on city-level variation in the day that monitoring was automated, they find an immediate and lasting increase of 35 percent in reported PM10 concentrations post-automation. Moreover, they find that automation's introduction increases online searches for face masks and air filters by 200 percent and 28 percent, respectively, using an RDD. Results are consistent when using an event study design. First, we were able to computationally replicate the results. Second, we find that results are robust to more flexible specifications of the weather variables, to re-constructed weather variables using the same matching procedure as the authors (i.e., closest station) and meteorological data with additional weather stations, to alternative construction of the weather variables using an inverse distance weighted approach of the surrounding weather stations, and to more flexible choices of fixed effects (up to the city level). Finally, we find limited evidence of discontinuity in objective measures of ground pollution (i.e., AOD) for a sub-sample using alternative weather variables. The estimate, however, is economically insignificant. Moreover, no discontinuity is observed in the full sample. Therefore, we believe this result does not invalidate the original study's findings.

Link to Full Report: https://osf.io/b7dn2/

Link to Replicators' Package: https://osf.io/m8hfr/?view_only=9f6632ec96c0451daf0f8889b9ad2b25

Original Authors' Response: https://osf.io/b7dn2/

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/125321/version/V1/view

Reproduction Report

Title Original Study: Can't We All Just Get Along? How Women MPs Can Ameliorate Affective Polarization in Western Publics

doi: https://doi.org/10.1017/S0003055422000491, American Political Science Review

Report's Abstract: We present a replication and extension of Adams et al. (2023), examining the influence of women Members of Parliament (MPs) on affective polarization. Conducted during the 2023 Montreal Replication Games, our analysis reaffirms the original findings through the authors' base R code and a tidyverse simplification. Our results highlight that the mitigating effect on polarization is predominantly observed among left-wing respondents, with null effects noted for centrist and right-wing parties. This discrepancy is attributed to left-wing parties' explicit commitment to gender equality. Further analysis reveals the study's robustness across different countries and years (1996-2007) while addressing data structure and imputation methods to ensure reliability. Our findings underscore the nuanced role of women MPs in political dynamics, particularly among left-wing voters, against democratic backsliding concerns.

Link to Full Report: https://osf.io/69px3/

Link to Replicators' Package: https://osf.io/69px3/

Original Authors' Response: Thank you for replicating our paper Can't We All Just Get Along? How Women MPs Can Ameliorate Affective Polarization in Western Publics (APSR 2023) as part of the Montreal Replication Games. We appreciate the attention to detail and rigor applied to the replication project. We are pleased that our initial results replicate well. We appreciate your robust approach to testing the stability of our findings using a country and year 'leave-one-out' cross-validation strategy. We also thank you for catching the coding error which dropped a handful of cases from the original analysis; we are glad that the results remain substantively the same when this error is corrected. We also are interested in the results from the extension you undertook, finding that our results are primarily driven by left-wing parties' supporters, in particular parties from the green, radical left and social Democratic parties. On the other hand, the point estimates are positive for all parties excepting the conservative and radical right parties, which can be expected to have the most conservative views on gender roles. We note that the authors' interpretation, that "the portion of women MPs affects the attitudes of left-wing voters and not the attitudes of the voters most likely to undermine democracy" is true, but that the results also suggest that far-right parties, who most aggressively challenge liberal democratic norms, may be able to "soften" their image among left-wing voters by running female candidates. This is consistent with the argument made by Catalano Week et al (2023), that radical right parties strategically run women to broaden their appeal. Again, we deeply appreciate your replication and insightful extension of our research.

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AHQRVR

Reproduction Report

Title Original Study: Changing Hearts and Minds? Why Media Messages Designed to Foster Empathy Often Fail
doi: https://doi.org/10.1086/719416, Journal of Politics
Report's Abstract: This paper focuses on computational reproducibility and robustness replicability of Gubler et al.'s(2022) studies which examine the effect of media messages on empathic concern, dissonance, and out-group policy attitudes. The original paper tests four hypotheses using two online experiments with large samples from one US state ($N1$ = 5,800; $N2$ = 2,200). Regarding the first experiment, we successfully reproduced the effect that initial antipathy weakens the effect of humanizing treatment on empathic concern (H1). However, we show that the moderating effect is negligible and has little practical significance. Moreover, the individual effect estimates in our analyses slightly differed from the original paper due to different procedure of data cleaning and minor coding errors in the original paper. The most relevant difference was the opposite effect of gender than reported in the original paper. We also show that empathic concern might mediate the effect of humanizing treatment on attitudes toward immigrants (H3). The original study rejected the mediation hypothesis due to not finding a total effect of humanizing treatment on attitudes. In contrast, we found that humanization treatment has a positive indirect effect on attitudes through empathic concern. At the same time, it also has a direct negative effect on attitudes. For the second experiment (H1, H2a, H2b, H3), we attempted to reproduce the results using a different software. We partially succeeded once receiving support from the authors of the original study. We note throughout the report issues we have encountered.
Link to Full Report: https://osf.io/zes6g/
Link to Replicators' Package: See Report's Online Appendix for the codes.
Original Authors' Response: https://osf.io/zes6g/
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FUCDTT

Reproduction Report

Title Original Study: Changing Tides: Public Attitudes on Climate Migration
doi: https://doi.org/10.1086/715163, Journal of Politics
Report's Abstract: See entry below.
Link to Full Report: https://www.socialsciencereproduction.org/reproductions/791/published/index
Link to Replicators' Package: https://github.com/alexkustov/Replication-of-Arias-and-Blair-2021
Original Authors' Response: "Thank you very much for reaching out! We are very pleased to hear that the results of our study were replicated, and do not need to provide an answer."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=name&fileSortOrder=desc

Reproduction Report

Title Original Study: Checking and Sharing Alt-Facts
doi: https://doi.org/10.1257/pol.20210037, American Economic Journal: Economic Policy
Report's Abstract: Henry, Zhuravskaya, and Guriev (2022) examine whether people are willing to share "alternative facts" espoused by right-wing populist parties before the 2019 European elections in France and how this interacted with the availability of fact-checking information. They find that both imposed and voluntary fact-checking reduce the likelihood of sharing false statements by approximately 45%, and that imposed and voluntary fact-checking have similar effect sizes. We reproduce these findings and introduce several alternative estimates to assess the robustness of the original results, including resolving an inconsistency in the handling of pre-treatment controls. Overall, our results align with the results of the original paper. The differences we find are small in absolute magnitude but, since many effects were small, not always trivial in terms of relative differences. This replication supports the conclusions of the original paper.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/34.htm
Link to Replicators' Package: https://doi.org/10.5281/zenodo.7858829
Link to Original Authors' Response: "Many thanks! No, we won't be writing a response."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/140161/version/V1/view

Reproduction Report

Title Original Study: Child Marriage Bans and Female Schooling and Labor Market Outcomes: Evidence from Natural Experiments in 17 Low- and Middle-Income Countries

doi: https://doi.org/10.1257/pol.20200008, American Economic Journal: Economic Policy

Report's Abstract: By studying child marriage bans in 17 developing countries, Wilson (2022) finds that raising the minimum legal age of marriage to 18 successfully increased the age at first marriage, the age at first birth, and the likelihood of employment. Additionally, the bans reduced child marriage and increased educational attainment in urban areas. We replicate these findings by collecting the raw data from the same sources as the paper and analysing the data following the procedures described in the paper, without referring to the data and codes provided by the author. Our findings are consistent with the results of the paper in terms of the statistical significance of point estimates and differ in magnitude by a negligible amount.

Link to Full Report: https://osf.io/5yhxc/

Link to Replicators' Package: https://osf.io/5yhxc/

Original Authors' Response: We could not reach out the author.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/130784/version/V1/view

Reproduction Report

Title Original Study: Concentration Bias in Intertemporal Choice
doi: https://doi.org/10.1093/restud/rdab043, Review of Economic Studies
Report's Abstract: Dertwinkel-Kalt et al. (2022) examine the effect of concentration bias - the tendency to overweight advantages that are concentrated in time relative to costs that are spread over multiple time periods - on intertemporal choice in a laboratory experiment. In their preferred empirical specification, the authors report that concentration bias leads to a 22.4% higher willingness to work than explained by a standard model of intertemporal discounting. We conduct a computational replication of the main results of the paper using the same procedures and original data. Our results confirm the sign, magnitude and statistical significance of the author's reported estimates across each of their five main findings.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/42.htm
Link to Replicators' Package: https://osf.io/d42xr/
Original Authors' Response: "We thank Deer, Ellingsrud, Heuer, and Kordt (2023) for conducting the Reproduction Report and appreciate that their "results confirm the sign, magnitude and statistical significance of [our] reported estimates across each of [our] five main findings" (p. 1). We don't have anything substantive to add to this. "
Original Authors' Package: https://zenodo.org/records/5091975

Reproduction Report

Title Original Study: Cooperative Property Rights and Development: Evidence from Land Reform in El Salvador
doi: https://doi.org/10.1086/717042, Journal of Political Economy
Report's Abstract: Montero (2022) explores a discontinuity in a land reform in El Salvador and reports two main findings. First, relative to outside-owned haciendas operated by contract workers, the productivity of worker-owned cooperatives is higher for staple crops and lower for cash-crop. Second, cooperative property rights increase workers' incomes and compress wage distributions. In this comment, we show that the latter result rests on two mistakes: three-quarters of the observations are duplicates and income inequality is calculated over too few workers to be meaningful. When corrected, the data sources and research design provide no credible evidence regarding the causal effects of ownership structure on income levels and inequality.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/20.htm
Link to Replicators' Package: https://doi.org/10.7910/DVN/AMD3NO
Link to Original Authors' Response: https://www.journals.uchicago.edu/doi/10.1086/725234
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/717042/suppl_file/20190161data.zip

Reproduction Report

Title Original Study: Decentralization Can Increase Cooperation among Public Officials

doi: https://doi.org/10.1111/ajps.12606, American Journal of Political Science

Report's Abstract: Molina-Garzón, Grillos, Zarychta, and Andersson (2022) examine how health sector decentralization affects cooperation between public officials. Using a public goods game conducted in Honduras, they find that officials who work under decentralized regimes contributed 0.8 more lempiras per round to a group solidarity fund, compared to officials who work under centralized regimes. They also find that most of this increase in investment under decentralized regimes occurred during rounds of the game in which the participants were able to communicate with each other. Finally, they find that decentralization was associated with a 14 percentage point increase in the proportion of potential cross-level network ties between participants that were realized. In this paper, I examine whether these results are robust to (1) the omission of some individual-level controls that may have been affected by the decentralization treatment, and (2) the use of a linear regression model instead of a Poisson regression model for the network analysis. I find that omitting the individual-level controls leads to similar conclusions about the effect of decentralization on individual contributions in the public goods game, but the interaction effect between decentralization and communication becomes statistically insignificant at the 0.05 level. For the network analysis, I find that using a linear regression instead of a Poisson regression has little bearing on the magnitude of the effect of decentralization on the proportion of ties realized, though the effect of decentralization becomes statistically insignificant for one version of the network model.

Link to Full Report: https://osf.io/q3dpt/

Link to Replicators' Package: https://osf.io/q3dpt/

Link to Original Authors' Response: https://osf.io/q3dpt/

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZLHYSZ

Reproduction Report

Title Original Study: Declining Worker Turnover: The Role of Short-Duration Employment Spells

doi: https://doi.org/10.1257/mac.20190230, American Economic Journal: Macroeconomics

Report's Abstract: Using a Diamond-Mortensen-Pissarides (DMP) model with noisy signals on worker-firm match quality calibrated on data from 30 US states for 1999 and 2017, Pries and Rogerson argue that improved screening may explain the decrease in short-term employment spells observed in the US labor market. Using a decomposition exercise in a "reduced form" model, the authors show that changes in short-term employment spells (and) are almost entirely accounted for by changes in the rate of learning on match quality and in the probability of a good match . Then, using a decomposition exercise in a "structural" model, they show in their main calibration strategy that changes in and are mainly driven by changes in and , parameters pertaining to learning about match quality. First, we reproduce the authors' codes in R and Python, two popular free open source programming languages. We find identical results to the paper. Second, we test the robustness of results to (1) using an earlier starting year, (2) adding additional states in the analysis, and (3) increasing the value of the 1999 mean vacancy duration parameter. The direction and relative size of the effect of each parameter on and is preserved in all robustness tests, corroborating the authors' argument.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/93.htm

Link to Replicators' Package: https://github.com/AlexandrePavlov/PriesRogerson2022Replication

Original Authors' Response: Declined to respond.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/120568/version/V1/view

Reproduction Report

Title Original Study: Digital Addiction
doi: https://doi.org/10.1257/aer.20210867, American Economic Review
Report's Abstract: Using an original economic model of digital addiction and a randomized experiment, Hunt Allcott, Matthew Gentzkow, and Lena Song (2022) isolate the effect of habit formation and self-control problems on how people use their smartphones. They find a persistent effect of temporary incentives on reducing social media usage. With the model-free results, the study shows that (after the incentive was in effect), participants in the bonus group reduced use by 56, 19 and 12 minutes in periods 3, 4 and 5,respectively, suggesting a persistent effect. But before the incentive was in effect in period 2, social media use reduced use by 5.1 minutes per day. Participants who used the limit functionality reduced FITSBY use by over 20 minutes per day, suggesting an impact of self-control problems on social media use. All these estimates are statistically significant. We perform a direct replication of the paper. Upon re-calculating the core dependent variable (FITSBY use by period), we find a small but concerning discrepancy: For a small number of observations, the aggregated dependent variable does not equal the sum of the disaggregated categories. Thankfully, this discrepancy does not have a major effect on the results. Using the provided data, we re-coded the core figures from scratch and found that we could replicate them all. We also compare the pre-analysis plan (PAP) with the main study to identify gaps and perform computational reproduction/replication of the structural model and model-free analysis. We only find minor differences between the PAP and the main paper, almost all of which are acknowledged in the paper.
Link to Full Report: https://osf.io/8kvdf/
Link to Replicators' Package: https://osf.io/8kvdf/
Link to Original Authors' Response: https://osf.io/8kvdf/
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/163822/version/V2/view

Reproduction Report

Title Original Study: Do Thank-You Calls Increase Charitable Giving? Expert Forecasts and Field Experimental Evidence

doi: https://doi.org/10.1257/app.20210068, American Economic Journal: Applied Economics

Report's Abstract: Samek and Longfield estimate the effect of 'thank you calls' on the extensive and intensive margins of subsequent donations. Based on a series of experimental interventions, the authors find no statistically discernable effect of thank-you calls on either the likelihood of donating again, or on the size of any subsequent donations made within the period of the study. In a companion exercise the researchers quantify the ability of experts in charitable fundraising and non-experts (using the Understanding America Survey) to predict the behaviours elicited by the experiment. Experts and non-experts (incorrectly) make the same predictions of an increase to the extensive margin of donation behaviour induced by the thank you call, and while both groups overestimate the intensive margin, the non-experts overestimated by a smaller magnitude. We were able to reproduce the papers findings completely, discovering only one difference in an appendix table related to the average gift amount — treatment for experiment 1 where only the constant term of the regression was affected. Upon careful examination of the code we found a few small errors that did not affect the results (one of the errors in the code did not seem to be carried through and used anywhere). Finally, we conducted several extensions of the original analysis which demonstrated that the findings are robust to heterogeneity of treatment effect by initial donation size, as well as different specifications of the regression analysis.

Link to Full Report: https://osf.io/fe2tr/

Link to Replicators' Package: https://gitlab.com/c3754/replication-games/-/tree/main/replication%20games%20MTL%202023%20charity?ref_type=heads

Link to Original Authors' Response: Waiting for the authors' response.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/149481/version/V1/view

Reproduction Report

Title Original Study: Do Transitional Justice Museums Persuade Visitors? Evidence from a Field Experiment

doi: https://doi.org/10.1086/714765, Journal of Politics

Report's Abstract: Balcells et al. (2022) explore the effect of transitional justice museums through a field experiment in Santiago, Chile, and attendance at the government's remembrance museum, the Museum of Memory and Human Rights which looks at the time of Pinochet's dictatorship. The authors want to understand how such experiences shape an individual's perceptions of trust in government institutions, and transitional justice policies, and how they are affected emotionally. Additionally, they seek to measure how long they last over time. They do this by creating treatment (museum attendance) and control (non-attendance) groups and administering pre-and post-treatment surveys and estimating the 'complier average causal effect' (CACE). They find that satisfaction with the current government significantly increases for the treatment group, looking over the entire population ($= 0.15, p = .04$) as measured with a 4-point Likert scale and support for a military government significantly drops by 11% ($= 0.11, p = .002$) across ideological stances. We first reproduce their results and find no major coding errors. Second, we test the robustness of the effects by 1) testing for heterogeneous effects by gender, 2) we combine the emotion variables into two indices, a mobilization and demobilization index, and 3) conduct a causal mediation analysis to see how confidence in the church may mediate effects found in the study.

Link to Full Report: https://osf.io/m3hwg/

Link to Replicators' Package: https://osf.io/m3hwg/

Original Authors' Response: "We thank all involved for their interest in our work. We are happy to hear that the results from our paper successfully replicated. We are intrigued by the additional analyses performed by the replicators. We hope their insights and results can inform future theorizing and empirical studies of the impact of Transitional Justice."

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=a651e0dbc9a5c140152aa84be2e0?persistentId=doi%3A10.7910%2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=type

Reproduction Report

Title Original Study: Does Competence Make Citizens Tolerate Undemocratic Behavior?

doi: https://doi.org/10.1017/S0003055422000119, American Political Science Review

Report's Abstract: We replicate the analysis conducted by Frederiksen, 2022a. We focus on assessing the computational and robustness replicability of their work. We find that their main exhibits and supplementary analysis are replicable, both when running their original Stata replication package, and when we attempt to replicate their findings from scratch in R. We also conduct additional robustness checks by estimating additional specifications and by subsetting the dataset by the time taken by the respondent to complete the survey. We again find that their work is robust to our battery of alternative specifications.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/28.htm

Link to Replicators' Package: https://github.com/tjbrailey/nottingham_replication_2023

Link to Original Authors' Final Response: "Thanks a lot for this initiative and not least for replicating my results."

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NGFLRO

Reproduction Report

Title Original Study: Does Patient Demand Contribute to the Overuse of Prescription Drugs?

doi: https://doi.org/10.1257/app.20190722, American Economic Journal: Applied Economics

Report's Abstract: We replicate Lopez et al.'s (2022) study on gatekeeping costs and the potential evidence for patient-driven and doctor-driven demand. Using their publicly available source materials, we first re-run their analysis "as is" to see if their results can be exactly replicated. We then expand the analysis to include patients previously excluded for not being acutely ill, offering a broader perspective on medication demand among all patient types. The findings confirm Lopez et al.'s results.

Link to Full Report: https://osf.io/x7g9z/

Link to Replicators' Package: https://osf.io/x7g9z/

Link to Original Authors' Response: Provided feedback to an initial report. Final response: " Thank you very much for sharing the updated report. We appreciate that the authors of the replication reworked the paper and have no further response or comments."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/126722/version/V1/view

Reproduction Report

Title Original Study: Does Public Opinion Affect the Preferences of Foreign Policy Leaders? Experimental Evidence from the UK Parliament
doi: https://doi.org/10.1086/719007, Journal of Politics
Report's Abstract: The study by Chu and Recchia (2022) tests the hypothesis that providing public opinion information can shift policymakers' opinions in the direction of what the public favors. They surveyed 101 British Members of Parliament (MPs) about their views regarding the United Kingdom's presence in the South China Sea. Their results demonstrated that MPs who received information about the public opinion poll expressed viewpoints closer to that of public opinion. The authors reported an effect that is "substantively meaningful and statistically significant at the .10 level." Our computational replication of the original study found that the paper is fully computationally reproducible. We successfully replicated the authors' results but found that the main findings are no longer significant when analyzed using unweighted data (see Table 1). We also conducted several robustness checks on sub-samples of the data to examine the key analyses both with and without weights. Here, we found that the results are once again robust and significant when weights are used, but no longer significant when weights are not used. As a further robustness check, we found no moderating effect of gender. Overall, our replication efforts suggest that the main finding of the original study may be sensitive to the use of survey weights.
Link to Full Report: https://osf.io/bqz6w/
Link to Replicators' Package: https://osf.io/vwt2n/?view_only=84e52a7c684942a4880410b3c89ff4c6
Original Authors' Response: " Thank you for your note and engaging with our work. We don't have a formal reply, though this is an honest question: isn't it standard practice to use weights when using YouGov's data, since making valid claims about representativeness depends on using their weights? YouGov's MP panels operate similarly to their public opinion poll, in that their claims to representativeness rely on using weights, provided by YouGov. I [Chu] think your write-up mentioned that there's a debate about using weights, and cited MTurk data, but I think that MTurk is quite different, and yes, I agree I do not use weights for MTurk data except unless requested by a reviewer for robustness checks, etc.. But I don't think MTurk and the MP representative poll we used is a good comparison in the context of evaluating the validity of weighting. In any case, happy to adapt if there is a clear consensus on this. Thanks again."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BNINNL

Reproduction Report

Title Original Study: Effective for Whom? Ethnic Identity and Nonviolent Resistance

doi: https://doi.org/10.1017/S0003055421000940, American Political Science Review

Report's Abstract: Manekin and Mitts (2022) investigate the success chances of minority ethnic groups when engaging in non-violent protests demanding political change. First, using observational data, the authors find that the success rate for nonviolent campaign tactics is lower for excluded/minority ethnic groups than for non-excluded/majority ethnic groups. Second, the authors use two original survey experiments to show that non-violent protest by ethnic minorities is perceived as more violent and requiring more policing than identical protest by majorities. This report reproduces the paper computationally and conducts several sensitivity analyses for both the observational and the experimental parts of the paper. We can confirm the general direction of the postulated effects, but evidence becomes less consistent (effect magnitudes and significance levels are not robust to some of the changes).

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/86.htm

Link to Replicators' Package: https://zenodo.org/records/10193470

Original Authors' Response: Cannot provide a response.

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SHHVCA

Reproduction Report

Title Original Study: Enabling or Limiting Cognitive Flexibility? Evidence of Demand for Moral Commitment

doi: https://doi.org/10.1257/aer.20201333, American Economic Review

Report's Abstract: We computationally reproduce Saccardo and Serra-Garcia (2023) where subjects exploit cognitive flexibility by viewing their incentives first and partially ignoring product quality information, and hence, recommend the incentivized product. We find one major coding error for the variable Selfishness. Additionally, two of the "moral cost" questions more likely capture spitefulness. After correcting the erroneous coding or dropping the two questions, we find stronger support for the authors' main conclusion regarding Selfishness driving incentive information avoidance with double effect size. Finally, we find weak evidence that subjects update their posterior beliefs differently depending on the product they are incentivized to recommend.

Link to Full Report: https://osf.io/nwds7

Link to Replicators' Package: https://osf.io/yfdet/

Link to Original Authors' Response: https://www.aeaweb.org/doi/10.1257/aer.20201333.appx

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/180741/version/V1/view

Reproduction Report

Title Original Study: Entertaining Beliefs in Economic Mobility
doi: https://doi.org/10.1111/ajps.12702, American Journal of Political Science
Report's Abstract: In Entertaining Beliefs in Economic Mobility (AJPS 2023) Kim finds that watching "rags-to-riches" style reality TV programs strengthens Americans' belief in the American dream. Through thoughtful and clever experimental and observational analysis, she demonstrates that exposure to television programs containing everyday people working hard to earn large prizes increases Americans' belief that success can be internally attributed and that economic mobility is possible. We computationally replicate Kim's results, finding no major errors in her coding or statistical procedure. We also include several robustness checks. First, we merge her two experimental samples, which increases the precision of her main quantity of interest such that it attains conventional levels of statistical significance. Second, we recreate tables and visualizations for alternative specifications of her main observational results. The original results are robust to these alternative models, but we do find that if sports programming is operationalized in the same manner as "rags-to-riches" programming, the sign, magnitude, and significance of watching either programming type are similar. We also uncover a partisan interaction effect, as only Democrats change their beliefs in economic mobility with increased TV viewing.
Link to Full Report: https://osf.io/xf5w2/
Link to Replicators' Package: https://github.com/jacobawinter/rep_games_2023
Original Author's Response: "Thanks for this! I have no particular response per se. I'm grateful for your collective efforts to make social science much more transparent."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FVRZYU

Reproduction Report

Title Original Study: Evaluating Deliberative Competence: A Simple Method with an Application to Financial Choice

doi: https://doi.org/10.1257/aer.20210290, American Economic Review

Report's Abstract: Ambuehl et al. (2022) explore ways to evaluate interventions designed to enhance decision-making quality when individuals misjudge the outcomes of their choices. The authors propose a novel outcome metric that can distinguish between interventions better than conventional metrics such as financial literacy and directional behavioral responses. The proposed metric, which transforms price-metric bias into interpretable welfare loss measures, can be applied to evaluate various training programs on financial products. Table 4 of the paper reports the authors' significant main point estimates at the 1% level. In this replication exercise, we first replicate the main findings of the original paper. Then, we modify the clustering method by using k-means with demographic variables as inputs, then we re-calculate standard errors with jackknife estimators. Finally, we include subjects who were excluded by the authors due to multiple switching in the multiple price lists. We find that all of these replications result in robust findings. Additionally, we successfully replicate Figure 4 from the paper. Notably, this replication demonstrates the insensitivity of the results to the choice of distance metric.

Link to Full Report: https://osf.io/scgbt/

Link to Replicators' Package: https://osf.io/scgbt/

Link to Original Authors' Response: Authors provided feedback.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/171681/version/V1/view

Reproduction Report

Title Original Study: Exposure and Preferences: Evidence from Indian Slums
doi: https://doi.org/10.1111/ajps.12570, American Journal of Political Science
Report's Abstract: Successful computational reproducibility. The replicators could not conduct the robustness checks without the help of the author.
Link to Full Report: No report.
Original Author's Response: "Thanks for your email. I am not interested in participating."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AV8PLT

Reproduction Report

Title Original Study: Finance and Green Growth
doi: https://doi.org/10.1093/ej/ueac081, Economic Journal
Report's Abstract: De Haas and Popov (2023) estimate the effect of country-level financial sector size and structure on decarbonization to show that countries with relatively more equity versus debt financing have more emission-efficient economies. We uncover multiple coding errors that change the magnitude and the precision of the coefficients of interest. These coding errors include misreporting of standard errors, and misspecifying generalized method of moments (GMM) estimators. We further provide robustness tests of the results to (1) restricting the sample to consistent sets of countries across the country and country-by industry samples, and (2) using a limited information maximum likelihood (LIML) estimator to address a weak-instrument problem. We find that the results from the robustness checks are qualitatively different from the original results but similar to the corrected results.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/95.htm
Link to Replicators' Package: https://osf.io/h8ct2/
Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/96.htm
Original Authors' Package: https://zenodo.org/records/7220094

Reproduction Report

Title Original Study: Flight to Safety: COVID-Induced Changes in the Intensity of Status Quo Preference and Voting Behavior

doi: https://doi.org/10.1017/S0003055421000691, American Political Science Review

Report's Abstract: Bisbee and Honig (2022) examine the effect of the COVID-19 pandemic on voting for Bernie Sanders in the 2020 Democratic Party primary using a difference-in-differences design, finding evidence that exposure to COVID-19 resulted in a 7-15 percentage point increase in voting for Biden. The study also uses a regression design with district-level fixed effects to estimate the effect of the COVID-19 pandemic on voting for anti-establishment candidates during the US 2020 House primaries. It finds evidence that an increase in COVID cases was associated with a decline in voting for anti-establishment candidates in general, and for those endorsed by the Tea Party. We re-run the code for all tests in this paper, successfully reproducing its results in a preliminary replication. We then use the De Chaisemartin and D'Haultfoeuille difference-in-differences estimator to replicate their main results, finding that though the coefficient remains negative, the results are not statistically significant. We also replicate their tests regarding US House primary candidates using a different measure of anti-establishment candidates. Here, we find that the interaction term between anti-establishment candidates and COVID-19 remain statistically significant, with the same sign. Finally, we employ an expanded dataset that includes Congressional primary candidates that were omitted in the initial dataset, as well as a re-coded extremism variable that also includes candidates endorsed by Donald Trump. These updated findings corroborate the paper's initial results. However, due to a restrictive number of observations that interfered with our application of the De Chaisemartin and D'Haultfoeuille estimator, we believe that the expanded U.S. House primary results constitute the more robust half of our replication.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/36.htm

Link to Replicators' Package: https://github.com/Dmscates/Bisbee-and-Honig-2022-Flight-to-Safety-Replication

Original Authors' Response: "You guys are amazing. Thank you for doing this! We are impressed by your rigor and grateful for the introduction to DCD'H DiD estimator that we'll have to add to the repertoire. We were working on the conditional accept when the flurry of generalized DiD work (Goodman-Bacon, Callaway, etc.) was blowing up [...] We also appreciate the manner in which you communicated with us during the course of your re-analysis, and the thoughtfulness of your report. [...] Although I'm sure it is a logistical nightmare and likely would add even more delays to the publication pipeline, it would be very pro-science if this type of replication were part of a journal's own pre-publication process. (This is what I naively thought replication meant back when I got my first publication, and have been disappointed in the process ever since.)"

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S5YMS7

Reproduction Report

Title Original Study: Gender Differences in Cooperative Environments? Evidence from The U.S. Congress

doi: https://doi.org/10.1093/ej/ueab069, Economic Journal

Report's Abstract: Gagliarducci and Paserman (2022) study gender differences in cooperative behavior among politicians using information from the U.S. House of Representatives between 1988 and 2010 on (i) the number of co-sponsors on bills and (ii) the share of co-sponsors from the rival party. Through different empirical strategies, they show that women-sponsored bills tend to have more co-sponsors, but the gap is only statistically significant among Republicans. Moreover, Republican women recruit a significantly larger share of co-sponsors from the rival party than Republican men, whereas the opposite is true among Democrats. GP argue that the observed pattern is consistent with a commonality of interest driving cooperation, rather than gender per se, since during this period Republican women were ideologically closer to the rival party than their male colleagues, while female Democrats were further away. We examine the robustness of these findings to (i) the correction of some errors in two control variables of the dataset used by GP and (ii) clustering the standard errors at the individual level, instead of individual-term. These changes have a relatively minor impact on results: most coefficients are still statistically significant and the main conclusions from the analysis are confirmed. Furthermore, we extend the analysis to the 2011-2020 period. The analysis of gender differences in bipartisan cooperation confirms GP's hypothesis that ideological distance plays an important role. However, results are slightly different when we analyze overall cooperation. The gender gap in favor of women is larger in magnitude than in GP and it is statistically significant in several specifications, providing support for the hypothesis that gender also matters for cooperation.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/75.htm

Link to Replicators' Package: https://www.dropbox.com/scl/fo/dmvgx5wlgql3tz98dx47u/h?rlkey=af83xiqrkw70rbjicdunoach9&dl=0

Link to Original Authors' Response: https://osf.io/w48tf/

Original Authors' Package: https://zenodo.org/records/5111360

Reproduction Report

Title Original Study: Good Reverberations? Teacher Influence in Music Composition since 1450

doi: https://doi.org/10.1086/718370, Journal of Political Economy

Report's Abstract: Borowiecki (2022) studies the influence of teachers on the style of their students in the domain of musical composition. The author finds that realized student-teacher pairs are on average 0.2-0.3 standard deviations more similar to unrealized, but possible, student-teacher pairs. In this report we provide the results of our replication of Borowiecki (2022). We direct our attention to the following tasks: 1) Replicating the outcome variables used in the paper, starting from the raw data, and generating alternative measures of similarity between students and teachers 2) Testing the validity of the random teacher-student pairing, a key assumption for the validity of the estimation strategy employed in the paper. We can replicate most of the outcome variables, but not all of them, due to incomplete raw data. Our alternative measures of similarity confirm the robustness of the original results. We find significantly different characteristics between paired and unpaired students, suggesting that matching between students and teachers does not occur randomly. However, controlling for these characteristics in the main regressions leads to quantitatively similar results to the ones reported in the original paper.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/27.htm

Link to Replicators' Package: https://www.dropbox.com/scl/fo/6hecmgjsq3mjo5ekkv8lm/h?rlkey=ftuoe4mf5f9jon0hiabb4brtn&dl=0

Link to Original Authors' Response: https://osf.io/79e2z/

Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/718370/suppl_file/20210405data.zip

Reproduction Report

Title Original Study: Hate Crimes and Gender Imbalances: Fears over Mate Competition and Violence against Refugees

doi: https://doi.org/10.1111/ajps.12595, American Journal of Political Science

Report's Abstract: Dancygier et al. (2022) ascribe anti-refugee hate crime in Germany from 2015 to 2017 to the fear of mate competition felt by native German men, amplified by growing refugee populations and existing gender gaps. In a replication of this article, we discovered that the substantively and statistically significant relationship between perceptions of mate competition and support for anti-refugee violence found in a 2016–17 survey of adults in Germany were robust when analyzed with ensembles of regression trees permitting arbitrary interactions in a large design matrix. However, statistically significant pairwise comparisons between survey respondents' perceptions of mate competition across strata of the municipality-level gender gap as recorded by German censuses were not robust to controlling the family-wise Type I error rate. Moreover, statistically significant relationships between the gender gap and the incidence of hate crime in Germany in the authors' panel regressions vanished in a wide range of specifications with municipality fixed effects—in certain cases, being replaced with statistically significant estimates of the opposite sign.

Link to Full Report (and Initial Version of the Report): https://osf.io/5n3ds/

Link to Replicators' Package: https://osf.io/5n3ds/

Link to Original Authors' Response: https://osf.io/5n3ds/

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QXJDJ5

Reproduction Report

Title Original Study: Historical Lynchings and the Contemporary Voting Behavior of Blacks

doi: https://doi.org/10.1257/app.20190549, American Economic Journal: Applied Economics

Report's Abstract: Williams (2022) ties the political participation of Blacks to historical lynchings that occurred in the United States. Her findings document lower Black voter registration rates in southern counties with greater number of historical lynchings. We show that this effect is driven by four outlier counties with relatively high Black lynching rates. Excluding these counties from the analysis yields a point estimate that is no longer statistically significant. Dropping the ninety-fifth percentile lynching rates and correcting the errors in voter registration rates rule out the effect size reported by Williams (2022), which now becomes close to zero and statistically insignificant. We also show that the main results are highly sensitive to the way lynching and voter registration rates are measured.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/32.htm

Link to Replicators' Package: https://osf.io/hv7wp/

Original Author's Response: No response to our emails.

Original Author's Package: https://www.openicpsr.org/openicpsr/project/136741/version/V1/view

Reproduction Report

Title Original Study: Hobo Economicus
doi: https://doi.org/10.1093/ej/ueab103, Economic Journal
Report's Abstract: Peter Leeson, August Hardy and Paola Suarez (2022) test maximizing behaviour of panhandlers at several Metrorail stations in Washington, D.C. Their main findings are that "stations with more panhandling opportunities attract more panhandlers" (the first statement) and that "cross-station differences in hourly panhandling receipts are statistically indistinguishable from zero" (the second statement). We test computational reproducibility and robustness replicability of their results. We can reproduce both statements, in Stata and R. Our robustness replications for the first statement confirm the authors' results in the vast majority of cases (replication was successful in 91% of the cases). Our robustness replications for the second statement might raise doubts on this finding. We run weighted ANOVA tests, we change the bounds in minutes used by authors by 5 minutes in their robustness checks, we run Bartlett's tests of equality of variances of means, and run pair-wise tests of equality of means. In three out of four cases we cannot replicate the results, and the differences (of either means, medians or variances of donations) across Metrorail stations are statistically different from zero. We hypothesize that panhandlers have a general idea about which stations have more passers-by, and will rationally go more often there. However, they are unlikely to have information about smaller variations in the number of passers-by (e.g., variations in passers-by at the same station over time due to non-public events), and therefore might find it difficult to perfectly maximize donations.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/55.htm
Link to Replicators' Package: https://osf.io/s4bca/
Original Authors' Response: Declined to respond.
Original Authors' Package: https://zenodo.org/records/5719541

Reproduction Report

Title Original Study: How Do Beliefs about the Gender Wage Gap Affect the Demand for Public Policy?
doi: https://doi.org/10.1257/pol.20200559, American Economic Journal: Economic Policy
Report's Abstract: We conduct a replication of Settele (2022), a online survey experiment designed to find out how individual's beliefs about the gender wage gap affect their policy preferences. We reproduce Results 1 and 2 of the study: how prior beliefs around the wage gap are distributed among individuals and how a information treatment causally affects the policy demand. Our re-coded replication shows that the reported results are robust.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/12.htm
Link to Replicators' Package: https://osf.io/j2ubt/
Original Authors' Response: "I very much appreciate the effort of you and your team to replicate not just my paper, but many others too. It is quite impressive to see the scope of your project and I am curious about your future plans with this initiative.

I just read the Reproduction Report for my paper and I think it is great. In particular, Figures 2 and 3 are really cool. (They are actually new, and offer a really insightful way of looking at the data. I should have come up with them myself!)

Regarding your question, I don't think the Reproduction Report requires a formal response from my side. I fully agree with the authors' interpretation of the results, and just want to say thank you for their great work. Please go ahead and publish the report whenever it suits you."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/134041/version/V1/view

Reproduction Report

Title Original Study: How Effective Are Monetary Incentives to Vote? Evidence from a Nationwide Policy?

doi: https://doi.org/10.1257/app.20200482, American Economic Journal: Applied Economics

Report's Abstract: Successful computational reproducibility. No re-analyses conducted.

Link to Original Authors' Response: Not contacted.

Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip

Reproduction Report

Title Original Study: How Much Should We Trust the Dictator's GDP Growth Estimates?

doi: https://doi.org/10.1086/720458, Journal of Political Economy

Report's Abstract: In this brief commentary, we have conducted a robustness reproducibility and replicability of Martinez's 2022 paper entitled "How much should we trust the dictator's GDP growth estimates?" by selecting different clusters and omitting fixed effect terms. Concurrently, we conduct sub-sample analyses and employ alternative measurements for the sake of robustness and direct replicability. Our results are generally robust, yet they also raise some intriguing questions. First, we attempt to remove the year fixed effect in the model specifications, but the elimination of the year fixed effect from the baseline equation did not account forunobserved variables across year, suggesting the variable bias by Oster (2019). Second, the entirety of the baseline results is influenced by the periods 2007-2013 (for a five-year interval) and 2010-2013 (for a three-year interval). Third, when utilizing a more varied dataset for the autocracy measurement, the effect vanished for countries that are partially unfree.

Link to Full Report: https://osf.io/4sk52/

Link to Replicators' Package: https://osf.io/4sk52/

Original Author's Final Response: "As before, please extend my gratitude to the replicators for their thoughtful work and for taking into consideration my previous comments. I am reassured by the fact that they were able to replicate all the original results in the paper. I also find it reassuring that the results prove robust to additional robustness tests concerning alternative clustering structures for the standard errors or alternative data sources on political regimes (albeit with some loss of precision). The heterogeneous effects by subperiod are also quite intriguing and potentially reflect changes in the geopolitical incentives to overstate GDP growth in non-democracies. My paper is certainly not the final word on this topic and these results could be the first step towards new and exciting research."

Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip

Reproduction Report

Title Original Study: Ideological Asymmetries and the Determinants of Politically Motivated Reasoning

doi: https://doi.org/10.1111/ajps.12624, American Journal of Political Science

Report's Abstract: Guay and Johnston (2022) examine asymmetric politically motivated reasoning on the part of liberals and conservaites. In our replication of the paper we examine four potential issues with the analysis: confounding in the numeracy task, heterogeneity across ideological constraints, the use of control variables, and heterogenity in the moderator index items. None of these potential issues are in fact issues. The results are quite robust. We found only one minor issue with the codebook, which does not affect the results.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/79.htm

Link to Replicators' Package: https://osf.io/mh5sk/

Link to Original Authors' Response: "Thank you again for examining our paper so closely [...] we changed the codebook and appreciate this replication effort."

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CGHTPZ

Reproduction Report

Title Original Study: Immigration and Redistribution
doi: https://doi.org/10.1093/restud/rdac011, Review of Economic Studies
Report's Abstract: Alesina et al. (2023) examine how people perceive the number and characteristics of migrants and how those perceptions affect their support for redistribution. They find that respondents from the United States, United Kingdom, Sweden, Italy, Germany and France markedly overestimate the share of immigrants in each country, with the average respondent in all countries except Sweden overestimating by more than a factor of two. We reproduce these results using the original code and data and test the robustness by (i) including participants excluded for time to complete the survey, (ii) extending the analysis of misperceptions to all survey respondents, and (iii) using alternative authoritative estimates of the proportion of immigrants. We find that these checks marginally change the estimates of the size of the misperception but do not change the conclusions to be drawn from the analysis. Alesina et al. (2023) also test the effect on support for redistribution of showing videos on immigrant characteristics. We computationally reproduced the treatment effects on support for redistribution.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/40.htm
Link to Replicators' Package: https://osf.io/ajm9g/
Original Authors' Response: "Dear Institute for Replication team,

Thank you very much for taking the time to replicate our paper. We appreciate the important work you do. We are happy to see that our results replicated well and that our robustness checks were confirmed.

With best wishes, Armando Miano and Stefanie Stantcheva"
Original Authors' Package: https://zenodo.org/records/5997521

Reproduction Report

Title Original Study: Indecent Disclosures: Anticorruption Reforms and Political Selection

doi: https://doi.org/10.1111/ajps.12646, American Journal of Political Science

Report's Abstract: This short report summarises a replication exercise performed on data from Szakonyi (2021). The original work applies a difference-in-differences design to the case of an anti-corruption reform implemented in Russia for local election candidates, mandating financial disclosures. The author applies this design by comparing the electoral outcomes of municipalities that happened to hold elections right after the reform with those that held elections right before the reform. For both groups, the design uses information from the previous electoral cycle as a pre-treatment benchmark. Using only data provided by the author in the original dataset, I first verified that results are reproducible when using alternative software. Second, I performed two simple placebo tests to obtain evidence on violations of the design's identifying assumptions. These placebo tests return null results, reassuring on the reproducibility of the original findings.

Link to Full Report: https://osf.io/gx4d6/

Link to Replicators' Package: https://osf.io/gx4d6/

Original Author's Response: "Many thanks to the replicator for taking the time to replicate and extend the paper. The placebo tests are very helpful in illuminating whether the identifying assumptions hold. I will make sure to run versions of these in future analyses."

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KDUMRM

Reproduction Report

Title Original Study: Inflammatory Political Campaigns and Racial Bias in Policing

doi: https://doi.org/10.1093/qje/qjac037, Quarterly Journal of Economics

Report's Abstract: Grosjean et al. (2023) (GMY2023) estimate the causal effect of a Trump rally on the number of black drivers stopped by police officers, using a difference-in-difference approach. In their preferred specification, the authors find that after a Trump rally, the probability that a stopped driver is black increases by 5.74%. This effect is significant at the 1% level. In this report we focus on reproducing the main claim of the paper. First, we reproduce the paper's main findings and uncover an issue with counties that experience multiple Trump rally treatments, given the original modelling choices taken in GMY2023. When we remove counties that experience multiple rallies, the estimated effect size drops to 2.46% and loses statistical significance. Second, we attempt to conduct a direct replicability check, by employing a new data set as a source for the dependent variable. We use data from the National Incident Based Reporting System (NIBRS). We observe no effect of Trump rallies both on the original data, covered by NIBRS and on the NIBRS data. Third, we conduct a robustness replicability exercise by coding an event-study difference-in-difference design at the day level. We estimate the event-study in a $[7; +7]$ window. We do not discover any systematic effect of Trump rallies on the dependent variable from GMY2023.

Link to Full Report: https://osf.io/xadb6/

Link to Replicators' Package: https://osf.io/c7j58/

Original Authors' Response: https://osf.io/xadb6/

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/A3B9HE

Reproduction Report

Title Original Study: Interaction, Stereotypes, and Performance: Evidence from South Africa

doi: https://doi.org/10.1257/aer.20181805, American Economic Review

Report's Abstract: Corno, La Ferrara and Burns (2022) exploit the random allocation of freshman roommates in a large South African university to gauge the impact of a roommate's race on racial attitudes as measured by an implicit association test, and on school performance. They notably find that (a) white students randomly assigned to black roommates have less negative racial stereotypes, and (b) black students randomly assigned to live with white students have higher GPAs. We first reproduce all regression tables in Corno et al. (Corno et al. (2022)), and then test for robustness by varying the controls and conducting influential analysis. Overall, we find the results for finding (a) and (b) and robust in 15% and 40% of the robustness checks we ran, and the t/z scores are on average 78% and 85% as large as the original study.

Link to Full Report: https://osf.io/w7vpu/files

Link to Replicators' Package: https://osf.io/w7vpu/files

Link to Original Authors' Response: Did not receive formal response as of November 2025.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/174501/version/V1/view

Reproduction Report

Title Original Study: Interventions and Cognitive Spillovers
doi: https://doi.org/10.1093/restud/rdab087, Review of Economic Studies
Report's Abstract: In the paper of, Altmann et al. (2022) the authors investigate whether positive effects which are due to behavioral policy interventions in policy-targeted domains come along with negative effects in policy non-targeted domains. Using lab and online experiments where subjects have to solve one policy-focused decision task and one non-focused background task, the authors show that increasing incentives or steering attention to the former led to higher attention spans, lower default adherence rates, and a higher choice quality in the decision task. However, because of steering participants focus to the decision task, lower choice quality and lower attention spans in the background task emerged as a consequence, which was particularly pronounced among individuals with lower cognitive capabilities and complex decision tasks. Essentially, the authors also describe that the negative effects in the background tasks offset the positive effects in the decision task, ultimately yielding a net-zero effect overall. Therefore, the authors emphasize policymakers to also consider the potential negative cognitive spillovers in order to not overestimate the benefits of behavioral policy interventions. All the results the authors in the main text report are significant on 5% and 1% significance levels. All findings presented in the main text of the paper can be replicated using the original Stata code and verified thoroughly using R. Additionally, we performed two robustness tests to ensure the reliability of the paper's main results, and they remained consistent. Hence, the reported findings in the paper appear to be robust.
Link to Full Report: https://www.econstor.eu/bitstream/10419/272845/1/I4R-DP043.pdf
Link to Replicators' Package: https://osf.io/kugbs/
Original Authors' Response: "We do not have any comments regarding the replication. We just want to briefly express our gratitude for the thorough and excellent work of the authors of the replication study."
Original Authors' Package: https://zenodo.org/records/5652808

Reproduction Report

Title Original Study: Jumping the Gun: How Dictators Got Ahead of Their Subjects

doi: https://doi.org/10.1093/ej/ueac073, Economic Journal

Report's Abstract: Hariri and Wingender add new nuance to the traditional wisdom that economic modernisation is a path to democracy. They show that the diffusion of repressive, military technologies, causes a decline in the number of democratisations in the following years, and argue that this is because of a greater ability to forcefully oppress popular dissent. We conduct a robustness replication exercise, focussed on three tests: i) Are findings robust to alternative weightings of individual technologies in the instrument for country-aggregate military technology? ii) Is high leverage in individual countries, regions or time periods driving the global findings? iii) Are the strength of the IV and its independence of important macroeconomic indicators a chance occurrence? The main findings of the paper are largely robust to these tests.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/50.htm

Link to Replicators' Package: https://osf.io/4cx86/

Original Authors' Response: "we do not have any comments to the Reproduction Report, so I'm just sending you this email to applaud the initiative. You and the Institute for Replication is doing a great service to the profession."

Original Authors' Package: https://zenodo.org/records/7077694

Reproduction Report

Title Original Study: Liquidity Constraints in the U.S. Housing Market
doi: https://doi.org/10.1093/restud/rdab063, Review of Economic Studies
Report's Abstract: Successful computational reproducibility. No re-analyses conducted.
Link to Original Authors' Response: Authors provided feedback and suggestions.
Original Authors' Package: http://doi.org/10.5281/zenodo.5112964

Reproduction Report

Title Original Study: Local Elites as State Capacity: How City Chiefs Use Local Information to Increase Tax Compliance in the Democratic Republic of the Congo
doi: https://doi.org/10.1257/aer.20201159, American Economic Review
Report's Abstract: Balán et al. (2022) evaluate the impact of "local elites" involvement in local tax collection in a large city in the Democratic Republic of Congo. Using a randomized controlled trial to vary the identities of tax collectors, they find that local elites' involvement raises tax compliance and total revenue by 50 and 44 percent, respectively. The paper argues that the primary mechanism behind the results is better targeting made possible by local elites' superior information about property holders' willingness and ability to pay. In this replication comment, we first reproduce the paper's main results. Then, we assess the robustness of the results by (1) employing randomization inference for statistical tests; (2) controlling for baseline characteristics that are not balanced; and (3) using an alternative method to examine the claims supporting the preferred mechanism of better targeting. We find robust estimates in (1). However, the results are less robust both in terms of statistical significance and magnitude for (2) and (3). We conclude that the average treatment effect is robust, while the main claim about mechanisms, the information channel, is less robust to alternative estimation approaches. We contextualize and discuss the significance of these results, including the negligible revenue potential even under full compliance.
Link to Full Report: https://ideas.repec.org/p/zbw/i4rdps/191.html
Link to Replicators' Package: https://github.com/SossouAdjisse/LocalTaxReplicationProject.git
Link to Original Authors' Response: https://ideas.repec.org/p/zbw/i4rdps/192.html
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/147561/version/V1/view

Reproduction Report

Title Original Study: Major Reforms in Electricity Pricing: Evidence from a Quasi-Experiment

doi: https://doi.org/10.1093/ej/ueab076, Economic Journal

Report's Abstract: Labandeira et al. (2022) examine the effect of a policy in Spain that modified the electricity bill structure for all Spanish households. The policy simultaneously increased fixed costs and decreased marginal costs on household electricity bills. Using fixed effects and instrumental variables (IV) specifications, the main causal finding in the paper is that the reform reduced house- hold electricity consumption for Spanish households by 15%. Their point estimate is statistically significant at the 1% level. In a similar specification, they find the reform reduced household expenditures on electricity by 9.8%, statistically significant at the 1% level. The code provided by the authors is computationally reproducible. We found two coding errors in different IV specifications, which had served as robustness checks to their main results. Correcting the errors removes statistical significance in two of four IV results, but increases the point estimates and statistical significance in the other two IV results. We also perform robustness checks. The IV estimates lose statistical significance in two of four robustness checks (with point estimates changing 1.1% to -39%). However, the OLS regressions are robust to changing covariates (sign and significance remained for 12 of 14 tests of the OLS specification, with changes in the estimates ranging from -157% to 64%, but averaging -3.3%).

Link to Full Report: https://osf.io/bysa7/

Link to Replicators' Package: https://osf.io/bysa7/

Original Authors' Response: Original authors provided feedback. Multiple rounds of back and forth with replicators.

Original Authors' Package: https://zenodo.org/records/5423782

Reproduction Report

Title Original Study: Market Access and Quality Upgrading: Evidence from Four Field Experiments

doi: https://doi.org/10.1257/aer.20210122, American Economic Review

Report's Abstract: Bold et al. (2022b) investigate the effect of providing access to a market (i.e. a buyer) which rewards quality with a premium on farm productivity and farming incomes from smallholder maize farmers in western Uganda, using a series of randomized experiments and a difference-in-differences approach. We successfully reproduce the results of this study using the publicly provided replication packet. Then test the robustness of these results by re-defining treatment and outcome variables, testing for model misspecification and the leverage of outliers, and testing for non-random selection in the Fisher-permutation process. Our results show that the findings in Bold et al. (2022b) are robust to a variety of decisions in the research process. This evokes confidence in the internal validity of the findings.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/72.htm

Link to Replicators' Package: https://journaldata.zbw.eu/dataset/bold-et-al-american-economic-review-2022

Original Authors' Response: "Thank you very much for sharing the report (and taking the time to replicate the study). We have no comments."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/158401/version/V1/view

Reproduction Report

Title Original Study: Market-Based Monetary Policy Uncertainty
doi: https://doi.org/10.1093/ej/ueab086, Economic Journal
Report's Abstract: Bauer et al. (2022) derive market-based monetary policy uncertainty and uncover an 'FOMC uncertainty cycle' characterized by a fall of uncertainty after FOMC announcements and its subsequent built-up. Then, the authors show that the financial markets' response to monetary policy announcements depends on the level of short-rate uncertainty on the day before the FOMC announcement. First, we reproduced the paper's findings, though with Matlab version-specific issues. Second, we tested the robustness of the two main results of the paper. We show that the uncertainty cycle in the monetary policy uncertainty is confirmed when the crisis period is included in the sample or when the median instead of the average of changes in the monetary policy uncertainty is considered. However, the FOMC uncertainty cycle does not appear when the monetary policy uncertainty index (Husted et al. 2020) or the daily economic policy uncertainty index (Baker et al. 2016) are used as uncertainty proxies.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/77.htm
Link to Replicators' Package: https://osf.io/qx8aw/
Original Authors' Response: "Thank you, glad to see that this work found our results to be rock solid!

We won't write a response. Do let us know if you have any other questions about our work."
Original Authors' Package: https://zenodo.org/records/5566246

Reproduction Report

Title Original Study: Market-Based Monetary Policy Uncertainty
doi: https://doi.org/10.1093/ej/ueab086, Economic Journal
Report's Abstract: This report replicates and examines Bauer et al.'s (2021) paper on monetary policy transmission to financial markets. The paper introduces novel measures of monetary policy uncertainty and analyses its drivers. It also investigates the impact of uncertainty changes on interest rates and financial asset prices. We assess reproducibility, consolidate market uncertainty measures using PCA and Factor Analysis, and rigorously test the reduction of uncertainty after Federal Market Open Committee (FOMC) announcements. Our findings support the paper's claim of reduced uncertainty on meeting days. Additionally, we explore the implications of the uncertainty channel on various financial assets, such as Gold, the Swiss Franc, European stock indexes, and Bitcoin.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/76.htm
Link to Replicators' Package: https://github.com/YaolangZhong/Nottingham_Replication_Game/tree/main/replication_code
Original Authors' Response: "Thank you, glad to see that this work found our results to be rock solid!

We won't write a response. Do let us know if you have any other questions about our work."
Original Authors' Package: https://zenodo.org/records/5566246

Reproduction Report

Title Original Study: Measuring the Welfare Effects of Shame and Pride
doi: https://doi.org/10.1257/aer.20190433, American Economic Review
Report's Abstract: This Reproduction Report examines and extends the research conducted by Butera, Metcalfe, Morrison, and Taubinsky (2022) on "The Welfare Effects of Pride and Shame." The original paper explores the welfare implications of public recognition as a motivator for desirable behavior and introduces an empirical methodology to measure Public Recognition Utility (PRU), which quantifies the utility individuals experience when their actions are publicly recognized. This report focuses on the real effort experiment reported in the paper that was conducted using a classroom sample, a lab sample, and an online sample. I computationally reproduce the original results and verify their robustness. While reproducing the results, I found two minor coding errors in the replication package. Correcting these errors slightly changes some estimates reported in the paper but does not turn over any results. The main treatment effect findings are further robust to using different sets of controls and sample selection criteria. Moreover, I conduct a heterogeneity analysis which reveals significant variations in how participants value public recognition. Overall, the replication study confirms the original conclusions while providing additional insights into the heterogeneity of PRU shapes on an individual level.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/64.htm
Link to Replicators' Package: https://github.com/tilmanfries/welfare-shame-pride-replication-report
Original Authors' Final Response: "Thanks again for all your hard work on this."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/145141/version/V1/view

Reproduction Report

Title Original Study: Mental Health Costs of Lockdowns: Evidence from Age-Specific Curfews in Turkey

doi: https://doi.org/10.1257/app.20200811, American Economic Journal: Applied Economics

Report's Abstract: This report presents a replication of Altindag et al. (2022) performed at the Olso Replication Games in 2022. Altindag et al. (2022) estimate the effects of an age-specific lockdown on mental health outcomes and mobility among adults aged 65 and older in Turkey, using a regression discontinuity design. The authors find a decline in mobility with a one-day decrease in the number of days being outside and an increase in the probability of never going out by 30 percentage points. These point estimates are statistically significant at the 1% level. The mobility restrictions lead to a worsening in mental health outcomes of approximately 0.2 standard deviations, statistically significant at the 10% level in their preferred specification. In this paper we accomplish two things. First, we successfully reproduce Altindag et al.'s main findings. Second, we test the ro-bustness of the results to a small number of changes to their preferred estimations by (1) not clustering the standard errors on the running variable, (2) not including control variables, and (3) calculating the optimal bandwidth using another technique. Point estimates for mobility outcomes are stable to all three manipulations, and standard errors only change marginally. Point estimates and standard errors for the mental health outcomes are somewhat more sensitive, especially to changing the optimal bandwidth selection method. However, the observed changes are reasonably expected when applying data-driven model selection methods to noisy data (to avoid over-fitting, it is likely preferable to apply a less data-driven approach like the original authors did). Our general impression is that the original analyses and results are both theoretically plausible and credible, despite some defensible model dependencies.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/16.htm

Link to Replicators' Package: https://osf.io/25u7b/

Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/17.htm

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/131981/version/V1/view

Reproduction Report

Title Original Study: Mortality, Temperature, and Public Health Provision: Evidence from Mexico

doi: https://doi.org/10.1257/pol.20180594, American Economic Journal: Economic Policy

Report's Abstract: Cohen and Dechezleprêtre (2022) investigate the heterogeneous impact of temperature on mortality across Mexico, and how affordable healthcare services that target the low-income population attenuate the mortality effects of weather events. They find that while extreme temperatures are more dangerous than less extreme temperatures, the increased frequency of non-extreme temperatures mean these temperatures cause more deaths. First, we reproduce the paper's main findings, uncovering a minor coding error that has a trivial effect on the main results. Second, we test the robustness of the results to clustering at the state level, omitting precipitation, and using a different weighting scheme. The original results are robust to all of these changes.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/90.htm

Link to Replicators' Package: https://osf.io/q52e4/

Original Authors' Response: Cohen: "We thank The Institute for Replication. Next time, I will make sure I do not forget Feb. 29th in the code!"

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/125201/version/V1/view

Reproduction Report

Title Original Study: Motivated Beliefs and Anticipation of Uncertainty Resolution

doi: https://doi.org/10.1257/aeri.20200829, American Economic Review: Insights

Report's Abstract: Drobner (2022) examines the effect of manipulating experimental subjects' expectations about uncertainty resolution in learning about their performance on their belief updating patterns in an ego-relevant domain. In their preferred empirical specification, the author finds that individuals update their beliefs optimistically as they exhibit a higher belief adjustment in response to good compared to bad news only when they do not expect resolution of underlying uncertainty about their performance in an IQ test and neutrally when they know they will find out their relative performance at the end of the experiment. First, we reproduce the all of the paper's findings without identifying any coding errors. Second, we test the robustness of the results to (1) adding individual covariates and (2) excluding subjects who exhibit a fundamental error in their belief updating from the analysis. We find no substantial changes in the main coefficients of interest with the inclusion of demographic variables in the analysis, consistent with demonstrated balance in covariates between the two experimental groups. Yet, several of the main estimates lose statistical significance and change from conservatism (under-updating) to over-inference (over-updating) in some conditions on the subset of participants excluding those who exhibit fundamental errors in belief updating.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/65.htm

Link to Replicators' Package: https://osf.io/evt3a/

Original Authors' Response: "Thanks for sharing the report. I think it's a great initiative and feel free to publish this report on your webpage. I will not be able to provide an "answer"."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/139262/version/V1/view

Reproduction Report

Title Original Study: Multinationals' Sales and Profit Shifting in Tax Havens
doi: https://doi.org/10.1257/pol.20200203, American Economic Journal: Economic Policy
Report's Abstract: We perform a robustness replication analysis of Laffitte and Toubal (2022), which considers how multinational corporations shift profit to "tax havens", jurisdictions where they face lower tax burdens. We find that the main results of Laffitte and Toubal (2022), are fairly robust to alternative versions of three important researcher choices: i) the definition of tax havens; ii) the use of a continuous measure of tax-friendliness rather than a binary classification of tax havens; and iii) a sample that omits two small but "extreme" tax havens: Bermuda and Barbados. In all cases, results remain of the same sign and retain statistical significance, though the magnitudes are somewhat attenuated in our robustness exercises.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/37.htm
Link to Replicators' Package: https://osf.io/3sbmr/
Original Authors' Response: "Thanks for your email and for replicating our exercise. Your work is useful. We recognize that the results remain consistent even when considering different interpretations of the haven concept and a smaller sample of observations.

We are also pleased to hear that the replication file we shared with the AEJ: Policy has proven helpful."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/148301/version/V1/view

Reproduction Report

Title Original Study: Multiracial identity and political preferences
doi: https://doi.org/10.1086/714760, Journal of Politics
Report's Abstract: The growing concern regarding reproducibility and replicability of social science re- sults has powered the adoption of open data and code requirements at journals and norms among researchers. However, even when these norms and requirements are fol- lowed, changes to the software used in data cleaning and analysis can render papers non-reproducible. This paper details the challenges of reproducibility in the face of software updates. We present a case study of a published article whose results are no longer reproducible due to changes in the software used. We then discuss the tools and techniques researchers can use to ensure that their research remains reproducible despite changes in the software used.
Link to Full Report: https://osf.io/ecymu/
Link to Replicators' Package: https://github.com/taylorjwright/r_and_p_versioning
Original Authors' Response: Back and forth between authors and replicators, but did not obtain a final response as of November 2025.
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BLVJJH

Reproduction Report

Title Original Study: News Shocks Under Financial Frictions
doi: https://doi.org/10.1257/mac.20170066, American Economic Journal: Macroeconomics
Report's Abstract: Görtz et al. (2022) estimate the effects of innovations to future total factor productivity (TFP) on financial markets. In a Bayesian vector autoregression, they identify a TFP news shock as one that explains the largest share of 40- quarter ahead forecast error variance (FEV) of TFP. Their estimated impulse responses functions show that a positive news shock significantly decreases credit market spreads and increases credit market supply. They also find that a shock that explains the maximum of the FEV of the "excess bond premium" (EBP) (Gilchrist and Zakrajsek 2012) causes similar responses. These results are consistent with an estimated DSGE model with financial frictions. We estimate the main IRFs of the study using the original data and a frequentist estimation approach. We obtain similar point estimates for the dynamic responses to TFP news and EBP max-share shocks. We also update their macroeconomic and financial time series, as some of the data has been revised substantially since their original estimate. We use the updated data to re-estimate the above-mentioned IRFs, and we find that the results are robust to this change in the data. Finally, we investigate the computational reproducibility of their DSGE results, and find that their provided code (consistent with warnings in their README file) does not execute in the most recent version of Dynare or Matlab. Using the version indicated in their replication files, we encounter issues estimating the posterior mode.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/51.htm
Link to Replicators' Package: https://github.com/gionikola/replication-game-ucsd
Original Authors' Final Response: "Thank you for the update and considering our work for replication."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/130141/version/V1/view

Reproduction Report

Title Original Study: Non-Linearities, State-Dependent Prices and the Transmission Mechanism of Monetary Policy
doi: https://doi.org/10.1093/ej/ueab049, Economic Journal
Report's Abstract: Ascari and Haber (2022) fill the gaps in the literature by showing evidence in favor of the state-dependent sticky price model's predictions using the macro-aggregates. We report a replication and robustness check of the study. We employ several additional macroeconomic control variables and different alternative measurements for monetary policy shocks and find that the original results remain qualitatively robust. Our analysis further shows that the turbulent periods of inflation in the 1970s and 1980s have an important role in claiming the robustness of the original results.
Link to Full Report: https://osf.io/kbwap/
Link to Replicators' Package: https://osf.io/kbwap/
Original Authors' Response: No response.
Original Authors' Package: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/ej/132/641/10.1093_ej_ueab049/1/ueab049_replication_package.zip?Expires=1765387359&Signature=Ysj-YfzVPuVPtgGm1ZBFtFNI1APv5x1Rgajkm2orCIsJXt2dZG3Amp92XbuA6m0iP-4LwOFv~PFKA4t1qXm_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA

Reproduction Report

Title Original Study: Not All Elections Are Created Equal: Election Quality and Civil Conflict

doi: https://doi.org/10.1086/714778, Journal of Politics

Report's Abstract: Utilizing a time-series cross-sectional dataset on developing countries, Donno et al. (2022) examine how variation in election quality shapes opportunities and incentives for civil conflict. Across a number of models in their analysis, they find that civil conflict is more likely when elections are not free and fair. They also find that for countries with low integrity elections, the probability of conflict occurring is higher if it has experienced conflict before. We begin by reproducing Donno et al.'s (2022) main models and findings, which yielded no coding errors or differences in effect estimates. Afterwards, for replication purposes we run a series of robustness and conceptual replication tests. For our first replication, we examine the heterogeneous effect between electoral integrity and ethnic fractionalization on conflict. Our second test examines whether a subsample of authoritarian regimes should have been included in the authors' original analysis.

Link to Full Report: https://osf.io/unhkr/

Link to Replicators' Package: https://drive.google.com/drive/folders/1Vlwfr3_Q7c56XFPsQuKUNSnEU_lMFUQz?usp=sharing

Link to Original Authors' Response: https://osf.io/unhkr/

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8B31FG

Reproduction Report

Title Original Study: Parties as Disciplinarians: Charisma and Commitment Problems in Programmatic Campaigning
doi: https://doi.org/10.1111/ajps.12638, American Journal of Political Science
Report's Abstract: Hollyer, Klašnja, and Titiunik (2022) analyse the trade-off that political parties face between running programmatic campaigns and fielding charismatic candidates, whose electoral appeal may come at the cost of undermining the party brand. They argue that higher electoral volatility prompts parties to rely on charismatic candidates, even though they might not be as loyal to the party's programmatic stance. They substantiate their argument with a cross-national dataset and a quantitative case study in Brazil. We computationally reproduced and conducted further robustness tests for their cross-national study by translating the Stata code to R. Next, we conducted a computational reproduction and some additional robustness tests for the quantitative case study. We find that their cross-national analysis is reproducible, albeit with some minor discrepancies. The quantitative case study is also largely reproducible and both are robust in several ways. We conclude by making some suggestions about data dissemination and robustness checks for authors of regression discontinuity designs.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/54.htm
Link to Replicators' Package: https://osf.io/93gfx/?view_only=7063353244d646ffaf7bfd53013e3143
Original Authors' Response: "Thanks for your note and for all the work of Kelly, Odermatt, and Metson in replicating our paper. [...] Our read of the Reproduction Reports that the findings in our paper hold in the Kelly et al replication. [...] Our sense is that the discrepancies between the replication and original paper are sufficiently small, and the task of comparing the replication R code to the original Stata code is likely to be sufficiently demanding of time, that the opportunity cost of a thorough response is high. So, I think we'll forgo the opportunity to draft a response, and just let the replication stand without reply.

We'll leave it to any sufficiently interested parties with expertise in both Stata and R to iron out the discrepancies between the replication and original paper."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AWSQTW

Reproduction Report

Title Original Study: Patience, Risk-Taking, and Human Capital Investment Across Countries

doi: https://doi.org/10.1093/ej/ueab105, Economical Journal

Report's Abstract: Hanushek et al. (2021) test how country-level measures of patience and risk-taking from the Global Preference Survey predict student performance on the Programme for International Student Assessment (PISA) math test. They find that country-level patience positively predicts math test scores and country-level risk-taking negatively predicts math test scores. They find similar results when holding country of residence characteristics constant and focusing on the preferences of the country of origin of migrants. We have checked the computational reproducibility and find that the data and analysis script provided by the authors allowed us to exactly reproduce the main tables in the paper. We also checked the robustness replicability by testing how robust the results are to decisions about imputation, weighting, operationalization of dependent variables, choice of control variables, and the inclusion of highly leveraged observations. We see that results are generally robust, though statistical significance of the risk-taking coefficient in the migrant analysis hinges on whether a control for OECD country of residence is included. Finally, we check the conceptual replicability of the results by using data from the Trends in International Mathematics and Science Study (TIMSS) instead of PISA - a different dataset with a different standardized test. This exercise shows that their results are robust to expanding the analysis to countries participating in both PISA and TIMSS.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/48.htm

Link to Replicators' Package: https://osf.io/kgt8z/

Link to Original Authors' Response: "We would like to thank the replicators and compliment them for their thoughtful replication and extension of our paper. We are particularly impressed by the extension to the TIMSS data, which is actually great support for the underlying idea. We do not see a reason to formulate a formal response for your website.

Thank you all for your valuable work!"

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/153101/version/V2/view

Reproduction Report

Title Original Study: Peer Effects in Academic Research: Senders and Receives
doi: https://doi.org/10.1093/ej/ueac031, Economical Journal
Report's Abstract: In this report, we provide an overview from reproducing and replicating Bosquet et al. (2022). As a first step, we successfully reproduce all the results in the paper, as well as figure A1. All results were fully reproducible and match the published version of the paper. Next, we carry out three sensitivity analysis. We examine how the main results change from the weights used, additional controls, and author-university pairs. The main results are robust to these checks.
Link to Full Report: https://osf.io/czkgw/
Link to Replicators' Package: https://osf.io/czkgw/
Link to Original Authors' Response: The authors responded to the replicators' questions. Bosquet then responded to the final report: "I would simply thank the team of replicators and I am happy to see that the tested results are robust to the tested alternatives. As written in my previous email, I think those kinds of efforts are very useful for the community and the credibility of published results so thanks as well for that."
Original Authors' Package: https://zenodo.org/records/6457037

Reproduction Report

Title Original Study: Playing Politics with Environmental Protection: The Political Economy of Designating Protected Areas
doi: https://doi.org/10.1086/718978, Journal of Politics
Report's Abstract: Mangonnet et al. (2022) examine whether political alignment at the national and sub-national levels explain the spatial designation of Protected Areas (PAs) in Brazil. Their identification relies on spatial discontinuities in political alignment across municipalities. They find that a president-mayor coalition alignment reduces the incidence of PAs by about one percentage point, whereas they find no party alignment effects. We were able to reproduce the paper's findings using the same code and software. Alternative software routines reproduce their results with small and inconsequential numerical differences. Moreover, robustness replications find consistent results for one out the two treatments. Finally, we find no evidence of fabrication of data.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/73.htm
Link to Replicators' Package: https://osf.io/t76jd/
Original Authors' Response: "We are grateful to Laura Villalobos, Jill Caviglia-Harris, Tharaka Jayalath, and the team at the Institute for Replication for generously replicating our work. We encourage readers to follow their alternative software routines for faster estimations."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N6LIMH

Reproduction Report

Title Original Study: Policy Deliberation and Voter Persuasion: Experimental Evidence from an Election in the Philippines
doi: https://doi.org/10.1111/ajps.12566, American Journal of Political Science
Report's Abstract: I would characterize my robustness replication as almost entirely successful. The design checks I report all support a straightforward understanding of the design. My effect and uncertainty estimates barely differ from the original estimates (when compared with like estimation procedures), with any discrepancies attributable to simulation error. One small area of difference was the weighting scheme employed by the authors to correct for "over-representation" of meeting attendees in the treatment group. As discussed below, I do not understand the design reason for this choice and when I simulate its properties, I can obtain small amounts of bias. The net consequence of their approach was usually to make coefficient estimates smaller, so we don't have a major difference in conclusion except perhaps in a secondary analysis of mechanisms.
Link to Full Report: https://osf.io/y8ubt/
Link to Replicators' Package: https://osf.io/y8ubt/
Link to Original Authors' Response: https://osf.io/y8ubt/
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3HACJ

Reproduction Report

Title Original Study: Political Turnover, Bureaucratic Turnover, and the Quality of Public Services

doi: https://doi.org/10.1257/aer.20171867, American Economic Review

Report's Abstract: The politically motivated replacement in local governments is a pervasive fact in our modern democracies. Whether it has causal effects on the quality of public services, such as education, is a critical question and yet understudied. This paper uses a regression discontinuity design (RDD) for close elections to replicate Akthari, Moreira and Trucco (2022) who find negative effects on the quality of public education in Brazil (.05-.08 standard deviations of lower test scores). I first reproduce these main results, finding minor computational differences that have no effect on the conclusions. I also show that the estimates for Brazil are in general robust to different specifications following Brodeur, Cook and Heyes (2020). Finally, I implement the same RDD framework now applied to Chilean administrative records to find null effects on test scores. Taken together, these results suggest that political turnover has weakly negative effects on service quality.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/39.htm

Link to Replicators' Package: https://osf.io/q43vz/

Link to Original Authors' Response: https://osf.io/kv4pj/

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/150323/version/V1/view

Reproduction Report

Title Original Study: Pre-Colonial Warfare and Long-Run Development in India
doi: https://doi.org/10.1093/ej/ueab089, Economic Journal
Report's Abstract: We test the reproducibility and replicability of Dincecco et al. (2022), which reports a positive relationship between pre-colonial interstate warfare and long-run development patterns across India. Overall, we confirm that all of the study's estimates are computationally reproducible by using both the provided replication package in Stata and code written by the present authors in R. We test for and find no evidence of data manipulation in the final datasets. Concerning direct replicability, we consider different ways of measuring distance to conflicts and also alternative proxies for both the dependent variable and variables which capture channels by which the main effects operate. We are able to replicate the magnitude and significance of the estimated coefficient on conflict exposure in most of the tests, noting that while most estimates are substantively in line with the original study, some alternative measures of distance to conflict imply different magnitudes for estimates, and proxy estimates are sensitive to both the time period and type of conflict considered.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/35.htm
Link to Replicators' Package: https://osf.io/af6m2/
Link to Original Authors' Response: https://osf.io/af6m2/
Original Authors' Package: https://zenodo.org/records/5583263

Reproduction Report

Title Original Study: Public Infrastructure and Economic Development: Evidence from Postal Systems

doi: https://doi.org/10.1111/ajps.12594, American Journal of Political Science

Report's Abstract: Rogowski et al. (2022) use secondary data to study the impact of historic postal infrastructure on economic development, both cross-country and within the US. Their results suggest a large positive effect of post offices on economic development that is robust across various sensitivity checks. We successfully computationally reproduce all results. In a robustness assessment, we find the results to be robust to simple changes in the analysis but observe some sensitivity to accounting for spatial trends in the cross-country analysis. Additionally, we correct a coding inconsistency, showing that in the corrected version, one main robustness check for the US-analysis is no longer supporting the result. Despite this, we find the results to be overall robust given the numerous analyses and robustness checks in the original paper.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/92.htm

Link to Replicators' Package: https://osf.io/j3ydr/?view_only=ad14a07cb3a741ca9bbfab391ad7c6fe

Original Authors' Response: "Thanks so much for reproducing the findings in our paper and exploring extensions of our results. We also appreciate your sharing the report with us. [...] I [Rogowski] confirm that we are comfortable letting your report stand and that we will not write a response to it. "

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/33K3EF

Reproduction Report

Title Original Study: Re-Assessing Elite-Public Gaps in Political Behavior
doi: https://doi.org/10.1111/ajps.12583, American Journal of Political Science
Report's Abstract: Kertzer (2022) conducts a meta-analysis of parallel experiments on samples of political elites and ordinary citizens. He examines whether the average treatment effect for elites is significantly different from the average treatment effect for citizens, finding that only 19 of 162 (11.7%) difference-in-difference estimates are statistically significant after adjusting for the false discovery rate. He also finds that elites and masses hold similar foreign policy attitudes after controlling for their demographic characteristics. In this Reproduction Report, we begin by running robustness and heterogeneity tests for the first claim. We find that the results survive many robustness tests. We also find, however, that only a small number of the these treatments significantly affected masses (N=28) or elites (N=30). This low rate suggests the possibility that almost all of these experiments failed to successfully manipulate either masses or elites. If so, we may not be able to conclude that masses and elites respond similarly to experiments with confidence until political scientists produce more experiments with actual treatment effects or with successful manipulation checks in cases of null effects. In the second part of this Reproduction Report, we conceptually replicate the second Kertzer analysis, finding a strong correlation between elite and mass political decisions and attitudes, thus confirming Kertzer's analysis.
Link to Full Report: https://www.econstor.eu/bitstream/10419/266385/1/I4R-DP010.pdf
Link to Replicators' Package: https://osf.io/93urk/
Original Authors' Response: "Thank you for your email and for the invitation. [...] please send my appreciation to the authors for their interest in the manuscript; I find their analysis very interesting."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LHOTOK

Reproduction Report

Title Original Study: Rebel on the Canal: Disrupted Trade Access and Social Conflict in China, 1650–1911

doi: https://doi.org/10.1257/aer.20201283, American Economic Review

Report's Abstract: Cao and Chen (2022a) study the role of disruption of trade on social conflict in China in the 19th century. Identification builds on the closure of China's Grand Canal in 1826 in a difference-in-differences framework. In their preferred analytical specification, the authors find that counties along the canal experienced a 117 percent increase in rebelliousness after the initial closure of the canal in 1826 relative to their non-canal counterparts. First, we reproduce the paper's main findings using the official replication package. Second, we examine whether a sub-sample of counties/prefectures/provinces drives the result. Third, we test the robustness of the results to alternative treatment periods.

Link to Full Report: https://osf.io/dhn6e/

Link to Replicators' Package: https://osf.io/dhn6e/

Link to Original Authors' Response: https://osf.io/dhn6e/

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/157781/version/V1/view

Reproduction Report

Title Original Study: Recessions, Mortality, and Migration Bias: Evidence from the Lancashire Cotton Famine

doi: https://doi.org/10.1257/app.20190131, American Economic Journal: Applied Economics

Report's Abstract: Vellore Arthi, Brian Beach and W. Walker Hanlon (2022) investigate the effect of the Lancashire Cotton Famine on mortality, accounting for the migration response to the downturn. They use difference-in-differences to estimate the effect of the cotton famine on mortality. To account for the migration response to the cotton famine, they construct a linked dataset giving mortality rates by district of residence during the cotton famine, rather than by district of residence at the time of death. They find that the cotton famine increased mortality in cotton-textile producing districts, and that accounting for migration matters, in the sense that their estimates would have been markedly different had they not accounted for it. I check that ABH results are fully reproducible using their data and code, and that their claims are robust to (1) decreasing the age window for building the linked dataset, (2) modifying the specification and (3) computing different standard errors. The only significant discrepancy in results is that I find stronger effects of the cotton famine when I decrease the age window for building the linked dataset, likely because this reduces measurement errors.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/25.htm

Link to Replicators' Package: https://www.openicpsr.org/openicpsr/project/192272/version/V1/view

Original Authors' Response: "Thanks for the interest in our work. We've had a chance to review the report and it looks like everything replicated. Since there are no outstanding queries, we are happy to sign off on this."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/128521/version/V1/view

Reproduction Report

Title Original Study: Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India

doi: https://doi.org/10.1257/aer.20201112, American Economic Review

Report's Abstract: Dhar et al. (2022) examine the effect of a gender attitude change program in secondary schools in India. In their preferred specification, the authors show that the program made the students report more gender-egalitarian attitudes by 0.18 of a standard deviation, and shifted self-reported behaviors to be more aligned with gender-progressive norms by 0.20 standard deviations (both significant at 1% level). In contrast, they found no effect on girls' aspirations, as these were already high before the intervention. The effects did not attenuate between the first end-line (right after the programme was completed) and the second (two years later). To put the paper's results in perspective, we first comment on the authors' deviations from their pre-registration and pre-analysis plans, provide detailed power calculations, and add multiple-hypothesis-testing-adjusted standard errors. Second, we show that the paper's results are perfectly reproducible. Third, we show that the results are robust to excluding control variables, and alternative ways of constructing indices and dealing with non-response.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/24.htm

Link to Replicators' Package: https://osf.io/r5jfe/

Final Original Authors' Response: "Thanks. the revision looks good. I actually don't think we need to have a formal response any more. [...] Thus, I don't think there is anything substantive for us to include in a discussion paper/response. That reflects the fact that the Reproduction Reports fair and there is nothing major to respond to, so it's good news, from both the perspective of the integrity of our original paper and the professionalism of the replication."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/149882/version/V1/view

Reproduction Report

Title Original Study: Run-off Elections in the Laboratory
doi: https://doi.org/10.1093/ej/ueab051, Economic Journal
Report's Abstract: Bouton et al. (2022) make a causal claim by manipulating the voting system under which participants vote (runoff or plurality) and examining whether this manipulation affects the proportion of strategic voting. They estimate the effect of the voting system on the proportion of strategic voting for the participant population, using random effect regression where standard errors are clustered on group level. Regarding replication results, we reproduced the original study's main findings. Our analysis confirms that there are minor and mostly non-significant disparities in electoral outcomes and voters' welfare between the two voting systems, consistent with the original study's conclusions. Specifically, we conducted tests to assess the study's computational reproducibility and direct replicability. While the authors provided the raw data, they did not include a script for cleaning it or a codebook describing its contents. Consequently, we developed a data cleaning script to ensure accurate and consistent data processing.
Link to Full Report: https://osf.io/a8cev/
Link to Replicators' Package: https://github.com/carinahausladen/runoff-elections
Original Authors' Response: The authors provided feedback which was taken into account.
Original Authors' Package: https://doi.org/10.1093/ej/ueab051

Reproduction Report

Title Original Study: School Spending and Student Outcomes: Evidence from Revenue Limit Elections in Wisconsin
doi: https://doi.org/10.1257/pol.20200226, American Economic Journal: Economic Policy
Report's Abstract: Baron (2022) explores the independent effects of operational expenditure and capital expenditure on student outcomes in school districts across Wisconsin from the outcomes of close referendum approvals. By utilizing a dynamic regression discontinuity framework and cubic specification, the author finds that narrowly passing an operational referendum, increases operational expenditure per pupil by $298 each year on average, following the referendum over a ten year period. From this $198 are spent on instructional expenses. These point estimates are statistically significant at the 10% and 5% level, respectively. We first reproduce the main results from the paper without any issues arising. Secondly, we conduct a robustness replicability to (1) dropping school districts from the top and bottom 5% of the revenue limits distribution, categorically, and (2) dividing the time frame of the study into two periods: 1996-2005 and 2005-2014. We find that dropping the top 5% of the school districts by revenue limits reduces the additional operational expenditure by $140 per pupil (lower by 50 percent) and the effects of passing an operational referendum were nearly double in the former period compared to the latter period. Lastly, we find that the estimated effects on student outcomes rely heavily on recent observations.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/88.htm
Link to Replicators' Package: https://osf.io/m2w4x/
Original Author's Response: "Thank you for sharing the Reproduction Report. Please pass on my thanks to the replicators for their important work. First and foremost, I'm glad to see that the results in the paper are reproducible without any issues arising. The report explores two additional sources of heterogeneity. I have no additional comments on these. I do briefly want to clarify the last sentence in the report's abstract, which reads "Lastly, we find that the estimated effects on student outcomes rely heavily on recent observations." While I am not entirely sure what the replicators are referring to, my guess is that they refer to Table 2 in the report. In this table, they discuss that they are unable to study heterogeneity in the impacts of passing a referendum on test scores and postsecondary enrollment from 1996-2005, because data on these outcomes are unavailable prior to 2005. The availability of each dataset was discussed in the published version of the paper (see, for example, Table 1). Perhaps a more accurate statement would be to explain that the replicators couldn't explore the impact of passing a referendum on these specific outcomes in the early period due to data constraints—and that this was acknowledged in the published version."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/125821/version/V1/view

Reproduction Report

Title Original Study: Social Class and (Un)Ethical Behaviour: Causal and Correlational Evidence
doi: https://doi.org/10.1093/ej/ueac022, Economic Journal
Report's Abstract: The relationship between social status and ethical behavior is a widely debated topic in research. In their study, Gsottbauer et al. (2022b) investigate whether higher socio-economic status is linked to lower ethical behavior, using data from two large survey experiments involving over 11,000 participants. In this replication project, we test the computational reproducibility and robustness to the replication of their study, using the provided data and code from the replication package (Gsottbauer et al., 2022a). Nearly all the figures and tables were reproducible-in the process of reproducing the results, some minor rounding or transcription errors were discovered. In testing the robustness replicability, we find consistent results for our extensions. The effort for the replication was manageable, even though the authors treat categorical variables as numeric, or use manually-coded interaction variables (i.e., in regression models). In summary, we applaud the transparency of Gsottbauer et al. (2022b) in facilitating replications, and make some general recommendations for further improvements for data-analysis studies.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/29.htm
Link to Replicators' Package: https://github.com/ha0ye/replication-gsottbauer-2022
Original Authors' Response: Declined to respond.
Original Authors' Package: https://zenodo.org/records/6226207

Reproduction Report

Title Original Study: Sorting or Steering: The Effects of Housing Discrimination on Neighborhood Choice
doi: https://doi.org/10.1086/720140, Journal of Political Economy
Report's Abstract: This comment revisits the analysis in Christensen and Timmins (2022). We identify two critical errors used in the original analysis, one with the data and the other with coding. When either error is corrected several major results in the paper change, either in statistical significance or in effect size. The data error is a result of including fixed effects for the string variable 'city'. The raw variable is case sensitive and has many spelling mistakes. The coding error involves assigning a value of zero for the variable "of color" to both individuals identified as 'white' and as 'other' in the raw data. The level of clustering in the paper is also arguably too fine. Many of the results are not robust to clustering at the city level, as opposed to the subject pair level. In total, we affirm the authors' overarching claim of substantial and nuanced housing discrimination against racial minorities generally, and African Americans in particular; however, the effect sizes and significance are generally (although not always) smaller than the original authors findings. Additionally, there are several instances where the effects of discrimination on African Americans are no longer statistically significant but the effect of discrimination on Hispanics becomes significant.
Link to Full Report: https://osf.io/vwgxd/
Link to Replicators' Package: https://github.com/mattdwebb/HUDreplication
Original Authors' Response: Authors mentioned that they are currently writing a response.
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/720140/suppl_file/20191181data.zip

Reproduction Report

Title Original Study: Spillover Effects of Intellectual Property Protection in the Interwar Aircraft Industry
doi: https://doi.org/10.1093/ej/ueab091, Economic Journal
Report's Abstract: We are attempting to reproduce the results of Hanlon and Jaworski (2022) based on their dataset. Our work is conducted in two different ways: (i) computational reproducibility, aiming to produce the same results using different software, notably R, with the given data; and (ii) checking the robustness of the results. For (i), the estimated coefficients are consistent based on the R software. For (ii), we carefully examine the given datasets of Hanlon and Jaworski (2022) and review the economic history of the US Interwar aircraft industry between 1918 and 1935 to identify potential confounding variables (apart from IPP strengthening in the year 1926) that might affect both the controls and error term, and thus the results. We identify some confounding variables that may affect the results and attempt to illustrate them before and after 1926 when IPP is strengthened. Overall, we find that the results are replicable by utilizing the codes and datasets of Hanlon and Jaworski (2022), which is encouraging.
Link to Full Report: https://osf.io/t4avf/
Link to Replicators' Package: https://osf.io/t4avf/
Link to Original Authors' Response: https://osf.io/t4avf/
Original Authors' Package: https://zenodo.org/records/5627298

Reproduction Report

Title Original Study: State Action to Prevent Violence against Women: The Effect of Women's Police Stations on Men's Attitudes toward Gender-Based Violence

doi: https://doi.org/10.1086/714931, Journal of Politics

Report's Abstract: Córdova and Kras (2022) examine how the existence of a women's police station (WPS) in the place of residence influences citizens' attitudes toward gender-based violence in Brazil. In their analytical specification, the authors find that men are more likely to reject violence against women (VAW) and support bystander intervention in municipalities with a WPS, especially if the WPS has been operating for a long time. This paper examines the replicability and robustness of Córdova & Kras' (2022) findings. First, we reproduce the paper's main findings and uncover one minor coding error and three estimates that have been reported with the opposite sign compared to that in our reproduction; neither is of consequence for the study's main results. Second, we test the robustness of the results by (1) recoding one of the main explanatory variables and several of the control variables to account for non-linear trends, (2) using alternative techniques to estimate clustered standard errors, (3) consistently applying a 95% confidence level in the presentation of the results, (4) altering the propensity score matching (PSM) procedure as well as the composition of the variables used in the PSM robustness check, (5) using an alternative technique to test for multicollinearity, (6) excluding potential endogenous control variables, and (7) using an alternative coding for computing margins. Reassuringly, the results are robust to most of these tests. However, two of the robustness checks challenge parts of the paper's main findings. First, allowing for non-linearity in the effect of time since the establishment of WPS shows (a) a non-linear effect on VAW and (b) no apparent changes in either male or female attitudes over time once the WPS has been established. Second, the inclusion of other variables in the PSM procedure renders part of the main estimates of interest statistically nonsignificant ($p < 0.1$). Our findings highlight the need for further re-analyses and replications for investigating the preventive effects of women's police stations.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/67.htm

Link to Replicators' Package: https://osf.io/yjwr8/

Link to Original Authors' Response: Responded to our emails but no formal response as of February 2024.

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/D2WL5I

Reproduction Report

Title Original Study: Student Performance, Peer Effects, and Friend Networks: Evidence from a Randomized Peer Intervention

doi: https://doi.org/10.1257/pol.20200563, American Economic Journal: Economic Policy

Report's Abstract: Wu et al. (2023) estimate the effect of classroom seating arrangements in China using a randomized control trial with two treatment schemes. The first treatment scheme involves seating high and low achieving students together, and the second treatment involves this same seating arrangement with financial incentives for the high-achieving students, if their deskmates' test scores improved. All statistically significant impacts come from the incentivized treatment scheme. Wu et al. (2023) find that low-achieving students sitting next to incentivized high-achieving students perform 0.24 SD (p-value=0.018) better on math exams. In addition, being assigned to the incentive treatment scheme increased extraversion and agreeableness for low and high achieving students. Lastly, they do not find much evidence of peer effects on test scores nor personality traits. This study is computationally reproducible using their provided replication package. We ran their code using Stata 14, 17, and 18. After running their replication package, we further investigated Tables 2-5. The main conclusions are generally robust to various coding decisions. Notably, in investigating the peer effects, when we change the specification to also control for the difference in baseline scores between the student and their deskmate, we find that the more dissimilar deskmates are at baseline, the bigger the peer effects.

Link to Full Report: https://osf.io/9hx3b/

Original Authors' Response: The authors provided feedback which was taken into account.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/149262/version/V2/view

Reproduction Report

Title Original Study: Talking Shops: The Effects of Caucus Discussion on Policy Coalitions
doi: https://doi.org/10.1111/ajps.12636, American Journal of Political Science
Report's Abstract: In Talking Shops: The Effects of Caucus Discussion on Policy Coalitions, Zelizer analyzes the causal effect of caucus deliberations on legislative policy coalitions. In practice, political scientists have little empirical evidence on how policy discussions actually work among sitting legislators and whether these discussions have an effect on policy making and policy opinion. Taking on this challenge, Zelizer conducted two field experiments in an American state legislature. In short, the experiments randomized whether a bill was selected for discussion among a bi-partisan legislative caucus. The paper then measures and reports the corresponding effects of that discussion around the bill. Zelizer finds that deliberation increased the amount of co-sponsorship for a given bill, among both co-partisans and counter-partisans, but deliberation did not effect whether a bill was passed by the legislature or whether the bill received more amendments. We conduct a robustness replication of the main results of Talking Shops. Specifically, we reproduce Tables 3 and 4 of the paper under alternative specifications. We find that the main results of the paper are reproducible and robust to multiple alternative specifications.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/69.htm
Link to Replicators' Package: https://osf.io/tmfyj/
Link to Original Authors' Response: "One purpose of replication, among others, is to evaluate whether published results are sensitive to modeling decisions. Do alternative, reasonable approaches generate the same, or different, results? Did the author's approach provide an outlier estimate that is indicative of p-hacking or, to be kinder about it, sensitivity of results to modeling decisions? That seems incredibly useful. That purpose is not advanced, in my view, by testing 'incorrect' methods or models. We do not learn about the robustness of results from testing alternative approaches that introduce bias, or by estimating different estimands that are a combination of treatment effects and selection bias. While it doesn't seem to matter too much in this case — selection bias appears relatively small, and in the same direction as treatment effects — I think this issue matters for the exercise in general for several reasons. First, do the analyses justify the inferences being made? In my view, changing the estimand or estimating biased models cannot justify saying anything about the robustness of the original results. Second, what would have happened if the new results did not match the original? Are we willing to claim published results are not robust when applying estimators with known flaws generates different results? And third, shouldn't we just generally aim to use 'correct' estimators for a given situation? While IPW is not perfect, ignoring differential treatment probabilities is a conscious decision to ignore selection bias. Why would we want to run that model if our goal is inference about treatment effects? I appreciate the work everyone is doing on this enterprise. Hopefully these comments, whether correct or not, help advance the goal of publishing robust, valid empirical research."

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3M5AX

Reproduction Report

Title Original Study: Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field

doi: https://doi.org/10.1257/aer.20200751, American Economic Review

Report's Abstract: Hussam et al. (2022a) use a cash grant experiment in India to demonstrate that community knowledge can help target high-growth microentrepreneurs. In their preferred specification, the authors find that the average marginal return to the grant is 9.4 percent per month, while estimated returns for entrepreneurs reported by peers to be in the top third of the community are between 24 percent and 30 percent. First, we reproduce the paper's main findings and uncover one minor coding error, which affects the estimates for one of the main tables but does not change the overall conclusions of the paper. Second, we test the robustness of the results to: (1) different treatment of outliers, (2) dropping surveyor and survey month fixed effects, and (3) using quartiles instead of terciles for grouping the ranking of entrepreneurs. The paper's results are robust to these robustness checks. Finally, we test heterogeneity of results by gender, which was not reported in the original study.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/49.htm

Link to Replicators' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DI7RR9

Link to Original Authors' Response: "We are very grateful to Isabella Masetto, Diego Ubfal, and to the team at I4R for their excellent work. We verified the coding error and we agree that it did not meaningfully alter the conclusion of our paper that community information is informative over and above the predictive power of observable characteristics. We will post a link to this correction on our websites and will consult the editors of the AER as to whether this error rises to the level of requiring a formal correction."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/151841/version/V1/view

Reproduction Report

Title Original Study: Teaching Norms: Direct Evidence of Parental Transmission
doi: https://doi.org/10.1093/ej/ueac074, Economic Journal
Report's Abstract: This paper is a replication study of Brouwer, T., Galeotti, F., & Villeval, M. C. (2023), using the original data. The study explores how social norms are transmitted from one generation to another, specifically from parents to children. The authors conducted a field experiment involving 601 parents of children aged 3 to 12 in Lyon, France, to examine whether parents engage more in norm enforcement in the presence of their child, and whether the nature of punishment changes in the presence of the child. The study found that parents do engage more in norm enforcement in the presence of their child, and tend to use more indirect punishment when their child is present. This study highlights the role that parents play in transmitting social norms to their children. The replication analysis was successful, with the results of the original study being robust to changes in the model specification.
Link to Full Report: https://osf.io/qnbfa/
Link to Replicators' Package: https://zenodo.org/records/8114738
Original Authors' Response: The replicators took into account the authors' feedback. They wrote at the end of the back and forth: "We thank you and the replication team for the replication and the successive interactions. We created an OSF project including the data replication package enabling the reproduction of the analysis presented in our article. The package comprises a source file (in Stata format and in TXT) and a Stata do-file that allows the reconstruction of the master file used in the replication package submitted to the Economic Journal."
Original Authors' Package: https://zenodo.org/records/7045559

Reproduction Report

Title Original Study: Technological Change and the Consequences of Job Loss
doi: https://doi.org/110.1257/aer.20210182, American Economic Review
Report's Abstract: Braxton and Taska (2023) find that technological change accounts for 45 percent of the decline in earnings after job loss. We first reproduce all regression tables in Braxton and Taska (2023), and then test for robustness by controlling for the initial level of wages, additional fixed effects, multi-way clustering, and conducting influential analysis. We find that the paper's original results are sensitive to controlling for initial wages and some additional fixed effects. Overall, we find the results are robust with a coefficient in the same direction and significant at 5% in 33% of the robustness checks we ran, with average t/z scores 28% as large as the original study.
Link to Full Report: https://osf.io/qws2p/
Link to Replicators' Package: https://osf.io/qws2p/
Original Authors' Response: Did not get a response as of November 2025.
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/181166/version/V1/view

Reproduction Report

Title Original Study: The Common-Probability Auction Puzzle
doi: https://doi.org/10.1257/aer.20191927, American Economic Review
Report's Abstract: Ngangoué and Schotter (2023) investigate common-probability auctions. By running an experiment, they find that, in contrast to the substantial overbidding found in common-value auctions, bidding in strategically equivalent common-probability auctions is consistent with the Nash equilibrium. We reproduce their results in R, conduct robustness checks on how their sample was constructed, and consider possible heterogeneity. We confirm their documented qualitative results.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/74.htm
Link to Replicators' Package: https://osf.io/7bq4s/
Original Authors' Response: "Thank you for putting the effort in replicating our study! Your results are also quite interesting to us as we haven't thought of all the robustness checks you've made. At this point, we do not have any major comments to make and refrain from submitting a response."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/184041/version/V1/view

Reproduction Report

Title Original Study: The Curious Case of Theresa May and the Public That Did Not Rally: Gendered Reactions to Terrorist Attacks Can Cause Slumps Not Bumps

doi: https://doi.org/10.1017/S0003055421000861, American Political Science Review

Report's Abstract: Holman et al. (2022; HMZ) propose women (compared to men) political leaders experience significant drops in public approval ratings after a transnational terrorist attack. After documenting how survey-based evaluations of then-Prime Minister Theresa May suffered after the 2017 Manchester Arena attack, HMZ assemble a country-quarter level panel database to explore the generality of their hypothesis. They report evidence suggesting women (compared to men) leaders systematically experience decreased public approval rates after major transnational terrorist attacks (p-value of 0.020). We find that result disappears once any of the following adjustments is implemented: (i) excluding election quarter covariates ($p = 0.104$); (ii) correcting objective coding errors in the election quarter covariates ($p = 0.058$); (iii) excluding the May-Manchester observation ($p = 0.098$); or (iv) clustering standard errors at the country level ($p = 0.558$). Exploring all 2^5 combinations of the five control groups HMZ incorporate in their specification, none of them clears the 5% threshold of statistical significance once the corrected election quarter variables are employed. We conclude that the empirical evidence does not provide sufficient support for HMZ's abstract claim that "conventional theory on rally events requires revision: women leaders cannot count on rallies following major terrorist attacks."

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/41.htm

Link to Replicators' Package: https://doi.org/10.5683/SP3/6SYCML

Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/44.htm

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VHNPUO

Reproduction Report

Title Original Study: The Dynamics and Spillovers of Management Interventions: Evidence from the Training within Industry Program
doi: https://doi.org/10.1086/719277, Journal of Political Economy
Report's Abstract: Bianchi and Giorcelli (2022) study the long-term and spillover effects of a management intervention program on firm performance in the US, between 1940 and 1945. The authors find that the Training Within Industry (TWI) program led to positive effects which lasted for at least 10 years. Firm sales of treated firms increased by 5.3% in the first year after implementation, peaking at 21.7% after 8 years, before reducing to 16% gains after a decade. The authors claim that the program generated long-lasting changes in managerial practices. Finally, the program also led to positive spillover effects on the supply chain of treated firms. First, we reproduce the paper's main findings. Second, we test the robustness of the results to (1) changing the main specification sample and (2) testing other difference-in-differences estimators, using the same data, provided by the authors. We find that the results are robust to these changes. All point estimates in the study remain statistically significant and of similar magnitude. While the paper's finding reproduce and replicate, challenges in reproducing results we encountered lead us to recommend improvements to journals' code policies.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/66.htm
Link to Replicators' Package: https://github.com/cwestheide/i4r_dp66_code
Original Authors' Final Response: "Thanks a lot for sharing the updated report with us. We don't have anything to add at this point."
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/719277/suppl_file/20200781data.zip

Reproduction Report

Title Original Study: The Economic Effects of Long-Term Climate Change: Evidence from the Little Ice Age
doi: https://doi.org/10.1086/720393, Journal of Political Economy
Report's Abstract: Waldinger (2022) finds significant negative economic effects (proxied by city size) from gradual climate change which occurred during the Little Ice Age (1600-1850) and offers two potential mechanisms (agricultural productivity and mortality) and two potential adaptations (trade and land use). In this comment, we show that while Waldinger (2022)'s findings can be replicated, the main result relies on a critical author assumption: Cities with 0 inhabitants in the original data are instead assumed to have 500. This assumption affects 23.5% of observations and 49.6% of cities in the sample. When these "missing data" are excluded from the analysis, the effect estimated by otherwise identical research methods is of similar magnitude and statistical significance but of opposite sign.
Link to Full Report: https://osf.io/tmn2j/
Link to Replicators' Package: https://osf.io/tmn2j/
Link to Original Authors' Response: https://osf.io/tmn2j/
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/720393/suppl_file/2015548data.zip

Reproduction Report

Title Original Study: The Effects of Banking Competition on Growth and Financial Stability: Evidence from the National Banking Era
doi: https://doi.org/10.1086/717453, Journal of Political Economy
Report's Abstract: Carlson et al. (2022) examine the causal impact of banking competition by investigating a unique circumstance in the National Banking Era of the nineteenth century in the US, where a discontinuity in bank capital requirements occurred. On the one hand, their findings suggest that banks operating in markets with fewer barriers to entry tend to increase their lending activities, promoting real economic growth. On the other hand, banks in less restricted markets also exhibit a higher propensity for risk-taking, posing risks to financial stability. First, we fully reproduce the paper's outcomes apart from a minor discrepancy in the estimate of Table 9 attributed to issues in the provided codes. Second, we test the robustness of the results by (i) changing the ranges used to select the sample of cities included in the analysis, (ii) adopting different options to address outliers' potential issues and (iii) introducing additional control variables. We observe that the estimation results remain mostly consistent when subjecting them to various robustness checks. However, it is worth highlighting that the results can be partially influenced by the criteria used to select the sample of cities and the inclusion of control variables.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/81.htm
Link to Replicators' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BB864R
Original Authors' Final Response: "We thank the replication team (Andrea Calef, Sya In Chzhen, Marco Mandas, and Fabio Motoki) for the detailed Reproduction Report. We are glad to hear that the replicating team affirms the robustness of the paper's findings. We are also glad that the replicators were able to successfully replicate all tables and figures. We thank the replicating team for identifying various smaller issues regarding the code structure which fortunately did not affect our original findings. We believe that the report as such does not require us to respond in any further detail. We highly appreciate the effort of both the replicating team and the I4R."
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/717453/suppl_file/20200610data.zip

Reproduction Report

Title Original Study: The Geography of Repression and Opposition to Autocracy
doi: https://doi.org/10.1111/ajps.12614, American Journal of Political Science
Report's Abstract: Analytic data sets and analysis code are available and they produce the same results as presented in the paper (CRA). Robustness checks involve the (i) use of matching estimators to address possible bias from misspecification, based on propensity score estimated from a random forest model, (ii) doubly robust (TMLE) estimation to address possible bias from misspecification in either the propensity score or outcome regression stages, using a super learner ensemble with random forest, GAM, mean, and non-parametric regression models and averaged over repeated runs to minimize randomness, (iii) define treated comunas as those within a fixed physical distance radius of the nearest military base, rather than only those that contain it, and (iv) instead of using 2SLS to assess the causally mediated effect of military bases on plebiscite outcomes via repression, we propose to conduct mediation analysis (Tingley et al 2013), implemented in the R 'mediation' package.
Link to Full Report: https://www.socialsciencereproduction.org/reproductions/789/published/index?step=4
Link to Replicators' Package: https://github.com/pjesscarter/repression-replication
Link to Original Authors' Response: We are happy that the replicators successfully reproduced all the analysis in our published paper. Moreover, additional robustness checks within the quantitative framework of the paper further confirm the empirical results. Two extensions using propensity score matching give somewhat different results. Unfortunately, these additional estimators violate standard requirements for credible matching designs, i.e., overlap in the propensity score distribution across treatment and control groups. As shown by previous research, this lack of overlap leads to unstable estimators with variance that may explode in finite samples such as ours (Frölich 2004, Khan and Tamer 2010). In another extension, the replicators employ a mediation analysis to re-interpret the empirical evidence in our paper. To support the use of our method, i.e., instrumental variables, we rule out alternative explanations and provide a range of historical evidence. Without historical and contextual support for alternative assumptions, we believe that the method used by the replicators is hard to interpret.
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EYAWES

Reproduction Report

Title Original Study: The Labor Market Impacts of Universal and Permanent Cash Transfers: Evidence from the Alaska Permanent Fund
doi: https://doi.org/10.1257/pol.20190299, American Economic Journal: Economic Policy
Report's Abstract: Jones and Marinescu (2022) study the employment effects of a universal cash transfer in Alaska. Using a synthetic control method, they find that the transfer had no negative effects on employment. We reproduce the results using their replication package and investigate if the results hold when using a different software to run the analysis. We also use different estimation techniques and perform sensitivity checks to assess robustness of the results. We find some differences in the size and significance of the average treatment effects on labor force participation and hours worked when we use a different software (R) and various extensions of the synthetic control method. We also find smaller coefficients on part-time employment when including more covariates. However, these differences do not contradict the main conclusion of the paper.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/80.htm
Link to Replicators' Package: https://osf.io/6atfw/
Original Authors' Final Response: "Thanks for putting in all this effort to evaluate the robustness of our results! I [Marinescu] think this is really a worthwhile endeavor."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/140121/version/V1/view

Reproduction Report

Title Original Study: The Long-Run Effects of Sports Club Vouchers for Primary School Children

doi: https://doi.org/10.1257/pol.20200431, American Economic Journal: Economic Policy

Report's Abstract: Marcus, Siedler and Ziebarth (2022 American Economic Journal: Economic Policy) examine the long-run health effects of a universal sports-club voucher program that was introduced in Saxony for primary school children in 2009. In 2018, the authors designed a survey that targeted the affected cohorts and nearby cohorts in Saxony and two neighboring states, and use a differences-in-differences identification strategy that exploits variation across states and cohorts in policy exposure. The authors document that treated individuals have knowledge of the program and recall receiving and redeeming the vouchers at higher rates, but find no effects on any health outcomes or behaviors. We successfully reproduce the main results of the paper exactly using data available in the paper's replication package and new Stata and R code. We also verify the robustness of the results using different outcomes, different control variables, different sample restrictions and different inference methods.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/46.htm

Link to Replicators' Package: https://osf.io/4bnjt/

Original Authors' Response: "We would like to thank the authors for their interest in our paper. We greatly appreciate their careful reading of the paper and the insightful robustness exercises they conducted. We are pleased that our results were successfully reproduced using different software packages, and that the additional robustness analyses performed by the authors further strengthen and support our conclusions."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/138922/version/V1/view

Reproduction Report

Title Original Study: The Long-Term Effects of Measles Vaccination on Earnings and Employment

doi: https://doi.org/10.1257/pol.20190509, American Economic Journal: Economic Policy

Report's Abstract: Atwood (2022) analyzes the effects of the 1963 U.S. measles vaccination on longrun labor market outcomes, using a generalized difference-in-differences approach. We reproduce the results of this paper and perform a battery of robustness checks. Overall, we confirm that the measles vaccination had positive labor market effects. While the negative effect on the likelihood of living in poverty and the positive effect on the probability of being employed are very robust across the different specifications, the headline estimate-the effect on earnings-is more sensitive to the exclusion of certain regions and survey years.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/33.htm

Link to Replicators' Package: https://osf.io/jv7kx/

Link to Original Authors' Response: https://osf.io/qxjnk/

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/138401/version/V1/view

Reproduction Report

Title Original Study: The Macroeconomics of Sticky Prices with Generalized Hazard Functions

doi: https://doi.org/10.1093/qje/qjab042, Quarterly Journal of Economics

Report's Abstract: We replicate the empirical results in Section 4 of Alvarez et al. (2022). First, we were able to reproduce the original authors' major empirical results, but only after editing the program for it to run on our computing platform. There are small discrepancies in the empirical estimates, e.g. bootstrapped standard errors, that involve the use of simulations. Second, we replicated Alvarez et al.'s results by adopting the data cleaning criteria used by their original data source (Cavallo 2018) to evaluate its robustness to data handling procedures. We found noticeable changes in the empirical results that can have important implications on the effects of monetary policy. To conclude, we propose using Docker container to promote research reproducibility, and more attention is needed on data handling to improve the robustness of empirical research.

Link to Full Report: https://github.com/atyho/Ottawa-Replication-Games-2023/blob/main/Ho_Huynh_Rea_Replication_Report.pdf

Link to Replicators' Package: https://github.com/atyho/Ottawa-Replication-Games-2023/

Link to Original Authors' Response:

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IBM0IE

Reproduction Report

Title Original Study: The Morning After: Cabinet Instability and the Purging of Ministers after Failed Coup Attempts in Autocracies
doi: https://doi.org/10.1086/716952, Journal of Politics
Report's Abstract: We replicate the analysis provided in Bokobza et al. (2022). They identify a causal effect of failed coup attempts on cabinet minister removals in autocracies on both the country and individual minister level and show that higher-ranking ministers and those holding strategic positions are more likely to be purged than more loyal and veteran ministers using fixed effects panel models. We focus on computational reproducibility and robustness replicability. In addition to reproducing the original results using Stata and R, we replicate analyses using random effects panel models and ordered beta regression models, reproduced analyses performed in R using different packages, replaced the main independent variable, clustered standard errors on a different level, and added independent variables related to coup-proofing. We find that the original findings were reproducible and robust.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/45.htm
Link to Replicators' Package: https://doi.org/10.7910/DVN/21HZCC
Link to Original Authors' Response: https://osf.io/sm526/
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GCDJ25

Reproduction Report

Title Original Study: The Origin of the State: Land Productivity or Appropriability?
doi: https://doi.org/10.1086/718372, Journal of Political Economy
Report's Abstract: This is a replication of Mayshar et al. (2022) (MPP).The article posits that the state (defined as societal hierarchy such as tax-levying elites) originated from cultivation of appropriable cereal grains, contrary to the conventional theory that the state originated from increased land productivity following the adoption of agriculture. The article uses multiple datasets to demonstrate a causal effect of cereal cultivation on hierarchy (Claim 1) without finding a similar effect for land productivity (Claim 2), and that societies based on roots or tubers display levels of hierarchy similar to nonfarming societies (Claim 3). The results of our replication in brief are: 1. The data and code provided by MMP closely reproduce the main results presented in their Table 1 (see our Table 1). 2. Concurrently testing the cereal cultivation and land productivity claims leads to slightly less statistical significance, on average, than the published article (Table 2).3. Removing the inherited 1-5 scale of the dependent variable (hierarchy) finds that cereal production is not as effective at moving across all levels of hierarchy compared to the more general claim (Table 3 and 4).4. Using the same procedures with an aim to confirm the conventional hypothesis (land productivity leads to increased hierarchy conditional on cereal growth) offers statistically significant evidence both for and against Claims 1 and 2 and against Claim 3 (Table 6). 5. The statistical significance of Claim 1 is sensitive to the removal of the top 3% of observations outliers (Table 7). 6. Correction of mis-classified 'none or none specified' crop societies alters the interpretation of coefficients behind Claim 3. Societies that rely more on agriculture among farming societies experience more complex hierarchies, irrespective of being primarily cereal producing or tubers growing (Table 8 and 9). (...)
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/82.htm
Link to Replicators' Package: https://osf.io/ekzdg/
Original Authors' Response: Comments taken into account in the report.
Original Authors' Package: https://www.journals.uchicago.edu/doi/suppl/10.1086/718372/suppl_file/2018030data.zip

Reproduction Report

Title Original Study: The Power of Hydroelectric Dams: Historical Evidence from the United States over the Twentieth Century
doi: https://doi.org/10.1093/ej/ueac059, Economic Journal
Report's Abstract: Successful computational reproducibility. No coding errors uncovered.
Original Authors' Package: https://doi.org/10.1093/ej/ueac059

Reproduction Report

Title Original Study: The Relative Efficiency of Skilled Labor across Countries: Measurement and Interpretation
doi: https://doi.org/10.1257/aer.20191852, American Economic Review
Report's Abstract: Rossi (2022) examines the relative efficiency of skilled workers across countries. He finds the elasticity of skill efficiency with respect to GDP per worker is 1.4 and that the relative human capital accounts for only about 9 percent. We reproduce the paper's main findings and test the sensitivity of the results to (1) alternative samples and (2) additional controls for determining wages. We find the results remain robust to these alternative specifications, and the estimated values of the key elasticities remain nearly unchanged.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/59.htm
Link to Replicators' Package: https://osf.io/fge7z/
Original Author's Response: "Thanks for replicating the paper. I don't have any comments to add to the report."
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/146041/version/V1/view

Reproduction Report

Title Original Study: The Side Effects of Immunity: Malaria and African Slavery in the United States

doi: https://doi.org/10.1257/app.20190372, American Economic Journal: Applied Economics

Report's Abstract: Esposito (2022) documents the role of malaria in the diffusion of African slavery in the US. She finds that the introduction of malaria triggered a demand for malaria-resistant labour, which led to a massive expansion of African enslaved workers in more malaria-infested areas. Further results document that, among African slaves, more malaria-resistant individuals commanded significantly higher prices. We reproduce the paper's main findings, uncovering only one minor coding error that has no effect on the study's main results. We then test the robustness of the results to (1) varying the set of control variables used in various analyses; (2) conducting permutation tests; and (3) conducting event studies that account for time-varying controls. We generally find that the author's results are robust to all of these alternative specifications, though there are some sets of controls that cause estimates to become small and statistically insignificant.

Link to Full Report: https://osf.io/728ud/

Link to Replicators' Package: https://osf.io/728ud/

Original Authors' Response: Original author provided feedback. No final response on the updated version.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/120483/version/V1/view

Reproduction Report

Title Original Study: The Wheels of Change: Technology Adoption, Millwrights and the Persistence in Britain'S Industrialisation
doi: https://doi.org/10.1093/ej/ueab102, Economic Journal
Report's Abstract: Mokyr et al. (2022) estimate the effects of early technology adoption on industrialization. Authors argue that human capital was the main determinant of the location of the industry in the first decades of the Industrial Revolution. They document that the location of mills in the eleventh century (reported in the Doomsday Book) has a positive and statistically significant impact on the number of wrights in the early eighteenth century. We confirm the computational reproducibility of the paper. The estimates are not sensitive to outliers, which are common in the data. The results are also robust to changes in the control variables. The results remain robust if we adjust the estimated p-values for the low number of clusters, and if we include county fixed effects. We conduct a placebo experiment with a present-day outcome (the Brexit referendum) to check if the results are picking up on a more general demographic and economic correlation pattern; the experiment shows no spurious correlations.
Link to Full Report: https://osf.io/gdne3/
Link to Replicators' Package: https://osf.io/tws8n/
Original Authors' Response: No response.
Original Authors' Package: https://zenodo.org/records/5734954

Reproduction Report

Title Original Study: Understanding Ethnolinguistic Differences: The Roles of Geography and Trade

doi: https://doi.org/10.1093/ej/ueab065, Economic Journal

Report's Abstract: Dickens (2022) studies the role of trade on long-run inter-ethnic linguistic differences. He establishes that neighboring ethnolinguistic groups have smaller (lexicostatistical) linguistic distances when there is a larger agricultural productivity variation between them. Specifically, he establishes that pre-1500 land productivity variation (CSI SD) and its change due to Columbian Exchange in the post-1500 (CSI SD CHANGE) era decrease linguistic distances between groups. In what can be considered his main specification, which includes geographical controls, spatial controls, and language family fixed effects (Table 1 column 5), he estimates that a one standard deviation increase in the change in land productivity variation (post-1500) decreases linguistic distances by 0.11 standard deviations (p-value ¡ 0.01) and a one standard deviation increase in land productivity variation (pre-1500) decreases linguistic distances by 0.06 standard deviations (p-value = 0.12). We conduct a direct replication of the paper by (i) reconstructing the main independent variables using the same original sources and following the procedures explained in the original study, (ii) using an updated version of the linguistic map (Ethnologue v17 instead of v16), and (iii) constructing alternative measures of inter-ethnic potential gains from trade. Our results basically confirm the sign, magnitude, and statistical significance of the point estimates in the original study.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/62.htm

Link to Replicators' Package: https://osf.io/k3p7g/

Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/63.htm

Original Authors' Package: https://doi.org/10.1093/ej/ueab065

Reproduction Report

Title Original Study: Vulnerability and Clientelism
doi: https://doi.org/10.1257/aer.20190565, American Economic Review
Report's Abstract: The paper estimates the effect that changes in household vulnerability have on citizens' participation in clientelist relationships. The authors exploit two sources of variation in household vulnerability: rainfall shocks, and a randomized intervention that provided cisterns in drought-prone areas. We reproduce all the findings presented in the four main results tables presented in the paper. The results of our robustness replication show that the results in the original paper are robust to variations in the rainfall period used as a baseline to assess changes in household vulnerability, and to exclusions that eliminate individuals in the sample who may have been substituted with others at different survey points. However, some of the original results that explain the underlying mechanisms are sensitive to how "clientelist relationships" are defined. When more frequent interactions with politicians are used as the defining characteristic of households in clientelist relationships, we find that the original results suggesting clientelism as a significant mechanism are no longer statistically significant at any standard significance level. We note, however, that the authors, in a reply to questions we sent them after the Replication Games, convincingly show that their results are robust to changing the definition of the clientelist marker.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/83.htm
Link to Replicators' Package: https://osf.io/q2tw6/
Link to Original Authors' Response: https://econpapers.repec.org/paper/zbwi4rdps/84.htm
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/173341/version/V1/view

Reproduction Report

Title Original Study: Wage Cyclicality and Labor Market Sorting
doi: https://doi.org/10.1257/aeri.20210161, American Economic Review: Insights
Report's Abstract: Figueiredo (2022) examines wage cyclicality across the skill mismatch distribution finding large differences. Some key results include finding that wages are acyclical in good labor market matches but procyclical in poor matches. Using the public replication material provided by the authors, we were able to exactly duplicate the results of the study. Further, using several further robustness checks, such as subtracting (potentially correlated) covariates in the regressions, using different standard errors (rather than clustered ones), or different time periods of the data left the key results largely unchanged with some minor caveats.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/78.htm
Link to Replicators' Package: https://osf.io/a8hcg/
Original Authors' Response: "I have read the report and I do not wish to write a reply.

Congratulations on this initiative – it is great!"
Original Authors' Package: https://www.openicpsr.org/openicpsr/project/150581/version/V1/view

Reproduction Report

Title Original Study: War, Socialism, and the Rise of Fascism: an Empirical Exploration

doi: https://doi.org/10.1093/qje/qjac001, Quarterly Journal of Economics

Report's Abstract: In this report, we present the results from a replication of Acemoglu et al. (2022). The authors suggest that the emergence of the 'Red Scare' in the aftermath of World War I led to a rise of fascism in Italy in the early 1920s. Their approach uses the war casualties as an instrument for the rise in socialist voting. We performed a series of replication strategies, including pre-trend controls, applying an alternative instrument and modifying the first-stage specification, as well as investigating the long-run political alignment. Based on our findings, the original authors' results are replicable under a variety of alternative specifications.

Link to Full Report: https://osf.io/a672c/

Link to Replicators' Package: https://osf.io/a672c/

Link to Original Authors' Response: No response.

Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CLJTSC

Reproduction Report

Title Original Study: What Makes Anticorruption Punishment Popular? Individual-Level Evidence from China
doi: https://doi.org/10.1086/715252, Journal of Politics
Report's Abstract: t also indirect effects through affecting evaluations of competence and morality. Conducting a conjoint study in China where respondents were asked to choose between two potential local officials, Tsai et al. found that 26% of the total effect of these officials punishing corrupt subordinates was estimated to come through indirect effects that go through evaluations of morality and competence. Using their code, I reproduced their original findings, and did not find any notable coding errors while doing so. Then, taking advantage of the fact that Tsai et al. included several additional covariates beyond punishment in their experiment, I engaged in an extension of the original model, using the same method, to examine whether economic performance characteristics have indirect effects on evaluation through competence and morality as well. I found results that suggest that economic performance does have an indirect effect on preferences through competence and morality. I then tested the robustness of Tsai et al.'s original heterogeneous sensitivity tests by varying cut points on two demographic variables and found that their findings of a lack of heterogeneous sensitivity remain robust to different cut-points. In all, my efforts suggest that Tsai et al.'s methods are valid and their findings robust.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/7.htm
Link to Replicators' Package: https://osf.io/czs6j/
Original Authors' Response: "We appreciate your efforts, both in replicating our paper and in doing so systematically for other studies in leading political science and economic journals. Your contribution is valuable to the entire academic community and to us especially.

We also appreciate your sharing Reproduction Reports with the original authors prior to dissemination and are glad to see from the Reproduction Report that our results and methods appear to be both valid and robust. Although a longer follow-up may not be necessary, we do wish to convey our gratitude to the replicator(s) and to the editorial team."
Original Authors' Package: https://dataverse.harvard.edu/dataset.xhtml;jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=date

Reproduction Report

Title Original Study: When a Doctor Falls from the Sky: The Impact of Easing Doctor Supply Constraints on Mortality

doi: https://doi.org/10.1257/aer.20210701, American Economic Review

Report's Abstract: Okeke (2023) evaluates a policy experiment conducted in Nigeria, whereby communities were randomly allocated to receive a new doctor at the local public health center. The performance of these centers was compared to other sites which were allocated either a new midlevel health-care provider, or no additional staff. The study finds that communities assigned a new doctor were associated with a decrease in seven-day infant mortality, such a decrease was not observed in communities assigned a midlevel health-care provider. This suggests that it is the 'quality' of the additional doctor driving the effects rather than due to a quantity increase of an additional health worker. The size of the mortality reduction increased with increased exposure to the intervention. We first conduct a computational reproduction, rerunning the original code and data, finding that the results reported in the original study are reproducible. Second, we test the robustness of the results in several ways, by 1) adapting the existing controls to make the results robust to contamination bias, 2) altering and adding to the control variables included, 3) changing the specification or regression technique used, and 4) testing coding grouping and changing how service use was coded. These changes cause little change to the point estimates, although we find that the original paper's standard errors were overly conservative, and thus the statistical significance of some results was understated.

Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/53.htm

Link to Replicators' Package: https://github.com/e-mcmanus/Okeke23_Replication

Original Authors' Response: "Thank you for sharing the Reproduction Report (and please pass on my thanks to the replicators). There does not appear to be much for me to respond to. It is gratifying to see that the results have held up well to additional scrutiny."

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/181581/version/V1/view

Reproduction Report

Title Original Study: Who Chooses Commitment? Evidence and Welfare Implications

doi: https://doi.org/10.1093/restud/rdab056, Review of Econoimc Studies

Report's Abstract: We conduct a computational reproduction and a robustness replication of Carrera et al. (2022) by using the same dataset and similar procedures as specified in their paper (i.e., method and analysis). Instead of using STATA, we use R and code the results from scratch. We also replicate the MATLAB code used for simulations and test whether it produces reasonable results for different parameter values. We confirm all of the main results and do not find high sensitivity of the model to changes in parameters.

Link to Full Report: https://osf.io/752q9/

Link to Replicators' Package: https://osf.io/752q9/

Link to Original Authors' Response: The authors provided feedback which was taken into account.

Original Authors' Package: https://zenodo.org/records/5173081

Reproduction Report

Title Original Study: Who Sells During a Crash? Evidence from Tax Return Data on Daily Sales of Stock

doi: https://doi.org/10.1093/ej/ueab059, Economic Journal

Report's Abstract: Hoopes et al., (2021) analyze United States tax return data encompassing all individual taxable stock sales between 2008 and 2009, to investigate the individuals who increased their stock sales in response to market turbulence. Our findings reveal that such increases were notably prevalent among investors in the highest tiers of the income distribution, including the top 1% and 0.1%, as well as retirees and those at the uppermost levels of the dividend income distribution. We reproduce the paper's main findings and results are very similar.

Link to Full Report: https://osf.io/b6s9k/

Link to Replicators' Package: https://www.dropbox.com/scl/fo/c3ysdlenysq391mugzprm/h?rlkey=riooyohci7i5vwx475r13jaqq&dl=0

Original Authors' Response: The authors provided feedback which was taken into account.

Original Authors' Package: https://doi.org/10.1093/ej/ueab059

Reproduction Report

Title Original Study: Why Don't Firms Hire Young Workers During Recessions?
doi: https://doi.org/10.1093/ej/ueab096, Economic Journal
Report's Abstract: We gauge the replicability of the results of Forsythe (2022) studying the cyclicality of individuals' labor market transitions conditional on their experience. Using Current Population Survey (CPS) data and state-level variation in cyclical unemployment, this paper shows that the hiring probability of youths is more sensitive to business-cycle conditions than that of experienced individuals. We replicate the main results in this paper by reconstructing the dataset using data from the IPUMS-CPS database (Flood et al. (2020)) and recoding the paper's main regressions. We also conduct a robustness replicability analysis and show that the paper's main results are robust in terms of statistical significance to (i) extending the sample period from 1994-2014 to 1994-2019 and (ii) using MSA-level unemployment variation instead of state-level variation. However, these extensions reduce the magnitude of the main effects of interest. The paper's key conclusions are unaffected.
Link to Full Report: https://osf.io/3pqbt/
Link to Replicators' Package: https://github.com/jcrechet/replication_forsythe_2022_EJ
Link to Original Authors' Response: The author responded but did not provide a response.
Original Authors' Package: https://zenodo.org/records/5710784

Reproduction Report

Title Original Study: Yellow Vests, Pessimistic Beliefs, and Carbon Tax Aversion
doi: https://doi.org/10.1257/pol.20200092, American Economic Journal: Economic Policy
Report's Abstract: Douenne and Fabre (2022) implement a representative survey following the Yellow Vests movement in France that started in opposition to the carbon tax in 2018. They find that a majority of French citizens would oppose a carbon tax and dividend program with proceeds paid equally to each adult. The authors further find that respondents have pessimistic beliefs about several aspects of the policy. They then show how informational treatments cause respondents to update these beliefs, and they finally estimate the causal effect of these beliefs on support for the policy. In this note, we focus on the second section of this paper: the causal effects of feedback on beliefs. Based on elicited household characteristics, Douenne and Fabre (2022) estimate whether each household "wins" or "loses" from the carbon tax and dividend reform. They provide this binary (win vs. lose) information to households and subsequently ask households to evaluate whether they believe they would financially benefit from the policy. By exploiting the discontinuity in win vs. lose feedback, they assess the degree to which feedback affects subjective beliefs, finding that a household that is told it will "win" as a result of the reform increases its subjective belief that it will not lose by about 25 percentage points. The subset of households that is part of the Yellow Vests movement, however, revises its subjective belief of not losing upwards by only 10 percentage points after being told that it will "win" from the carbon tax reform. Conversely, households who initially support the tax increase this belief by 41 percentage points when told they will "win." In this note we replicate this second section of the paper- the causal effects of feedback on beliefs- using the processed data provided by the authors. We successfully replicate the average treatment effect, but we find that the heterogeneous treatment effects may be biased due to model misspecification. While our results support the conclusion that these estimated effects depend on a household's attitudes toward the policy, we find that the source of heterogeneity differs. Further, we note two changes to the analysis that we believe are appropriate (which do not affect the conclusions drawn): first, some (1.8%) of observations in the dataset appear to be misclassified-wrongly coded as if a household would "lose" when in fact they would "win"-and second, the main causal analysis is based on a regression discontinuity design, but does not include standard components of such a design (e.g., a RD plot, optimal selection of bandwidth, density analysis, placebo tests). We update the design to address both of these points. We find results that generally support the main conclusions of Douenne and Fabre (2022), but we urge caution when interpreting the heterogeneous treatment effects.
Link to Full Report: https://econpapers.repec.org/paper/zbwi4rdps/58.htm
Link to Replicators' Package: https://github.com/karemanyassin/Yellow-Vests-Pessimistic-Beliefs-and-Carbon-Tax-Aversion-2022-A-Comment
Original Authors' Response: Authors provided feedback which was taken into account. No response.

Original Authors' Package: https://www.openicpsr.org/openicpsr/project/128143/version/V1/view

References

1. Brodeur, A. *et al. Mass Reproducibility and Replicability: A New Hope* I4R Discussion Paper 107. 2024.
2. Clark, C. J. & Tetlock, P. E. in *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions* 905–927 (Springer, 2023).
3. Brodeur, A., Sung, S. Y., Miguel, E., Vilhuber, L. & de la Guardia, F. H. *Assessing Reproducibility in Economics Using Standardized Crowd-sourced Analysis* NBER Working Paper 33753. 2024.
4. Brodeur, A., Cook, N. & Neisser, C. P-Hacking, Data Type and Data-Sharing Policy. *Economic Journal* **134,** 985–1018 (2024).
5. Brodeur, A., Cook, N. & Heyes, A. Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* **110,** 3634–3660 (2020).
6. Gerber, A. S. & Malhotra, N. Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods & Research* **37,** 3–30 (2008).
7. Andrews, I. & Kasy, M. Identification of and Correction for Publication Bias. *American Economic Review* **109,** 2766–94 (2019).
8. Elliott, G., Kudrin, N. & Wüthrich, K. Detecting p-Hacking. *Econometrica* **90,** 887–906 (2022).
9. Fišar, M. *et al.* Reproducibility in Management Science. *Management Science* (2023).