

Brodeur, Abel et al.

Working Paper

Computational Reproducibility and Robustness of Empirical Economics and Political Science Research Between 2022 and 2023

I4R Discussion Paper Series, No. 287

Provided in Cooperation with:

The Institute for Replication (I4R)

Suggested Citation: Brodeur, Abel et al. (2026) : Computational Reproducibility and Robustness of Empirical Economics and Political Science Research Between 2022 and 2023, I4R Discussion Paper Series, No. 287, Institute for Replication (I4R), s.l.

This Version is available at:

<https://hdl.handle.net/10419/338965>

Standard-Nutzungsbedingungen:

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

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

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

Terms of use:

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

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

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

No. 287

I4R DISCUSSION PAPER SERIES

Computational Reproducibility and Robustness of Empirical Economics and Political Science Research Between 2022 and 2023

Abel Brodeur et al.

March 2026

I4R DISCUSSION PAPER SERIES

I4R DP No. 287

Computational Reproducibility and Robustness of Empirical Economics and Political Science Research Between 2022 and 2023

Abel Brodeur^{1,2} et al.

¹University of Ottawa/Canada

²Institute for Replication

MARCH 2026

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur
University of Ottawa

Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research

Computational Reproducibility and Robustness of
Empirical Economics and Political Science
Research Between 2022 and 2023
Forthcoming in *Nature*

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. We reproduced original analyses and conducted robustness checks of 110 articles recently published in leading economics and political science journals (all of which have mandatory data and code sharing policies). We found that over 85% of published claims were computationally reproducible. In robustness checks, our re-analyses led to 72% of statistically significant estimates to remain significant and in the same direction, and the median reproduced effect size is (nearly) the same as the originally published effect size (that is, 99% of the published effect size). Additionally, six independent research teams examined 12 pre-specified hypotheses about determinants of robustness. Research teams with more experience found lower levels of robustness, and robustness correlated with neither author characteristics nor data availability.

Keywords: Reproduction, Robustness, Research Transparency, Open Science, Economics, Political Science

25 1 Introduction

26 Science aspires to be cumulative. Reproducibility efforts strengthen science by testing
 27 the reliability of published findings, promoting self-correction, and informing policy-
 28 making [1]. Computational reproductions, whereby independent researchers reproduce
 29 the results of published studies, are an essential diagnostic tool [2–10]. Such efforts
 30 should have greater visibility [11–16]. However, there has been little social science
 31 reproduction and robustness conducted at scale [10, 13, 17–23].

32 This project is a mega-reproduction led by the Institute for Replication (I4R),
 33 which evaluates the reproducibility and robustness of 110 published studies in eco-
 34 nomics (79) and political science (31). Our focus is on studies published in 12
 35 prestigious journals between 2022 and 2023. While each of these journals has a data
 36 and code availability policy requiring authors to publicly share their materials upon
 37 publication, most (though not all) also appoint a dedicated data editor. This edi-
 38 tor is responsible for enforcing the journal’s data and code policy and conducting
 39 internal computational reproducibility checks for accepted studies (see Supplementary
 40 Materials 12.8).

41 Not all studies from our targeted journals were chosen for reproduction and
 42 robustness, and our sample is thus not a random representative sample of studies
 43 in economics and political science. Our approach leads to an over-representation of
 44 studies using publicly available data ([18]). Another feature of our sample is that the
 45 targeted journals have a data availability policy *and* enforce it. This is in contrast to
 46 many top field journals in both economics and political science. Our sample should
 47 thus be viewed as very selective both in terms of impact and high data and code avail-
 48 ability rates, and might present an optimistic upper bound on reproducibility rates.
 49 In fact, virtually all papers in our sample include replication packages with cleaned
 50 data and code to reproduce the paper’s results, and about 30% also provide the raw
 51 data and cleaning code used to generate the analytical data (Supplementary Materials
 52 Appendix Figure 5, Levels 8, 9, and 10).

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

65 2 Definitions

66 We follow [27]’s nomenclature and define a claim **computationally reproducible** if
 67 its results can be reproduced using the original study’s data and protocols. A claim

68 is **robust** if its results are robust to alternative reasonable analytical decisions on the
 69 same data. Last, a claim is **replicable** if its results can be repeated using new data.

70 3 Teams and Choice of Study

71 The reproductions and replications in this project are generated in one of two streams.
 72 First, I4R has a board of editors who recommend potential reproducers. Second, I4R
 73 held 11 events called replication games (Games) ([28]). Games are one-day events
 74 open to faculty, post-docs, graduate students and other researchers. Participants are
 75 assigned to a small team of about 3–5 other researchers all working in the same subfield
 76 (*e.g.*, development economics).

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

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

88 We survey the reasons why teams selected their paper (Supplementary Materials
 89 Figure 8). While 13.6% of teams were assigned a study (*i.e.*, did not choose which
 90 study to work on), about 45% of teams report “Methods used”, 36% of teams selected
 91 “because of the journal of publication” and about 25% due to the “length of time to
 92 reproduce results”.

93 If a large portion of reproducers select papers based on the assumption that their
 94 findings are questionable, it could skew reproducibility rates downward, as such studies
 95 might be more prone to revealing problematic outcomes. However, in this project,
 96 only a minimal fraction of teams indicated that they chose their paper because of *ex*
 97 *ante* beliefs that main results are (not) replicable (3.6%). We found that selecting a
 98 paper due to the reproducers’ belief the paper is not robust is *inversely* correlated
 99 with reproducer experience ($\rho = -0.19, p < 0.000$). A few teams (5%) indicated that
 100 their choice was based on statistical power/sample size and trust of original authors.

101 4 Data Availability and Computational 102 Reproducibility

103 We find a computational reproducibility rate of 85%. That is, when provided with the
 104 original data and code, independent researchers are able to reproduce the published
 105 results in economics and political science studies 85% of the time using either: (1)
 106 the raw and analytical data, or; (2) the analytical data when the raw data were not
 107 provided. The remaining 15% of cases involved studies with only partial availability
 108 of code or data, or instances where code failed to run or produced inconsistent results

(See Supplementary Materials 12.11 and Appendix Figure 5). Fixing paths, missing packages and software requirements were not considered failures of computationally reproduce. 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 12.16, 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. 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. 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.

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 reproducibility 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 *insignificant* 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, 45.33% of original-re-analyses pairs represent a statistically significant result that is the same under re-analysis. Of particular note is the 15.06% of re-analyses that find a statistically *insignificant* 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 $\chi^2 = 264.11$, $p < 0.001$, $n = 4750$).

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, 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. 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.

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 relationships are far more likely to be positive than negative. We interpret this result that reproducers that 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 teams had been assigned to each paper. According to the teams, the provision of raw or intermediate data, or the necessary cleaning codes, does not seem to affect the robustness of research.

When analysts examined these same 12 hypotheses for originally statistically insignificant results, the relationships 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 Under Reproduction

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).

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 average replicated effect is 9% larger than the original. Supplementary Materials Appendix Figure 6 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 12.13). 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

281 minor differences for 10 studies. Two of the major differences were uncovered when
282 using a different software and looking at the authors' code. Additionally, one team who
283 computationally reproduced the results using a different *version* of the software used
284 by the authors uncovered noteworthy differences in the magnitude and significance of
285 the estimates (Supplementary Materials [12.12](#)).

286 9 Communication with Original Authors

287 I4R shares completed reproduction reports with original authors before public release
288 ([28]). Reports are reviewed typically by A.B. or another board member mainly for
289 tone and structure. I4R then disseminates the report and any author response simul-
290 taneously (see Supplementary Materials for the full list of reports). Reproducers may
291 revise their reports after receiving feedback from original authors.

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

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

301 Lastly, we assess agreement between authors and reproducers. Authors' final
302 responses were coded for whether disagreements remained after mediation; only 23% of
303 articles showed any remaining disagreement. Further details appear in Supplementary
304 Materials [12.6](#).

305 10 Discussion

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

311 Large-scale reproducibility initiatives offer a promising way to address these chal-
312 lenges in the social sciences and beyond. Our project provides a systematic, scalable
313 approach to evaluating reproducibility and robustness, with the goal of increasing
314 transparency and improving the credibility of published research.

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

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

323 We also asked reproducers about the quality of the replication packages they exam-
324 ined. Just over 40% reported gaining a more optimistic view of the discipline, while
325 about 45% reported no change. This suggests that mass reproduction can directly
326 enhance researchers' trust in scientific findings.

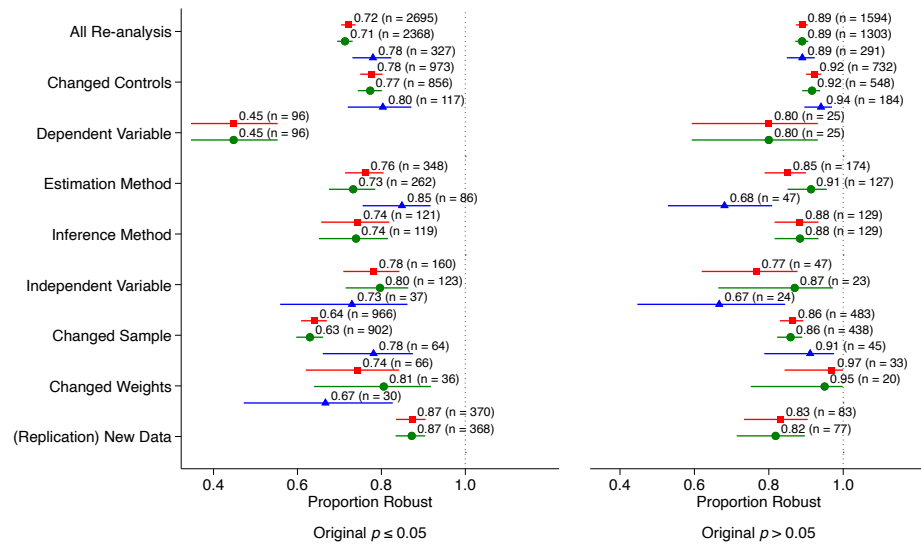
327 The initiative's success and scalability have been driven by the intrinsic motivation
328 of participating researchers to support open science and improve their technical skills.
329 By late 2025, I4R had organized 80 replication games involving over 3,500 researchers,
330 with events held every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

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

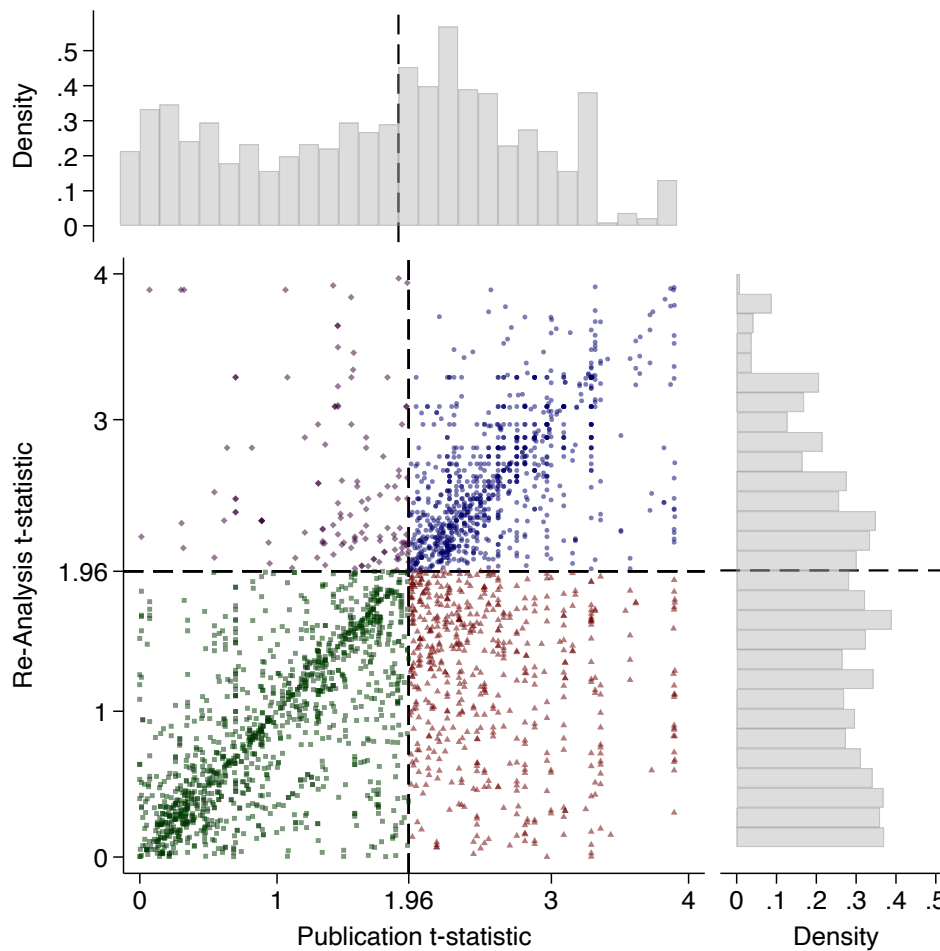
336 Our results have limitations. Only a small number of economics and political
337 science journals currently require data and code [17]; [18], and even fewer check
338 reproducibility [39]. Thus, our findings largely reflect leading journals with strong
339 data-sharing norms. Future research should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

342 11 Figures

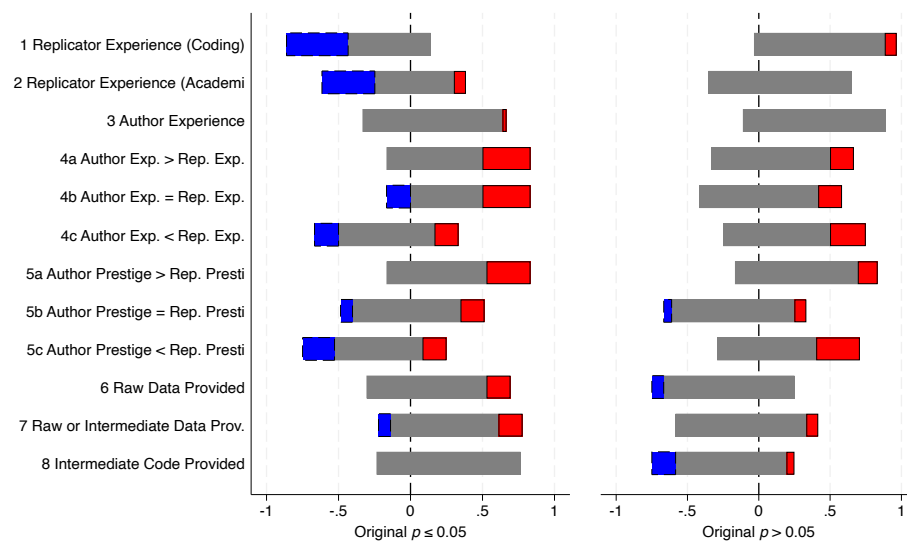
Fig. 1: Robustness Rate



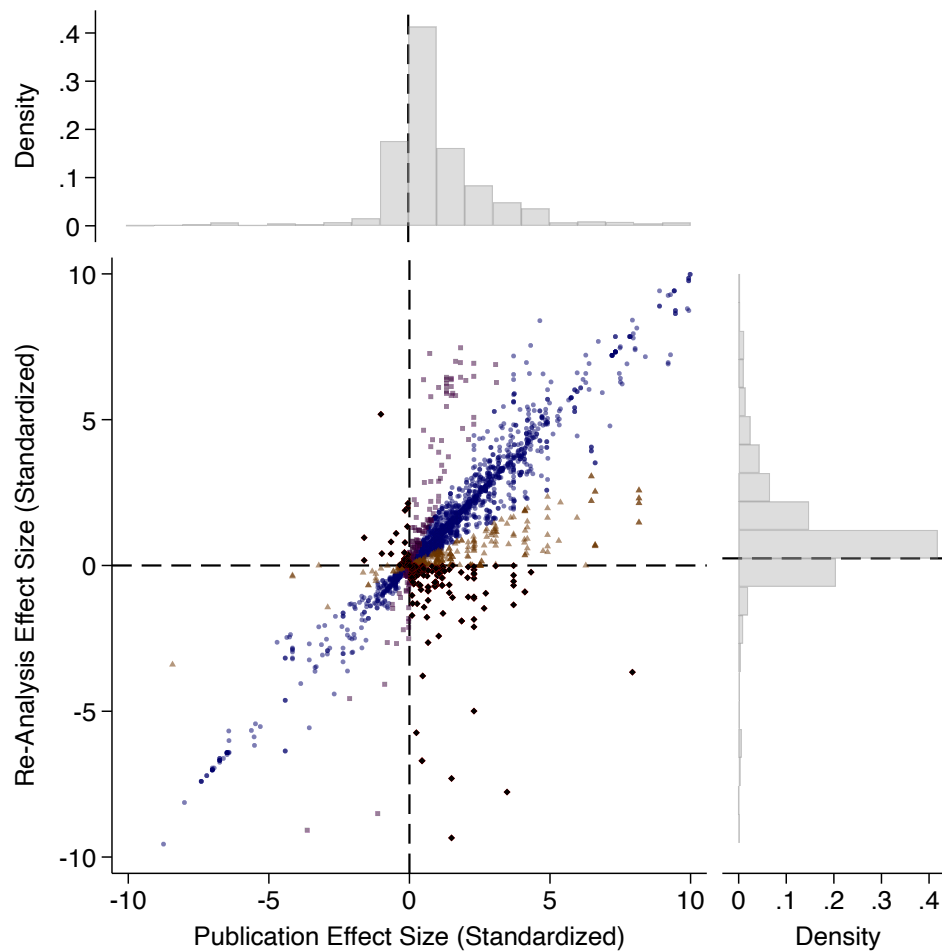
Robustness rate for ... **Left panel:** ... originally statistically significant research **Right panel:** ... originally statistically insignificant research **All panels:** Red squares represent full sample. Green circles represent economics subsample. Blue triangles represent political science subsample. 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. 95% confidence intervals presented in bar and whiskers.

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

Top histogram: Distribution of publication tests of significance. Tests 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 replication tests of significance. Tests over 4 truncated for exposition. **Scatterplot:** Each marker is a pair of test statistics, an originally published test statistic (horizontal value) and an associated replication 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** are the remaining observations are the 6% of test statistics that represent a sign reversal between the originally estimated effect and the effect estimated under re-analysis.

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.

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 the ratio of re-analysis 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 replication effect size (vertical value). If an originally estimated and replication 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.

343 Author list

344 Abel Brodeur, Derek Mikola, Nikolai Cook, Lenka Fiala, Thomas Brailey, Ryan Briggs,
 345 Alexandra de Gendre, Yannick Dupraz, Jacopo Gabani, Romain Gauriot, Joanne
 346 Haddad, Goncalo Lima, Jörg Ankel-Peters, Anna Dreber, Douglas Campbell, Lamis
 347 Kattan, Diego Marino Fages, Fabian Mierisch, Pu Sun, Taylor Wright, Marie Connolly,
 348 Fernando Hoces de la Guardia, Magnus Johannesson, Edward Miguel, Lars Vilhuber,
 349 Alejandro Abarca, Mahesh Acharya, Sossou Simplicie Adjisse, Ahwaz Akhtar, Eduardo
 350 Alberto Ramirez Lizardi, Sabina Albrecht, Synøve Nygaard Andersen, Zubaria Andlib,
 351 Falak Arrora, Thomas Ash, Etienne Bacher, Sebastian Bachler, Félix Bacon, Manuel
 352 Bagues, Timea Balogh, Alisher Batmanov, Mara Barschkett, B. Kaan Basdil, Jaromír
 353 Baxa, Sascha Becker, Monica Beeder, Louis-Philippe Beland, Abdel-Hamid Bello,
 354 Daniel Benenson Markovits, Grant Benjamin, Thomas Bergeron, Moussa Blimpo,
 355 Marco Binetti, Carl Bonander, Joseph Bonneau, Endre Borbáth, Nicolai Topstad Bor-
 356 gen, Solveig Topstad Borgen, Jonathan Borowsky, Elisa Brini, Myriam Brown, Martin
 357 Brun, Stephan Bruns, Nino Buliskeria, Andrea Calef, Alistair Cameron, Pamela
 358 Campa, Santiago Campos-Rodríguez, Giulio Giacomo Cantone, Fenella Carpena,
 359 Perry Carter, Paul Castañeda Dower, Ondrej Castek, Jill Caviglia-Harris, Gabriella
 360 Chauca Strand, Shi Chen, Sya In Chzhen, Jong Chung, Jason Collins, Alexan-
 361 der Coppock, Hugo Cordeau, Ben Couillard, Jonathan Crechet, Lorenzo Crippa,
 362 Jeanne Cui, Christian Czymara, Haley Daarstad, Danh Chi Dao, Daniel Dao, Marco
 363 David Schmandt, Astrid de Linde, Lucas De Melo, Lachlan Deer, Micolé De Vera,
 364 Velichka Dimitrova, Jan Fabian Dollbaum, Jan Matti Dollbaum, Michael Donnelly,
 365 Luu Duc Toan Huynh, Tsvetomira Dumbalska, Jamie Duncan, Kiet Tuan Duong,
 366 Thibaut Duprey, Christoph Dworschak, Sigmund Ellingsrud, Ali Elminejad, Yasmine
 367 Eissa, Andrea Erhart, Giulian Etingin-Frati, Elaheh Fatemipour, Alexa Federice,
 368 Jan Feld, Guidon Fenig, Mojtaba Firouzjaeiangalougah, Erlend Fleisje, Alexandre
 369 FortiFriter-Chouinard, Julia Francesca Engel, Nadjim Fréchet, Reid Fortier, Tilman
 370 Fries, Michael James Frith, Thomas Galipeau, Sebastián Gallegos, Areez Gangji,
 371 Xiaoying Gao, Cloé Garnache, Attila Gáspár, Evelina Gavrilova, Arijit Ghosh, Gar-
 372 reth Gibney, Grant Gibson, Geir Godager, Leonard Goff, Da Gong, Javier González,
 373 Jeremy D. Gretton, Cristina Griffo, Idaliya Grigoryeva, Maja Grøtting, Eric Gun-
 374 termann, Jiaqi Guo, Alexi Gugushvili, Hooman Habibnia, Sonja Häffner, Jonathan
 375 D. Hall, Olle Hammar, Amund Hanson Kordt, Barry Hashimoto, Jonathan S. Hart-
 376 ley, Carina I. Hausladen, Tomáš Havránek, Harry He, Matthew Hepplewhite, Mario
 377 Herrera-Rodriguez, Felix Heuer, Anthony Heyes, Anson T. Y. Ho, Jonathan Holmes,
 378 Armando Holzknecht, Yu-Hsiang Dexter Hsu, Shiang-Hung Hu, Yu-Shiuan Huang,
 379 Mathias Huebener, Christoph Huber, Kim P. Huynh, Zuzana Irsova, Ozan Isler,
 380 Niklas Jakobsson, Raphaël Jananji, Tharaka A. Jayalath, Michael Jetter, Jenny John,
 381 Rachel Joy Forshaw, Felipe Juan, Valon Kadriu, Sunny Karim, Edmund Kelly, Duy
 382 Khanh Hoang Dang, Tazia Khushboo, Jin Kim, Gustav Kjellsson, Anders Kjelsrud,
 383 Andreas Kotsadam, Jori Korpershoek, Lewis Krashinsky, Suranjana Kundu, Alexan-
 384 der Kustov, Nurlan Lalayev, Audrée Langlois, Jill Laufer, Blake Lee-Whiting, Andreas
 385 Leibing, Gabriel Lenz, Joel Levin, Peng Li, Tongzhe Li, Yuchen Lin, Ariel Listo,
 386 Dan Liu, Xuewen Lu, Elvina Lukmanova, Alex Luscombe, Lester R. Lusher, Ke
 387 Lyu, Hai Ma, Nicolas Mäder, Clifton Makate, Alice Malmberg, Adit Maitra, Marco

388 Mandas, Jan Marcus, Shushanik Margaryan, Lili Márk, Andres Martignano, Abi-
 389 gail Marsh, Isabella Masetto, Anthony McCanny, Emma McManus, Ryan McWay,
 390 Lennard Metson, Jonas Minet Kinge, Sumit Mishra, Myra Mohnen, Jakob Möller, Ros-
 391 alie Montambeault, Sébastien Montpetit, Louis-Philippe Morin, Todd Morris, Scott
 392 Moser, Fabio Motoki, Lucija Muehlenbachs, Andreea Musulan, Marco Musumeci,
 393 Munirul Nabin, Karim Nchare, Florian Neubauer, Quan M. P. Nguyen, Tuan Nguyen,
 394 Viet Nguyen-Tien, Ali Niazi, Giorgi Nikolaishvili, Ardyn Nordstrom, Patrick Nüß,
 395 Angela Odermatt, Matt Olson, Henning Øien, Tim Ölkens, Miquel Oliver i Vert,
 396 Emre Oral, Christian Oswald, Ali Ousman, Ömer Özak, Shubham Pandey, Alexan-
 397 dre Pavlov, Martino Pelli, Romeo Penheiro, RyuGyung Park, Eva Pérez Martel,
 398 Tereza Petrovičová, Linh Phan, Alexa Prettyman, Jakub Procházka, Aqila Putri,
 399 Julian Quandt, Kangyu Qiu, Loan Quynh Thi Nguyen, Andaleeb Rahman, Carson
 400 H. Rea, Adam Reiremo, Laëtitia Renée, Joseph Richardson, Nicholas Rivers, Bruno
 401 Rodrigues, William Roelofs, Tobias Roemer, Ole Rogeberg, Julian Rose, Andrew
 402 Roskos-Ewoldsen, Paul Rosmer, Barbara Sabada, Soodeh Saberian, Nicolas Sala-
 403 manca, Georg Sator, Daniel Scates, Elmar Schlüter, Cameron Sells, Sharmi Sen, Ritika
 404 Sethi, Anna Shcherbiak, Moyosore Sogaolu, Matt Soosalu, Erik Ø. Sørensen, Manali
 405 Sovani, Noah Spencer, Stefan Staubli, Renske Stans, Anya Stewart, Felix Stips, Kieran
 406 Stockley, Stephenson Strobel, Ethan Struby, John Tang, Idil Tanrisever, Thomas Tao
 407 Yang, Ipek Tastan, Dejan Tatić, Benjamin Tatlow, Féraud Tchuisseu Seuyong, Rémi
 408 Thériault, Vincent Thivierge, Wenjie Tian, Filip-Mihai Toma, Maddalena Totarelli,
 409 Van Tran, Hung Truong, Nikita Tsoy, Kerem Tuzcuoglu, Diego Ubfal, Laura Villalo-
 410 bos, Julian Walterskirchen, Joseph Tao-yi Wang, Vasudha Wattal, Matthew D. Webb,
 411 Bryan Weber, Reinhard Weisser, Wei-Chien Weng, Christian Westheide, Kimberly
 412 White, Jacob Winter, Timo Wochner, Matt Woerman, Jared Wong, Ritchie Woodard,
 413 Marcin Wroński, Myra Yazbeck, Chung Yang, Luther Yap, Kareman Yassin, Hao Ye,
 414 Jin Young Yoon, Chris Yurris, Tahreen Zahra, Mirela Zaneva, Aline Zayat, Jonathan
 415 Zhang, Ziwei Zhao, Yaolang Zhong

416 Declarations

417 Some journals require declarations to be submitted in a standardised format. Please
 418 check the Instructions for Authors of the journal to which you are submitting to see if
 419 you need to complete this section. If yes, your manuscript must contain the following
 420 sections under the heading ‘Declarations’:

- 421 • Funding:
 422 We acknowledge support from Coefficient Giving and the Social Sciences and
 423 Humanities Research Council.
- 424 • Conflict of interest/Competing interests:
 425 Any views expressed herein are the authors’ personal opinions and not those of
 426 Ontario Public Service. The work by Jeremy D. Gretton was not undertaken under
 427 the auspices of Ontario Public Service as part of his employment responsibilities.
 428 The views expressed in this paper are those of the authors. No responsibility for
 429 them should be attributed to the Bank of Canada. The findings, interpretations,
 430 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 and availability: The data and codes are available on zenodo (<https://zenodo.org/records/17792605>) and OSF (<https://osf.io/8wsqx/>). See OSF our pre-analysis plan.

- 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 (CNRS, Université Paris-Dauphine, PSL Research University, IRD, UMR LEDa, DIAL), 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 Autònoma 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

476 (Luxembourg Institute of Socio-Economic Research), Sebastian Bachler (Univer-
 477 sity of Innsbruck), Félix Bacon (Laval University), Manuel Bagues (University of
 478 Warwick), Timea Balogh (UC Davis), Alisher Batmanov (UC San Diego), Mara
 479 Barschkett (University of Bonn, IZA & DIW Berlin), B. Kaan Basdil (Risktürk),
 480 Jaromír Baxa (Institute of Economic Studies, Faculty of Social Sciences, Charles
 481 University, and Institute of Information Theory and Automation AS CR), Sascha
 482 Becker (University of Warwick and Monash University), Monica Beeder (University
 483 of Southampton), Louis-Philippe Beland (Carleton University), Abdel-Hamid Bello
 484 (Université Laval), Daniel Benenson Markovits (Columbia University), Grant Ben-
 485 jamin (University of Toronto), Thomas Bergeron (University of Toronto), Moussa
 486 P. Blimpo (University of Toronto), Marco Binetti (University of the Bundeswehr
 487 Munich), Carl Bonander (University of Gothenburg), Joseph Bonneau (UC Davis),
 488 Endre Borbáth (Ruprecht-Karls-Universität Heidelberg), Nicolai Topstad Borgen
 489 (Centre for Research on Equality in Education, University of Oslo), Solveig Topstad
 490 Borgen (University of Oslo), Jonathan Borowsky (University of Minnesota), Thomas
 491 Brailey (University of Oxford), Ryan Briggs (University of Guelph), Elisa Brini
 492 (University of Florence), Myriam Brown (Laval University), Martin Brun (Tampere
 493 University), Stephan Bruns (Hasselt University), Nino Buliskeria (Nazarbayev Uni-
 494 versity), Andrea Calef (University College London, School of Management), Alistair
 495 Cameron (Monash University), Pamela Campa (Stockholm Institute of Transi-
 496 tion Economics), Santiago Campos-Rodríguez (University of California, Irvine),
 497 Giulio Giacomo Cantone (Magna Graecia University), Fenella Carpena (Oslo Busi-
 498 ness School, Oslo Metropolitan University), Perry Carter (NYU Abu Dhabi), Paul
 499 Castañeda Dower (University of Wisconsin-Madison), Ondrej Casteck (Masaryk
 500 University), Jill Caviglia-Harris (Salisbury University), Gabriella Chauca Strand
 501 (Institute of Medicine, University of Gothenburg), Shi Chen (Queen's University),
 502 Sya In Chzhen (University of East Anglia), Jong Chung (Auburn University), Jason
 503 Collins (University of Technology Sydney), Alexander Coppock (Yale University),
 504 Hugo Cordeau (University of Toronto), Ben Couillard (University of Toronto),
 505 Jonathan Crechet (University of Ottawa), Lorenzo Crippa (University of Strath-
 506 clyde), Jeanne Cui (Beijing Normal University), Christian Czymara (Netherlands
 507 Interdisciplinary Demographic Institute (NIDI) & KNAW/University of Gronin-
 508 gen), Haley Daarstad (UC Davis), Danh Chi Dao (Queen's University), Daniel Dao
 509 (Oxford Sustainable Finance Group, University of Oxford), Marco David Schmandt
 510 (TU Berlin), Astrid de Linde (University of Oslo), Lucas De Melo (University of
 511 Nottingham, NICEP), Lachlan Deer (University of Melbourne), Alexandra de Gen-
 512 dre (The University of Melbourne), Micole De Vera (Banco de España), Velichka
 513 Dimitrova (UCL SRI), Jan Fabian Dollbaum (European University Institute), Jan
 514 Matti Dollbaum (University of Fribourg and LMU Munich), Michael Donnelly
 515 (University of Toronto), Luu Duc Toan Huynh (Queen Mary University of Lon-
 516 don), Tsvetomira Dumbalska (University of Oxford), Jamie Duncan (University of
 517 Toronto), Kiet Tuan Duong (University of York), Yannick Dupraz (CNRS, Univer-
 518 sité Paris-Dauphine, PSL Research University, IRD, UMR LEDa, DIAL), Thibaut
 519 Duprey (Bank of Canada), Christoph Dworschak (German Institute for Develop-
 520 ment Evaluation & University of York), Sigmund Ellingsrud (BI Norwegian Business

521 School), Ali Elminejad (Nazarbayev University), Yasmine Eissa (The American Uni-
 522 versity in Cairo), Andrea Erhart (University of Innsbruck), Giulian Etingin-Frati
 523 (ETH Zurich), Elaheh Fatemipour (University of Warwick), Alexa Federice (UC
 524 Davis), Jan Feld (Victoria University of Wellington), Guidon Fenig (University of
 525 Ottawa), Lenka Fiala (University of Ottawa, Tilburg University and Institute for
 526 Replication), Mojtaba Firouzjaeiangalougah (Masaryk University), Erlend Fleisje
 527 (Oslo Economics), Alexandre Fortier-Chouinard (Université Laval), Julia Francesca
 528 Engel (Kiel University), Nadjim Fréchet (University of Montreal), Reid Fortier
 529 (VisualAIM), Tilman Fries (LMU Munich), Michael James Frith (University of
 530 Edinburgh), Jacopo Gabani (World Bank & Centre for health economics, univer-
 531 sity of York), Thomas Galipeau (University of Toronto), Sebastián Gallegos (UAI
 532 Business School), Areez Gangji (Independent Researcher), Xiaoying Gao (Uni-
 533 versity of York), Cloé Garnache (Oslo Metropolitan University), Attila Gáspár
 534 (ELTE KRTK), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH
 535 Norwegian School of Economics), Arijit Ghosh (RWI - Leibniz Institute for Eco-
 536 nomic Research), Garreth Gibney (University of Galway), Grant Gibson (Canadian
 537 Research Data Centre Network and McMaster University), Geir Godager (Univer-
 538 sity of Oslo), Leonard Goff (University of Calgary), Da Gong (State University
 539 of New York, Geneseo), Javier González (Southern Methodist University), Jeremy
 540 D. Gretton (Ontario Public Service's Behavioural Insights Unit), Cristina Griffa
 541 (University of Nottingham), Idaliya Grigoryeva (UC San Diego), Maja Grøtting
 542 (The Norwegian Institute of Public Health), Eric Guntermann (UC Berkeley), Jiaqi
 543 Guo (University of Birmingham), Alexi Gugushvili (University of Oslo), Hooman
 544 Habibnia (WU Vienna University of Economics and Business), Sonja Häffner (Peace
 545 Research Institute Oslo), Jonathan D. Hall (University of Alabama), Olle Hammar
 546 (Linnaeus University and Institute for Futures Studies), Amund Hanson Kordt (Uni-
 547 versity of Oslo), Barry Hashimoto (Independent), Jonathan S. Hartley (Stanford
 548 University), Carina I. Hausladen (ETH Zurich, work conducted while at California
 549 Institute of Technology), Tomáš Havránek (Institute of Economic Studies, Fac-
 550 ulty of Social Sciences, Charles University), Harry He (University of California,
 551 San Diego), Matthew Hepplewhite (University of Oxford), Mario Herrera-Rodriguez
 552 (CREST-Ecole polytechnique, IP Paris), Felix Heuer (RWI – Leibniz Institute for
 553 Economic Research), Anthony Heyes (University of Birmingham), Anson T. Y.
 554 Ho (Toronto Metropolitan University), Jonathan Holmes (University of Ottawa),
 555 Armando Holzknecht (University of Innsbruck), Yu-Hsiang Dexter Hsu (Univer-
 556 sity of California, Davis), Shiang-Hung Hu (California Institute of Technology),
 557 Yu-Shiuan Huang (National Chengchi University), Mathias Huebener (Federal Insti-
 558 tute for Population Research (BiB)), Christoph Huber (Aalto University), Kim P.
 559 Huynh (Indiana University), Zuzana Irsova (Institute of Economic Studies, Faculty
 560 of Social Sciences, Charles University, and Anglo-American University, Prague),
 561 Ozan Isler (The University of Queensland), Niklas Jakobsson (Karlstad University
 562 & FBK-IRVAPP), Raphaël Jananji (Université de Montréal), Tharaka A. Jayalath
 563 (University of Saskatchewan), Michael Jetter (University of Western Australia),
 564 Jenny John (University of Ottawa), Rachel Joy Forshaw (Heriot-Watt University),
 565 Felipe Juan (Howard University), Valon Kadriu (University of Kassel and INCHER),

566 Sunny Karim (Carleton University), Edmund Kelly (University of Oxford), Duy
 567 Khanh Hoang Dang (King's College London), Tazia Khushboo (University of Cal-
 568 gary), Jin Kim (Chinese University of Hong Kong), Gustav Kjellsson (Centre for
 569 Health Governance & HEPER, School of Public Health & Community Medicine,
 570 University of Gothenburg), Anders Kjelsrud (Oslo Metropolitan University), Jori
 571 Korpershoek (Erasmus University Rotterdam), Andreas Kotsadam (Ragnar Frisch
 572 Centre for Economic Research), Lewis Krashinsky (Princeton University), Suran-
 573 jana Kundu (Indian Institute of Technology Delhi), Alexander Kustov (University
 574 of North Carolina at Charlotte), Nurlan Lalayev (University of Warwick), Audrée
 575 Langlois (Université Laval), Jill Laufer (UC Davis), Blake Lee-Whiting (Univer-
 576 sity of Toronto), Andreas Leibing (Dresden University of Technology), Gabriel Lenz
 577 (UC Berkeley), Joel Levin (UC San Diego), Peng Li (University of Bath), Tongzhe
 578 Li (University of Guelph), Yuchen Lin (University of Warwick), Goncalo Lima
 579 (European University Institute and University of Bologna), Ariel Listo (University
 580 of Maryland), Dan Liu (Australian National University), Xuewen Lu (University
 581 of Calgary), Elvina Lukmanova (New Economic School), Alex Luscombe (Univer-
 582 sity of Toronto), Lester R. Lusher (University of Pittsburgh), Ke Lyu (University
 583 of Nevada, Reno), Hai Ma (McGill University), Nicolas Mäder (Knauss School of
 584 Business, University of San Diego), Clifton Makate (Norwegian University of Life
 585 Sciences and Norwegian Geotechnical Institute), Alice Malmberg (UC Davis), Adit
 586 Maitra (The University of Melbourne), Marco Mandas (University of Cagliari), Jan
 587 Marcus (Freie Universität Berlin), Shushanik Margaryan (University of Potsdam),
 588 Lili Márk (Central European University), Diego Marino Fages (Durham Univer-
 589 sity), Andres Martignano (University of Nottingham), Abigail Marsh (Finance
 590 Canada), Isabella Masetto (London School of Economics and Political Science),
 591 Anthony McCanny (University of Toronto), Emma McManus (Health Organisation,
 592 Policy and Economics, The University of Manchester), Ryan McWay (University
 593 of Minnesota), Lennard Metson (London School of Economics), Fabian Mierisch
 594 (Independent Researcher), Jonas Minet Kinge (University of Oslo), Sumit Mishra
 595 (Krea University), Myra Mohnen (University of Ottawa), Jakob Möller (WU Vienna
 596 University of Economics and Business), Rosalie Montambeault (Université Laval),
 597 Sébastien Montpetit (University of Warwick), Louis-Philippe Morin (University
 598 of Ottawa), Todd Morris (University of Queensland), Scott Moser (University of
 599 Nottingham, School of Politics and International Relations), Fabio Motoki (Uni-
 600 versity of Texas Rio Grande Valley), Lucija Muehlenbachs (University of Calgary
 601 and Resources for the Future), Andreea Musulan (University of Montreal), Marco
 602 Musumeci (University of Padova), Munirul Nabin (Deakin University), Karim
 603 Nchare (Vanderbilt University), Florian Neubauer (RWI - Leibniz Institute for Eco-
 604 nomic Research), Quan M. P. Nguyen (University of Sussex), Tuan Nguyen (Hasselt
 605 University), Viet Nguyen-Tien (London School of Economics), Ali Niazi (Univer-
 606 sity of Calgary), Giorgi Nikolaishvili (Wake Forest University), Ardyn Nordstrom
 607 (Carleton University), Patrick Nüß (IWH Halle), Angela Odermatt (University of
 608 Oxford), Matt Olson (University of Pennsylvania Wharton), Henning Øien (Depart-
 609 ment of Health Management and Health Economics, University of Oslo), Tim

610 Ölkens (Humboldt University zu Berlin), Miquel Oliver i Vert (University of Not-
 611 tingham), Emre Oral (University of Mannheim), Christian Oswald (University of
 612 the Bundeswehr Munich), Ali Ousman (McGill University), Ömer Özak (Depart-
 613 ment of Economics, Southern Methodist University, IZA and GLO), Shubham
 614 Pandey (Indian Institute of Technology Bombay), Alexandre Pavlov (Université de
 615 Montréal), Martino Pelli (Asian Development Bank), Romeo Penheiro (University
 616 of Houston), RyuGyung Park (Government Department at William & Mary), Eva
 617 Pérez Martel (Universitat Autònoma de Barcelona), Jörg Ankel-Peters (RWI - Leib-
 618 nitz Institute for Economic Research), Tereza Petrovičová (UCSD), Linh Phan (UC
 619 Davis), Alexa Prettyman (Towson University), Jakub Procházka (Masaryk Univer-
 620 sity), Aqila Putri (University of Maryland), Julian Quandt (WU Vienna University
 621 of Economics and Business), Kangyu Qiu (University of Calgary), Loan Quynh Thi
 622 Nguyen (National Economics University), Andaleeb Rahman (Cornell University),
 623 Carson H. Rea (Emory University), Adam Reiremo (Norwegian School of Eco-
 624 nomics), Laëtitia Renée (Université de Montréal), Joseph Richardson (Lancaster
 625 University), Nicholas Rivers (University of Ottawa), Bruno Rodrigues (Ministry
 626 of Research and Higher Education, Luxembourg), William Roelofs (University of
 627 Toronto), Tobias Roemer (University of Oxford), Ole Rogeberg (Ragnar Frisch Cen-
 628 tre for Economic Research), Julian Rose (RWI - Leibniz Institute for Economic
 629 Research), Andrew Roskos-Ewoldsen (UC Davis), Paul Rosmer (Humboldt Univer-
 630 sity of Berlin & Berlin School of Economics), Barbara Sabada (Bank of Canada),
 631 Soodeh Saberian (University of Manitoba), Nicolas Salamanca (The University of
 632 Melbourne), Georg Sator (University of Nottingham), Daniel Scates (UC Davis),
 633 Elmar Schlüter (Justus Liebig University, Giessen), Cameron Sells (Indepenent
 634 Researcher), Sharmi Sen (Monash University), Ritika Sethi (University of Chicago),
 635 Anna Shcherbiak (WU Vienna University of Economics and Business), Moyosore
 636 Sogaolu (Rotman, University of Toronto), Matt Soosalu (Carleton University), Erik
 637 Ø. Sørensen (NHH Norwegian School of Economics), Manali Sovani (Tufts Univer-
 638 sity), Noah Spencer (University of Toronto), Stefan Staubli (University of Calgary),
 639 Renske Stans (Erasmus University Rotterdam), Anya Stewart (UC Davis), Felix
 640 Stips (Institute for Employment Research (IAB)), Kieran Stockley (University of
 641 Nottingham), Stephenson Strobel (McMaster University), Ethan Struby (Carleton
 642 College, Boston College, and Minnesota Supercomputing Institute), John Tang
 643 (Utrecht University), Idil Tanrisever (University of California, Irvine), Thomas Tao
 644 Yang (Australian National University), Ipek Tastan (University of Calgary), Dejan
 645 Tatić (WU Vienna University of Economics and Business), Benjamin Tatlow (Uni-
 646 versity of Nottingham), Féraud Tchuisseu Seuyong (Université de Montréal), Rémi
 647 Thériault (New York University), Vincent Thivierge (University of Ottawa), Wenjie
 648 Tian (University of Ottawa), Filip-Mihai Toma (Bucharest University of Economic
 649 Studies), Maddalena Totarelli (Ifo Institute & Ludwig Maximilian University of
 650 Munich), Van-Anh Tran (Monash University), Hung Truong (University of Ottawa),
 651 Nikita Tsoy (INSAIT, Sofia University), Kerem Tuzcuoglu (Bank of Canada), Diego
 652 Ubfal (World Bank), Laura Villalobos (Salisbury University), Julian Walterskirchen
 653 (University of Gothenburg), Joseph Tao-yi Wang (National Taiwan University),
 654 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 (University of Vienna and 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 (McMaster University), 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 Autònoma 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)

References

- [1] Vazire, S.: Quality Uncertainty Erodes Trust in Science. *Collabra: Psychology* **3**(1), 1 (2017)
- [2] Donoho, D.L., Maleki, A., Rahman, I.U., Shahram, M., Stodden, V.: Reproducible research in computational harmonic analysis. *Computing in Science & Engineering* **11**(1), 8–18 (2008)
- [3] King, G.: Replication, Replication. *PS: Political Science & Politics* **28**(3), 444–452 (1995)
- [4] Goodman, S.N., Fanelli, D., Ioannidis, J.P.: What does research reproducibility mean? *Science Translational Medicine* **8**(341), 341–1234112 (2016)
- [5] Marcoci, A., Wilkinson, D.P., Vercammen, A., Wintle, B.C., Abatayo, A.L., Baskin, E., Berkman, H., Buchanan, E.M., Capitán, S., Capitán, T., *et al.*: Predicting the replicability of social and behavioural science claims in covid-19 preprints. *Nature human behaviour* **9**(2), 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**(3), 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**(2), 28 (2007)
- [8] National Academies of Sciences, Engineering, and Medicine: Reproducibility and Replicability in Science. National Academies Press, ??? (2019). <https://doi.org/10.17226/25303> . <https://www.nap.edu/catalog/25303>
- [9] Peterson, D., Panofsky, A.: Self-Correction in Science: The Diagnostic and Integrative Motives for Replication. Social Studies of Science **51**(4), 583–605 (2021)
- [10] Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., Debonnel, E.: Certify reproducibility with confidential data. Science **365**(6449), 127–128 (2019)
- [11] Brandon, A., List, J.A.: Markets for Replication. Proceedings of the National Academy of Sciences **112**(50), 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**(7693), 417–9 (2018)
- [14] Maniadis, Z., Tufano, F.: The Research Reproducibility Crisis and Economics of Science. Economic Journal **127**(605) (2017)
- [15] Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.: A Manifesto for Reproducible Science. Nature Human Behaviour **1**(1), 1–9 (2017)
- [16] Nosek, B.A., Hardwicke, T.E., Moshontz, H., Allard, A., Corker, K.S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M.B., *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**(3), 1191–1226 (2023)
- [18] Brodeur, A., Cook, N., Neisser, C.: P-Hacking, Data Type and Data-Sharing Policy. Economic Journal **134**(659), 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**(1), 185–206

- (2022)
- [20] Christensen, G., Miguel, E.: Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature* **56**(3), 920–80 (2018)
- [21] Dafoe, A.: Science Deserves Better: the Imperative to Share Complete Replication Files. *PS: Political Science & Politics* **47**(1), 60–66 (2014)
- [22] McCullough, B.D., McGeary, K.A., Harrison, T.D.: Do Economics Journal Archives Promote Replicable Research? *Canadian Journal of Economics* **41**(4), 1406–1420 (2008)
- [23] Pérignon, C., Akmansoy, O., Hurlin, C., Dreber, A., Holzmeister, F., Huber, J., et al.: Computational Reproducibility in Finance: Evidence from 1,000 Tests. HEC Paris Paper (2023)
- [24] Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al.: Evaluating Replicability of Laboratory Experiments in Economics. *Science* **351**(6280), 1433–1436 (2016)
- [25] Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., et al.: Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015. *Nature Human Behaviour* **2**(9), 637–644 (2018)
- [26] Open Science Collaboration: Estimating the Reproducibility of Psychological Science. *Science* **349**(6251), 4716 (2015)
- [27] Dreber, A., Johannesson, M.: A Framework for Evaluating Reproducibility and Replicability in Economics. *Economic Inquiry* (2023)
- [28] Brodeur, A., Dreber, A., Guardia, F., Miguel, E.: Replication Games: How to Make Reproducibility Research More Systematic. *Nature* **621**(7980), 684–686 (2023)
- [29] Simonsohn, U., Simmons, J.P., Nelson, L.D.: Specification Curve Analysis. *Nature Human Behaviour* **4**(11), 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), 1–32 (2016)
- [31] Brodeur, A., Cook, N., Heyes, A.: Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review* **110**(11), 3634–3660 (2020)
- [32] Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J.A., et al.: Variability in the

- 764 Analysis of a Single Neuroimaging Dataset by Many Teams. *Nature* **582**(7810),
765 84–88 (2020)
- 766 [33] Breznau, N., Rinke, E.M., Wuttke, A., Nguyen, H.H., Adem, M., Adriaans, J., *et*
767 *al.*: Observing Many Researchers Using the Same Data and Hypothesis Reveals a
768 Hidden Universe of Uncertainty. *Proceedings of the National Academy of Sciences*
769 **119**(44), 2203150119 (2022)
- 770 [34] Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J.R., Burli,
771 P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., *et al.*: The Influence of Hidden
772 Researcher Decisions in Applied Microeconomics. *Economic Inquiry* **59**(3), 944–
773 960 (2021)
- 774 [35] Menkveld, A.J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M.,
775 Kirchler, M., Neusüss, S., *et al.*: Non-Standard Errors. *Journal of Finance*
776 (Forthcoming)
- 777 [36] Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E.,
778 Bahník, Š., *et al.*: Many Analysts, One Data Set: Making Transparent How Vari-
779 ations in Analytic Choices Affect Results. *Advances in Methods and Practices in*
780 *Psychological Science* **1**(3), 337–356 (2018)
- 781 [37] Fišar, M., Greiner, B., Huber, C., Katok, E., Ozkes, A.I., Collaboration, M.S.R.:
782 Reproducibility in Management Science. *Management Science* (2023)
- 783 [38] Ankel-Peters, J., Fiala, N., Neubauer, F.: Do Economists Replicate? *Journal of*
784 *Economic Behavior & Organization* **212**, 219–232 (2023)
- 785 [39] Vilhuber, L., Turrilo, J., Welch, K.: Report by the AEA Data Editor. *AEA Papers*
786 *and Proceedings* **110**, 764–75 (2020) <https://doi.org/10.1257/pandp.110.764>
- 787 [40] Brodeur, A., *et al.*: Mass Reproducibility and Replicability: A New Hope. I4R
788 Discussion Paper 107 (2024)
- 789 [41] Clark, C.J., Tetlock, P.E.: Adversarial collaboration: The next science reform. In:
790 *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions*, pp.
791 905–927. Springer, ??? (2023)
- 792 [42] Brodeur, A., Sung, S.Y., Miguel, E., Vilhuber, L., Guardia, F.H.: Assessing Repro-
793 ducibility in Economics Using Standardized Crowd-sourced Analysis. NBER
794 Working Paper 33753 (2024)
- 795 [43] Wood, B.D., Müller, R., Brown, A.N.: Push Button Replication: Is Impact Eval-
796 uation Evidence for International Development Verifiable? *PloS one* **13**(12),
797 0209416 (2018)
- 798 [44] Gerber, A.S., Malhotra, N.: Publication Bias in Empirical Sociological Research:

- 799 Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods*
800 & Research **37**(1), 3–30 (2008)
- 801 [45] Andrews, I., Kasy, M.: Identification of and Correction for Publication Bias.
802 *American Economic Review* **109**(8), 2766–94 (2019)
- 803 [46] Elliott, G., Kudrin, N., Wüthrich, K.: Detecting p-Hacking. *Econometrica* **90**(2),
804 887–906 (2022)

805 12 Methods

806 Our focus is on 12 journals. The journals are the following for economics: *Amer-*
807 *ican Economic Review*, *American Economic Review: Insights*, *American Economic*
808 *Journal: Applied Economics*, *American Economic Journal: Economic Policy*, *Ameri-*
809 *can Economic Journal: Macroeconomics*, *The Economic Journal*, *Journal of Political*
810 *Economy*, *Quarterly Journal of Economics*, *Review of Economic Studies*. For political
811 science, the journals are: *American Journal of Political Science*, *American Political*
812 *Science Review*, and *Journal of Politics*.

813 We have two streams to generate reproductions.

814 *I4R's Board.*

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

819 *Replication games.*

820 Our second stream to generate reproductions and replications is the replication games
821 (Games). Games are one-day meet-ups open to faculty, post-docs, graduate students
822 and other researchers. Participants join a small team of about 3–5 researchers all
823 working in the same subfield (*e.g.*, development economics).

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

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

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

835 Teams are asked to develop a game plan for the Games; each team member should
836 know what they are supposed to do during the Games. A virtual meeting with the
837 organizers before the Games allows each team to ask questions and discuss a game

838 plan. During the games, A.B., D.M. or one of I4R's co-directors, provide live assistance
839 to each team. Teams then have to write a (templated - <https://osf.io/8dkxc/>) report
840 summarizing their work and results in the following months. Of note, virtually all
841 teams kept working on their reproduction after the Games and some even started the
842 re-analysis prior to the Games.

843 Participants are offered the possibility to virtually attend Games. In our sample
844 of completed reports, about 68% of participants attended the games in-person, while
845 32% virtually attended the events. Most teams are fully virtual or in-person, with only
846 a small share of teams having a mix of virtual and in-person participants. Mixed teams
847 are typically due to a variety of reasons (*e.g.*, canceled flight for one participant), or
848 late registrations.

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

856 12.1 Reports

857 Teams have on average worked 13 active days on their reproduction (std. dev. of 24).
858 Supplementary Materials Appendix Figure 7 shows the distribution of days across
859 reports, trimmed at over 100 days.

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

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

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

880 In practice, some teams in both streams did not write a pre-re-analysis plan and
881 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.

12.2 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. We provide examples for each group in what follows.

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.

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

930 12.3 Robustness for Figures

931 While the bulk of our analysis compares coefficients and statistical significance from
 932 the original study and the work of reproducers, many results in papers are also dis-
 933 played in figures. For those which are plots of coefficients (i.e., event studies) we
 934 encouraged reproducers to give the underlying statistics used to create the graph. This
 935 was often at the discretion of the reproducers: it could be taxing to write new code
 936 to compare and extract those values. In one example, the underlying programs which
 937 were written by the original authors were too complicated to modify with robustness
 938 checks. Excepting anecdotal examples, many teams found it feasible to reproduce a
 939 figure as part of a robustness check or direct replication. In those circumstances, we
 940 (A.B. and D.M.) tried to subjectively describe if we believed the results were the same.
 941 This was usually taken with the discussion of the reproducers and reading the original
 942 paper. We find that 189 out of 263 figures—71.9 percent—we believe to have display
 943 the same result as the original paper and can be reasonably compared.

944 12.4 Non Comparable Re-Analyses

945 As mentioned earlier, a direct comparison is not possible between the original analysis
 946 and the reproducers' analysis for about 15% of re-analyses. In applied microeconomics
 947 and politics papers, this may be due to a change in the estimator or a change in
 948 the scale of the dependent or main independent variable. There are also scenarios
 949 where the original paper uses methods where coefficient estimates and p-values are
 950 not the objective of the analysis. This is apparent in a few empirical macroeconomics
 951 papers teams looked at. A common "robustness check" would be to adjust parameters
 952 which enter a model, possibly using accepted values in the field or estimated from an
 953 alternative dataset.

954 82 articles have at least one non-comparable estimate. Only a small proportion
 955 (10 re-analyses) were not directly comparable for all reported re-analysis estimates.
 956 For not directly comparable re-analyses, we report the proportion that reproducers
 957 indicated were of the same statistical significance as the original and same sign. For our
 958 four definitions of reproducibility and replication rates these are: When the original
 959 estimate is statistically significant at the 5% level, 85% of those we considered not
 960 directly comparable indicated their re-analysis was of the same significance (93% for
 961 the 10% level). When the original estimate was not statistically significant at the 5%
 962 level, 88% of those we considered not directly comparable indicated their re-analysis
 963 was of the same (non)significance (92% for the 10% level).

964 12.5 Types of Re-Analyses

965 One of our main objectives is to document the relative importance of several robustness
 966 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 described in Appendix 12.2 provides a description and examples for each group.

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.

12.6 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 [40] (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 [41]. 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.

12.7 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.

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. Supplementary Materials Appendix Figure 8 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 [42]. 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.

12.8 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.

1092 12.9 Many-Analysts Approach

1093 Our approach and research questions, which we detail below, were pre-registered. Our
 1094 pre-analysis plan was pre-registered here: <https://osf.io/8wsqx/>. The pre-analysis plan
 1095 was pre-registered prior to sharing the Meta Database with analysts.

1096 The six analyst teams tackled the following eight questions:

- 1097 1. “Does reproducibility/replicability rate depend on replicators’ experience coding?”
- 1098 2. “Does reproducibility/replicability rate depend on replicators’ academic experi-
 1099 ence?”
- 1100 3. “Does reproducibility/replicability rate depend on the authors’ experience?”
- 1101 4. “Does reproducibility/replicability rate depend on the interaction of the authors’
 1102 experience and replicators’ experience?” In particular:
 - 1103 (a) Are reproducibility/replicability rate higher when authors’ experience is
 1104 high, and replicators’ experience is low (in comparison to similar levels)?
 - 1105 (b) Are reproducibility/replicability rate higher when authors’ experience
 1106 and replicators’ experience is similar (in comparison to dissimilar
 1107 levels)?
 - 1108 (c) Are reproducibility/replicability rate higher when authors’ experience is
 1109 low, and replicators’ experience is high (in comparison to similar levels)?
- 1110 5. “Does reproducibility/replicability rate depend on the interaction of the authors’
 1111 prestige and replicators’ prestige?” In particular:
 - 1112 (a) Are reproducibility/replicability rate higher when authors’ have high
 1113 prestige, and replicators’ experience have low prestige (in comparison
 1114 to similar levels)?
 - 1115 (b) Are reproducibility/replicability rate higher when authors’ and replica-
 1116 tors’ prestige is similar (in comparison to dissimilar levels)?
 - 1117 (c) Are reproducibility/replicability rate higher when authors’ have low
 1118 prestige, and replicators’ experience have high prestige (in comparison
 1119 to similar levels)?
- 1120 6. “Does reproducibility/replicability rate depend on the original authors providing
 1121 raw data?”
- 1122 7. “Does reproducibility/replicability rate depend on the original authors providing
 1123 raw or intermediate data?”
- 1124 8. “Does reproducibility/replicability rate depend on the original authors providing
 1125 cleaning code?”

1126 12.9.1 Data for Analysts

1127 Analysts were not given access to raw data (database, team leader surveys, individual
 1128 surveys). Rather, they were given access to intermediate/analytical data which was
 1129 cleaned and merged in a manner which would be consistent for their analysis. Giving
 1130 researchers a downstream dataset allowed A.B. and D.M. to make restrictions on
 1131 what the analysts could do. The clearest example of this would be defining dependent
 1132 variables which were not allowed to be changed - providing a consistent definition
 1133 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.

12.9.2 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.

1177 12.9.3 Many Analysts: Additional Results

1178 Each row in Supplementary Materials Appendix Table 6 represents one of the
 1179 eight research questions. The four columns represent four broad categories regarding
 1180 research teams' coefficient estimate(s) to the research question: (1) negative and sta-
 1181 tistically significant, (2) negative and not-statistically significant, (3) positive and not
 1182 statistically significant and (4) positive and statistically significant. The left-to-right
 1183 order of the column categories corresponds to where the associated analyst t-statistic
 1184 would fall on the real number line. While the dependent variable (which does not
 1185 change in this table) is the same for each team, each team chooses their own primary
 1186 independent variable. Each cell represents the proportion of analyst-estimated rela-
 1187 tionships by category. The cells are team-weighted so that if a many-analyst team
 1188 presents three estimates and another team presents a single estimate, the first team's
 1189 estimates enter the proportion as 1/3 each.

1190 The cell in the first row and first column tells us that 42.8% of results from the
 1191 many-analysts find a negative and statistically significant relationship between the
 1192 coding experience of a reproducer and the reproducibility rate for estimates that were
 1193 originally statistically significant at the 5% level (i.e., lower reproducibility rate for
 1194 more experienced reproducers).

1195 From the second column, it becomes clear that, if there is a relationship between
 1196 reproducers experience coding and the reproducibility rate, it seems to be almost
 1197 definitively negative with a combined proportion of 86% of results returned as negative
 1198 and statistically significant or negative and not statistically significant at the 5% level.
 1199 Only 14% of estimates find a positive relationship between the reproducers' experience
 1200 coding and the reproducibility rate - of which none of the estimated positive relation-
 1201 ships estimated were statistically significant. (The associated row in Supplementary
 1202 Materials Appendix Table 7, which looks at the replication for the 10% threshold finds
 1203 the same pattern.) This result potentially suggests that reproducers with more experi-
 1204 ence coding are better suited to detecting and correcting less-than-robust estimations
 1205 - possibly because of having greater expertise with the methods used.

1206 Supplementary Materials Appendix Table 6 presents results where the dependent
 1207 variable takes a value of one if an originally 5% statistically significant result was
 1208 reproduced also at the 5% level. Supplementary Materials Appendix Table 7 has the
 1209 same structure, but uses the 10% threshold. Supplementary Materials Appendix Table
 1210 8 then examines whether an originally *not* 5% statistically significant result was repro-
 1211 duced, while Supplementary Materials Appendix Table 9 continues this with the 10%
 1212 threshold.

1213 For the second research question - whether the reproduction robustness rate
 1214 depends on the reproducers' academic experience, a somewhat similar albeit less
 1215 starkly negative result is found with some proportion moving into the positive and sta-
 1216 tistically significant category. That said, the ratio of negative-and-significant results
 1217 to positive-and-significant results remains above 4 to 1. The associated row in Supple-
 1218 mentary Materials Appendix Table 7, which looks at robustness for the 10% threshold
 1219 finds the same pattern, although with 75% of many-analysts results being negatively
 1220 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 [18] 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.

1264 12.10 Database

1265 In what follows, we describe our database. The database is mainly built from three
 1266 sources of raw data: (1) reports; (2) surveys of individual reproducers; and (3) surveys
 1267 of teams of reproducers. We also collected information from publicly available *curricula*
 1268 *vitae* of all original authors and reproducers.

1269 12.10.1 Reports

1270 Two of the lead authors (A.B. and D.M.) and research assistants read reports and
 1271 copied test statistics into an Excel file. We also coded and grouped robustness and
 1272 replicability exercises, and information on computational reproducibility and coding
 1273 errors. The work being entered by RAs was checked by A.B. or D.M. for completeness
 1274 and accuracy. If any part of any entry was unclear, they were checked again and
 1275 discussed.

1276 Only a subset of results was considered suitable for our research. We follow the
 1277 following criteria. We exclude extensions of the original authors' research, effects by
 1278 heterogeneity, or mediation analyses. These analyses correspond to situations where
 1279 there are no "original" estimates to which we can reasonably compare the reproducers'
 1280 estimates. Most often, reproducers included tables and figures which were the output
 1281 of a computational reproduction using the original authors' replication package. These
 1282 are always left out for the re-analyses. After being checked, reproducers would then
 1283 be contacted with their subset of the database and asked to confirm our transcribing
 1284 of their reports into the database.

1285 Coding errors and discrepancies are also excluded from the re-analyses. We discuss
 1286 coding errors and discrepancies between original authors' values in their published
 1287 paper compared to what their replication package produces in Methods Section 12.13.

1288 We report some additional information in our database. We collect information
 1289 on the journal, year of publication, number of Google Scholar citations at the time of
 1290 entry into the database, the research field, the position of the test in the original article
 1291 and the number of original authors and reproducers. We also collect information from
 1292 *curricula vitae* of all the original authors and reproducers. We obtained information
 1293 on their academic affiliation at the time of publication, their position at the main
 1294 institution and year the PhD was earned. In addition, we gather for each author
 1295 and reproducer the following information (at the time of completing the replication):
 1296 the total number of Google Scholar citations and whether they had published in a
 1297 Top-5 economic journal, a leading political science journal, and/or one of the other
 1298 economic journals we are reproducing/replicating. The Top-5 economic journals are
 1299 the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the
 1300 *Quarterly Journal of Economics* and the *Review of Economic Studies*. The leading
 1301 political science journals considered here are the *American Journal of Political Science*,
 1302 *American Political Science Review* and *Journal of Politics*.

1303 12.10.2 Surveys

1304 We asked all reproducers to fill out an individual survey. We also asked one author per
 1305 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?"

12.10.3 Descriptive Statistics

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

1349 using the raw data (either provided by the authors or re-downloaded by the repro-
 1350 ducers), while other teams have focused solely on conducting robustness checks for
 1351 multiple central hypotheses. As an illustrative example, imagine that an original arti-
 1352 cle has three main outcome variables and relies on two main specifications. If the
 1353 reproducers conduct five different robustness checks for each outcome variable and
 1354 specification, then this would lead to 30 re-analyzed test statistics.

1355 As a robustness check, we deal with this issue by adjusting the weight of each test
 1356 statistics by the inverse number of such statistics in the reproduction report such that
 1357 each reproduction report has the same weight.

1358 Supplementary Materials Appendix Table 2 provides descriptive statistics. The
 1359 articles in our sample are all recently published with a relatively small number of
 1360 Google Scholar citations (44 on average) as of the completion of a reproduction report.
 1361 The original authors are more experienced than reproducers with 11 years of expe-
 1362 rience (*i.e.*, years since completing their Ph.D.) against 3. Original authors have on
 1363 average 4,269 Google Scholar citations in comparison to 478 for reproducers. Those dif-
 1364 ferences are mostly driven by the larger share of graduate students among reproducers
 1365 than for original authors (49% against 6%). There are about 2.6 original authors per
 1366 article in comparison to 3.2 for reproducers. About 15% of reproducers have recently
 1367 published in a Top 5 or one of the three leading political science journals in our sample.
 1368 Approximately 30% have published in those journals or in one of the other journals
 1369 in our sample.

1370 While reproducers have less academic experience than original authors on average,
 1371 their level of expertise as programmers is quite advanced. About 10%, 48% and 33%
 1372 of reproducers report that their level of expertise is “Expert”, “Proficient” and “Com-
 1373 petent,” respectively. Moreover, about 55% of reproducers had already produced a
 1374 replication package for their own work or journal publication.

1375 12.10.4 Replication Packages and Expectations

1376 In an assessment of reproducers’ expectations regarding the quality of replication pack-
 1377 ages, we ask reproducers the following question in the individual survey: “Which of the
 1378 following best describes how the replication package aligned with your expectations”.
 1379 We find that more than half of reproducers report that the replication package aligned
 1380 reasonably with expectations, and an additional 26% of reproducers indicated that
 1381 the replication packages exceeded their initial expectations. Fewer than 10% report
 1382 that the replication package was worse than expected, possibly indicating that for this
 1383 small proportion of reproducers, the provided materials did not meet the anticipated
 1384 quality standards or may have lacked certain elements critical for an effective repro-
 1385 duction process. Overall, we find it encouraging that most reproducers found that the
 1386 provided materials exceeded or aligned well with their initial expectations.

1387 12.11 Computational Reproducibility

1388 We first evaluate computational reproducibility in our sample. We rely on the Social
 1389 Science Reproduction Platform (SSRP)’s 10-point scale to document computational
 1390 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 Supplementary Materials Appendix Figure 5. 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, 43]). 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.

12.12 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

1478 by translating it into R or Python makes the study itself accessible to many more
1479 researchers.

1480 Recoding also helps to assess the importance of differing assumptions embedded
1481 within programming languages (e.g., different types of Random Number Generations,
1482 rounding rules and numerical precision). We categorized recoding exercises done by
1483 reproducers into three categories: (i) identical numerical results, (ii) minor differences,
1484 and (iii) major differences. Minor differences involve small numerical discrepancies
1485 between the authors' estimates and those obtained by the reproducers. Those dif-
1486 ferences do not lead to important changes in significance or magnitude. In contrast,
1487 major differences lead to major differences in one or multiple claims.

1488 Supplementary Materials Appendix Table 12 shows our results. Out of 23 recoding
1489 exercises, we find major differences for three studies and minor differences for 10
1490 studies. Two of the major differences were uncovered when using a different software
1491 and looking at the authors' code.

1492 Additionally, one team that computationally reproduced the results using a differ-
1493 ent *version* of the software used by the authors uncovered noteworthy differences in
1494 the magnitude and significance of the estimates. About half the main claims were no
1495 longer reproducible (i.e., same sign and statistically insignificant or different sign) due
1496 to a change in the defaults used by base R when generating random numbers start-
1497 ing in version 3.6.0. This is the only instance where using a different version of the
1498 software led to major differences in the size and significance of the estimates.

1499 These results suggest that most teams who recoded using a different software
1500 language or without looking at the authors' code could obtain similar or very similar
1501 results.

1502 12.13 Coding Errors and Discrepancies

1503 We now turn to documenting the prevalence of coding errors and discrepancies between
1504 the code and the published article. Of note, a paper might be fully reproducible,
1505 but the programs may contain coding errors. Similarly, there might be important
1506 discrepancies between what the article states and what the programs compute, while
1507 remaining computationally reproducible.

1508 In what follows, we do not document trivial coding errors such as versioning issues
1509 and missing packages/paths. Those coding errors are typically easily fixed by the
1510 reproducers. We instead focus on coding errors which could have had an impact on
1511 claims and conclusions of articles.

1512 We uncover minor or major coding errors in 26 of the 110 studies in our sample,
1513 with some studies containing multiple errors. The errors can be broadly categorized
1514 into errors of the dependent variable (4 articles), main independent variable (5), control
1515 variables (10), estimation (2), inference (2), sample/observations (8) and other (5).
1516 While not all coding errors lead to changes in the conclusions of the original study,
1517 we uncovered several major coding errors worth discussing. Some examples of major
1518 errors include: a very large number of duplicated observations, failing to fully interact
1519 a difference-in-differences regression specification, miscoding the treatment variable
1520 for a large number of (or all) observations, and clear model misspecification.

1521 The prevalence of coding errors is larger for economics (26%) than political science
 1522 (16%). A plausible explanation is that replication packages from economic articles
 1523 have more lines of code than those in political science, mechanically increasing the
 1524 likelihood of at least one coding error.

1525 We also uncovered transcription issues for 13 studies, typically involving small
 1526 numerical differences or rounding errors not impacting the claims or conclusions of
 1527 the article.

1528 12.14 Re-Analyses, P-Hacking, and Publication Bias

1529 12.14.1 t- and p-Curves

1530 We present both t-statistics and p-curves in Supplementary Materials Appendix Figure
 1531 9. The top left panel provides the distribution of t-statistics from the *originally* pub-
 1532 lished estimates. We restrict the visualization to $t \in [0, 5]$, present bins of width 0.1,
 1533 and present an Epanechnikov kernel (with standard errors in blue, along with renor-
 1534 malization at zero) which softens valleys and peaks. We provide reference lines at the
 1535 conventional two-tailed significance levels. Roughly 60%, 50%, and 25% of test statis-
 1536 tics are significant at the 10%, 5% and 1% levels, respectively. We note especially
 1537 that the distribution exhibits a peak (global maximum) just above the two-star sta-
 1538 tistical significance threshold of $t = 1.96$ and a valley before the one-star statistical
 1539 significance threshold between $t = 1.0$ and $t = 1.65$. We take this as our first piece of
 1540 evidence that the original studies in our sample suffer from (marginal) p-hacking and
 1541 publication bias. The bottom left panel provides the equivalent p-curve for p-values
 1542 $\in [0.0025, 0.1500]$, with bins of width 0.0025. We have removed $p < 0.0025$ (for a two-
 1543 tailed test this is roughly $t = 3$) for illustrative purposes only, as inclusion of that mass
 1544 in the left-most bar of the p-curve leads the resolution of the remaining bars to be
 1545 quite low. We note that, much like the peak after $t = 1.96$ and the valley just before,
 1546 the p-curve exhibits a too-tall bar just to the left of the $p = 0.05$ threshold. Whether
 1547 interpreted through the t-curve or p-curve, we consider this to be our first piece of
 1548 evidence that the sample of original studies suffers from some form of p-hacking and
 1549 publication bias.

1550 We present t-and-p-curves using data from [31] in the right panels to serve as a
 1551 benchmark with which to compare the original studies. The top right panel presents
 1552 the distribution of t-statistics associated with hypothesis tests from articles published
 1553 in 25 leading economics journals in 2015 and 2018. These articles rely on one of four
 1554 popular identification methods (i.e., difference-in-differences, instrumental variable,
 1555 randomized controlled trials, and regression discontinuity design). Overall, the distri-
 1556 bution from our original studies sample is similar to that in [31], although with visually
 1557 markedly more bunching around the 5% significance threshold.

1558 This could be due to at least three reasons. First, the extent of p-hacking and
 1559 publication bias might be larger in our sample. Second, reproducers might focus on the
 1560 most central claim(s) in original studies, while [31] focus on all claims. Arguably, the
 1561 central claim(s) could be more p-hacked or suffer from more publication bias. Third,
 1562 reproducers might choose to reproduce studies finding an effect or focus on replicating
 1563 claims that reject the null hypothesis.

Supplementary Materials Appendix Figure 10 directly compares the distribution of test statistics for original studies and our re-analyses. Just as in Supplementary Materials Appendix Figure 9, 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 11 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 Supplementary Materials Appendix Figure 12 reproduce the top panel of Supplementary Materials Appendix Figure 10 for economics and political science while the bottom panels of Supplementary Materials Appendix Figure 12 reproduce the bottom panel of Supplementary Materials Appendix Figure 10. 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 13 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 13).

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 14 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.

12.14.2 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 ([44]) 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:

$$Pr(\text{Significant}_{pr} = 1) = \Phi(\alpha + \lambda \text{Reanalysis}_r) \quad (1)$$

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 $[0.05 \pm 0.04]$. 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 [45]. 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 [31] and [46] as our foundation. Our initial focus is on randomization tests, as designed by [31] 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 [46], 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 15. 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.

12.15 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.

12.15.1 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)”. Supplementary Materials Appendix Figure 16 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. ([37] 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.

12.15.2 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%).

12.16 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 ten 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.

12.16.1 Results

Replication Folder Availability Supplementary Materials Appendix Figure 17 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.

Supplementary Materials Appendix Figure 18 breaks down the previous figure's sample into those journals sampled from economics (Supplementary Materials Appendix Figure 18a) and political science (Supplementary Materials Appendix

Figure 18b). 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 Supplementary Materials Appendix Figures 19a through 20c 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 (Supplementary Materials Appendix Figure 19a) and cleaning code (Supplementary Materials Appendix Figure 19c). In earlier years, replication folders were more likely to include analysis code (Supplementary Materials Appendix Figure 19b) and final data Supplementary Materials Appendix Figure 20b). 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). Supplementary Materials Appendix Figure 20c 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.

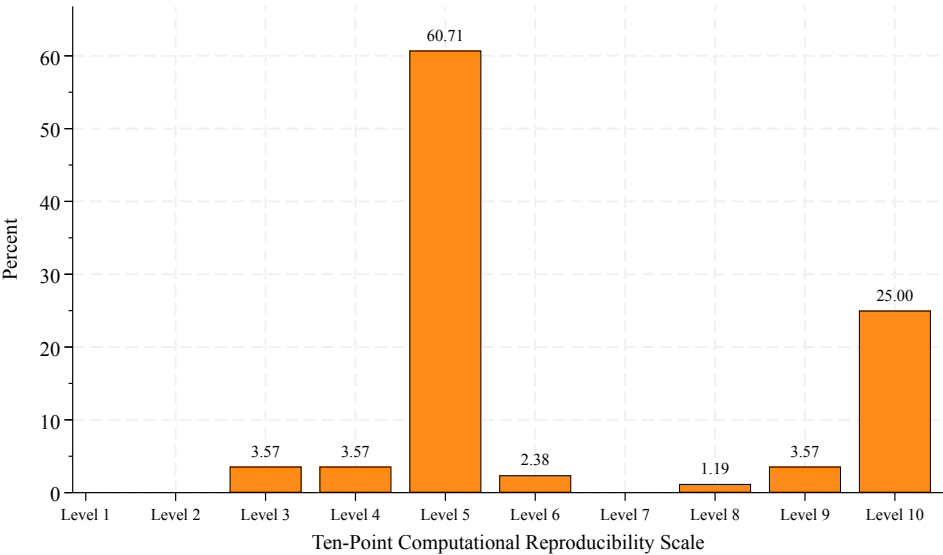
Comparison Between Journals with a Data Editor and Journals Without a Data Editor in 2023 One question raised is the importance of data editors in improving the quality of replication folders. We split the sample presented earlier in this section into those journals which did not have a data editor in 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

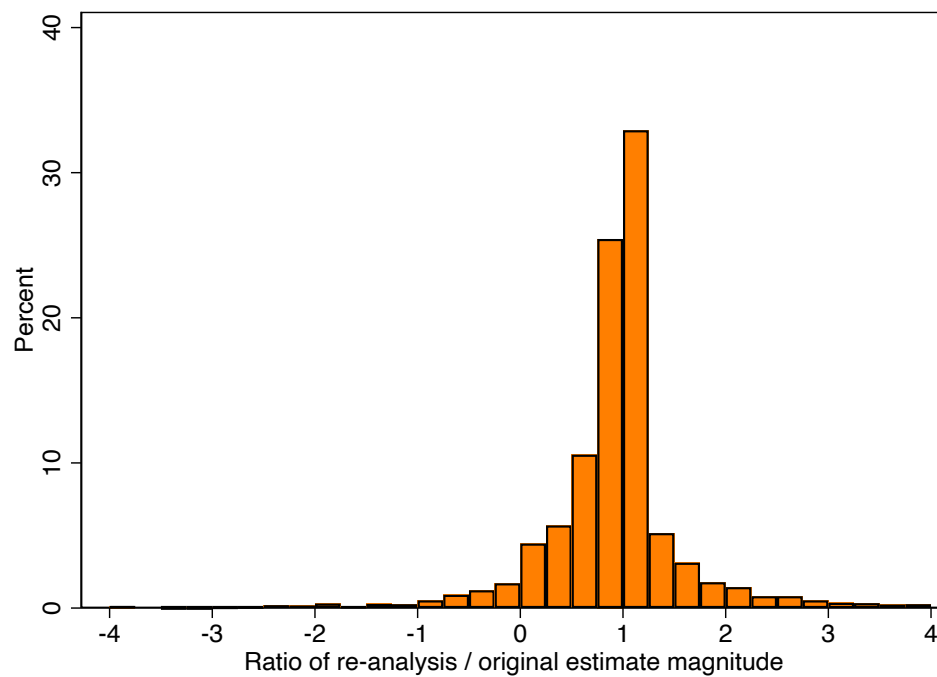
1867 believe this is likely due to raw data being sufficient for producing final data. When
1868 using our alternative definition of “final” data inclusion that takes into account raw
1869 (or intermediate data) with complete cleaning code, the difference between journals
1870 with data editors and those without data editors reduces to about -6%.

1871 **SUPPLEMENTARY MATERIALS APPENDIX**
1872 **FIGURES**

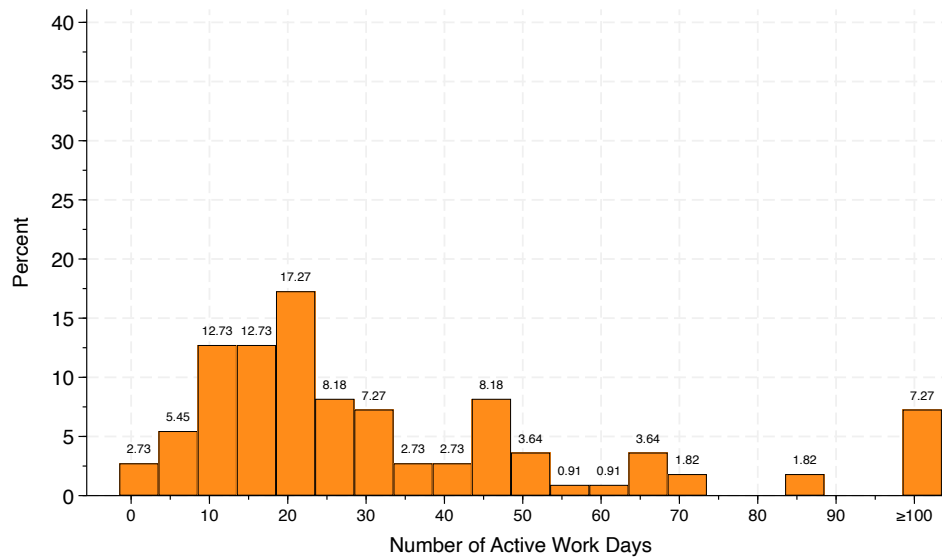
Fig. 5: 10-Point Computationally Reproducibility Score



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.

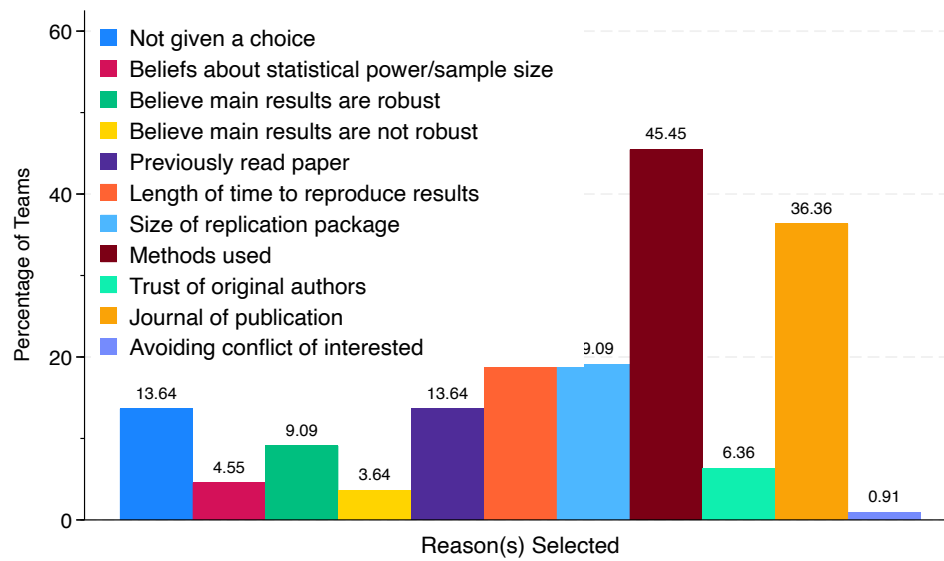
Fig. 6: Relative Reproduced Effect Size

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.

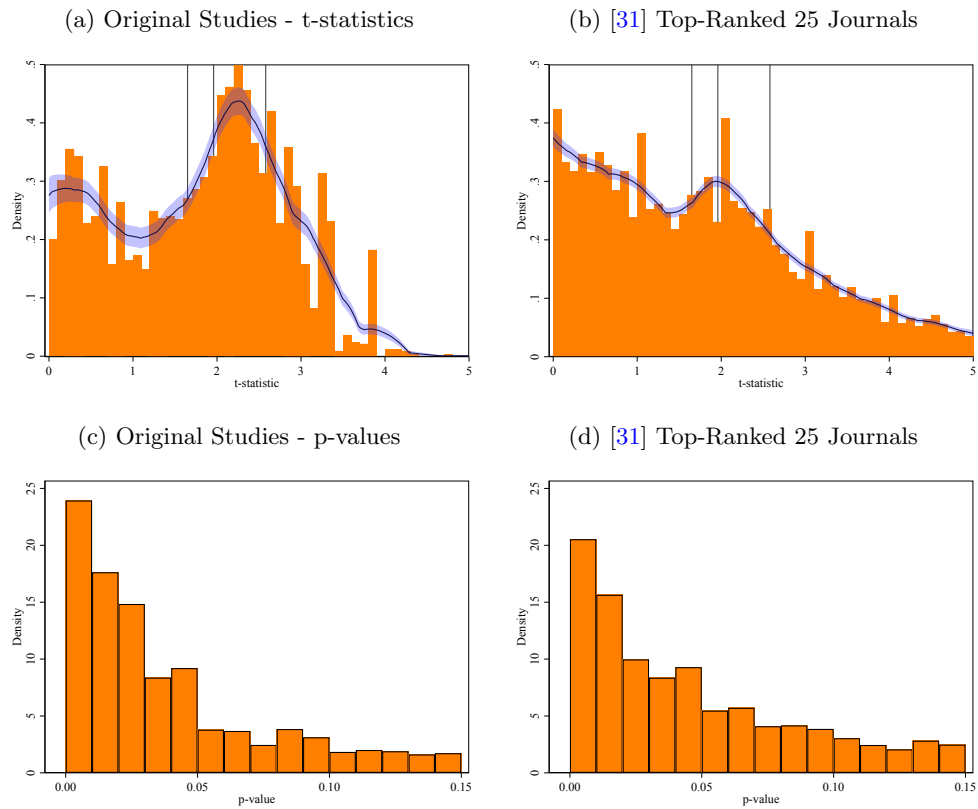
Fig. 7: Histogram of Number of Active Work Days

Notes: 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.

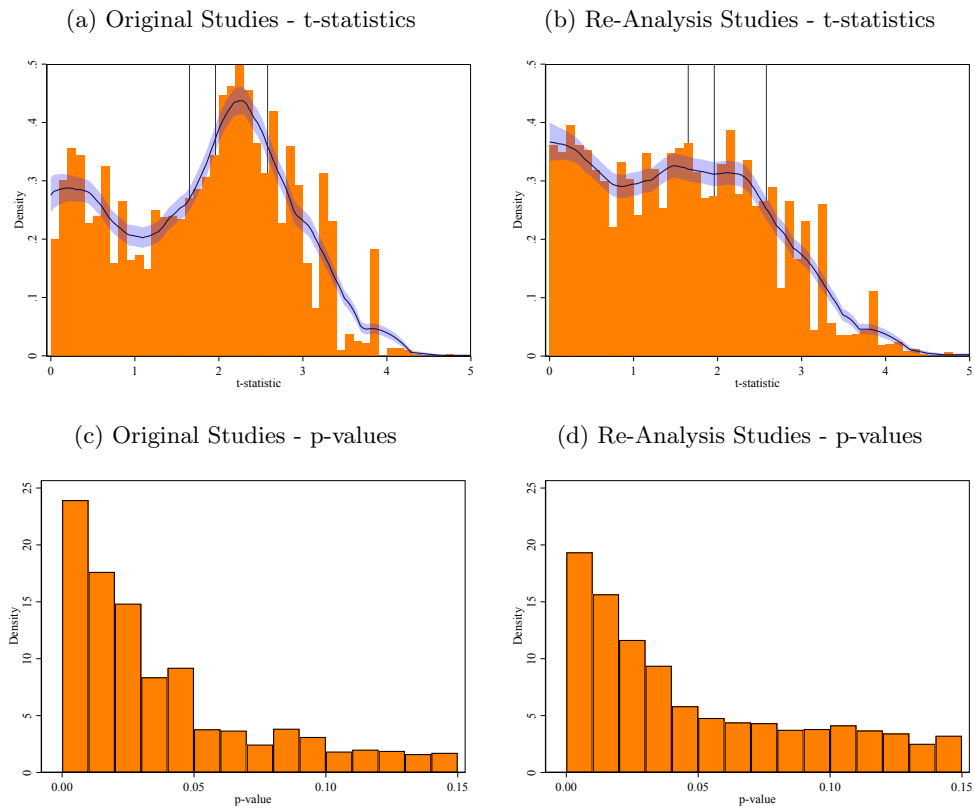
Fig. 8: For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided? (Select all which apply)



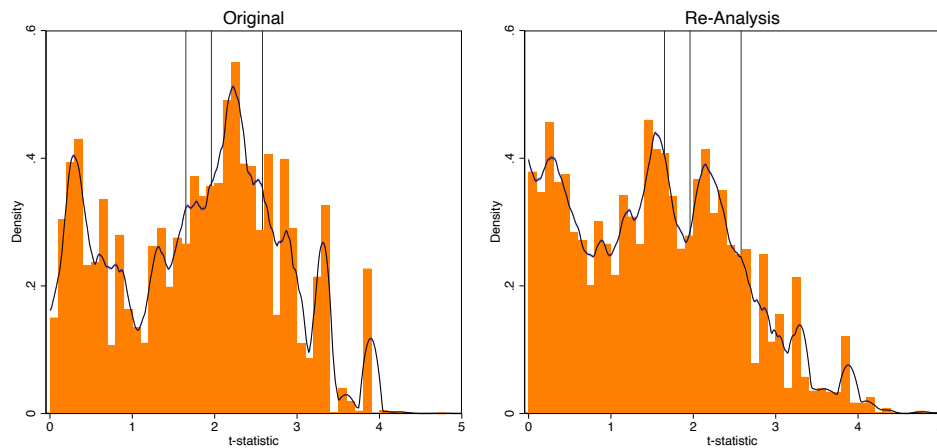
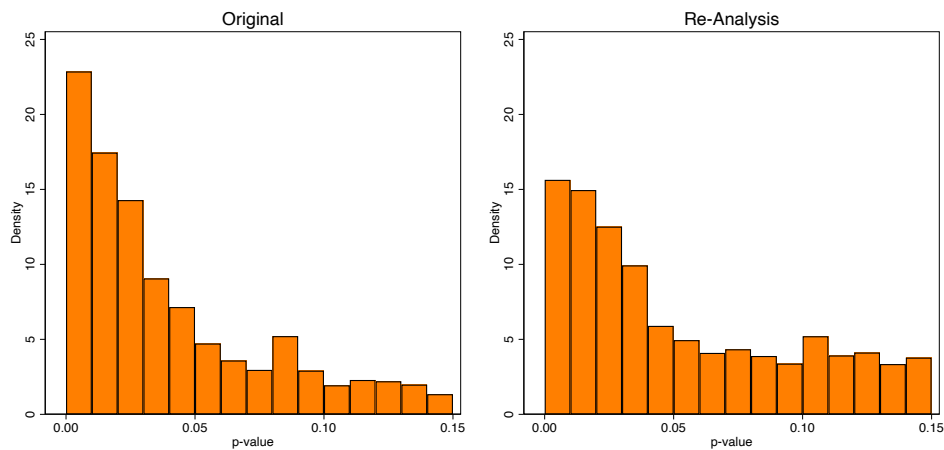
Notes: Data collected *via* survey of our reproducers after completing their reports. This figure illustrates the responses to the question: “For what reasons did you select your specific paper to reproduce and/or replicate from the list of papers provided?”

Fig. 9: Distributions of t-Statistics and p-Values for Original Studies and [31]

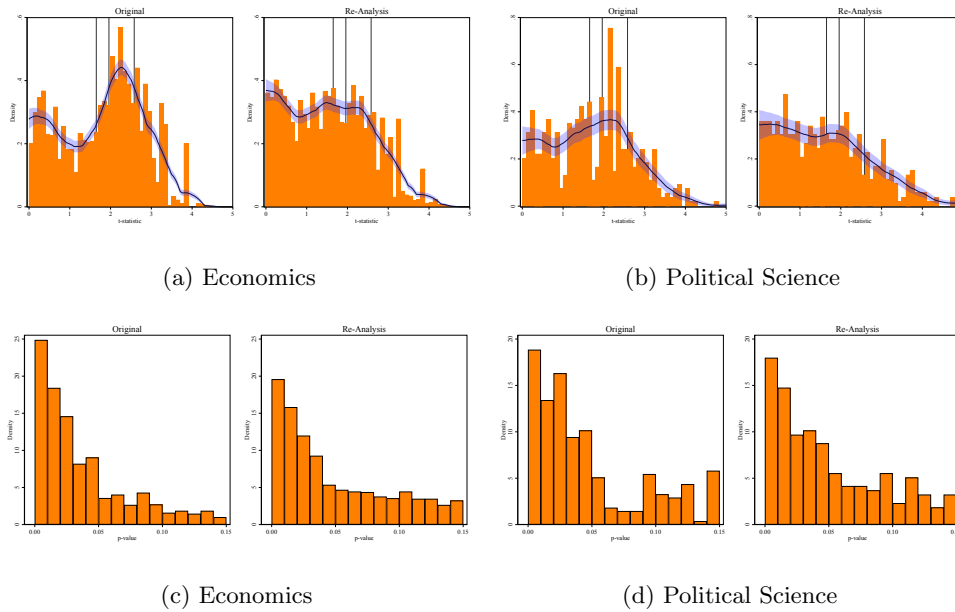
Notes: The top figures display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left figure includes all original studies in our data set. As a comparison, the top right figure plots the corresponding histogram of z-statistics from the top-ranked 25 economics journals published in 2015 and 2018 (from [31]). 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 [31], respectively.

Fig. 10: Distributions of t-Statistics for Original Studies and Re-Analyses

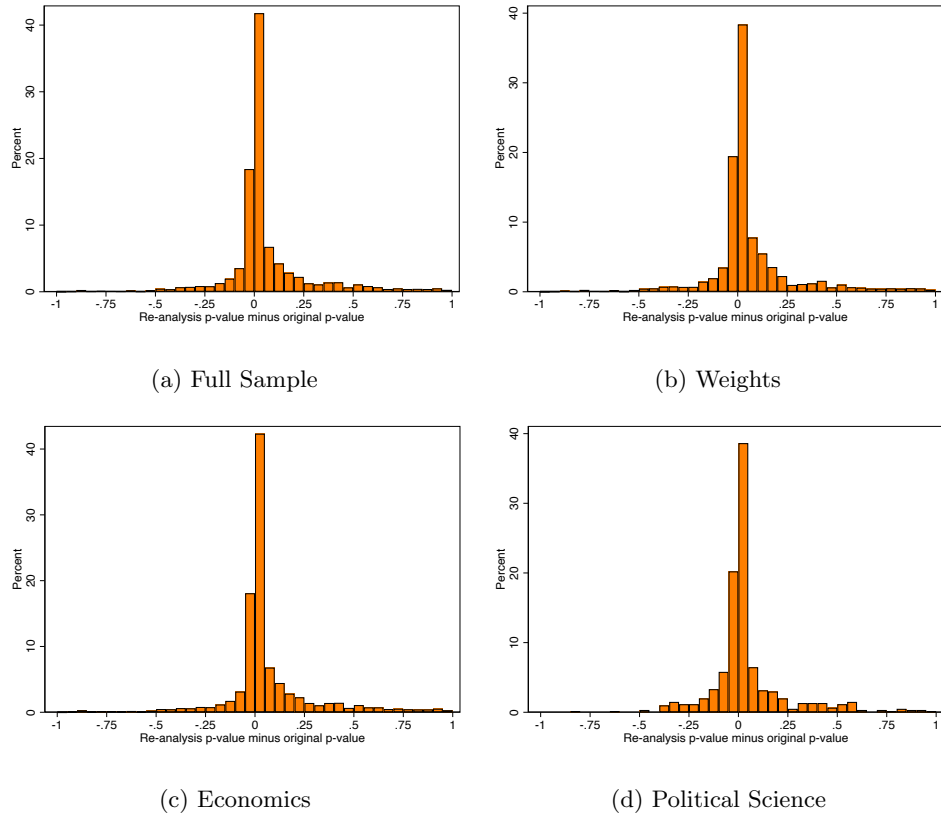
Notes: The top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left panel includes all original studies in our data set. The top right panel includes all re-analysis estimates in our data set. Vertical reference lines are displayed at conventional two-tailed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

Fig. 11: Weighted Distributions of Statistics for Original Studies and Re-Analyses(a) t -statistic(b) p -value

Notes: 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.

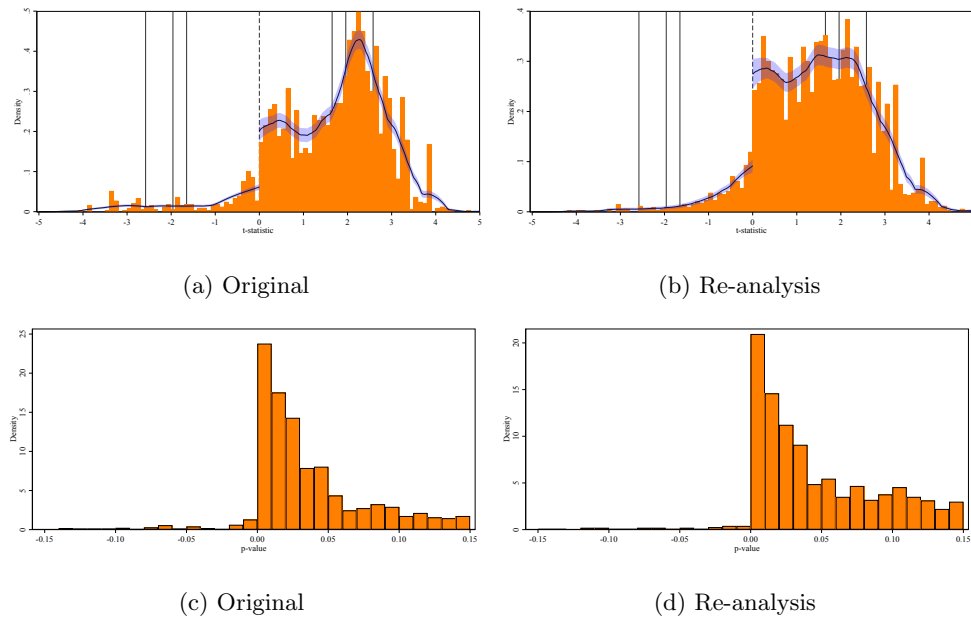
Fig. 12: Distributions of t -Statistics and p -values for Original Studies and Re-Analyses

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.

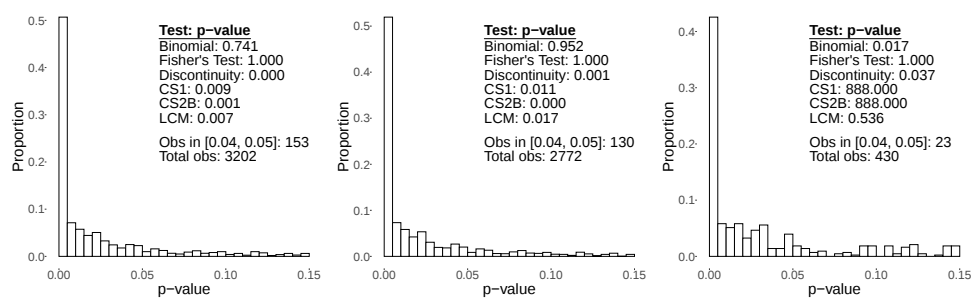
Fig. 13: Distribution of $p_{\text{re-analysis}} - p_{\text{original}}$ by Weights and Fields

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

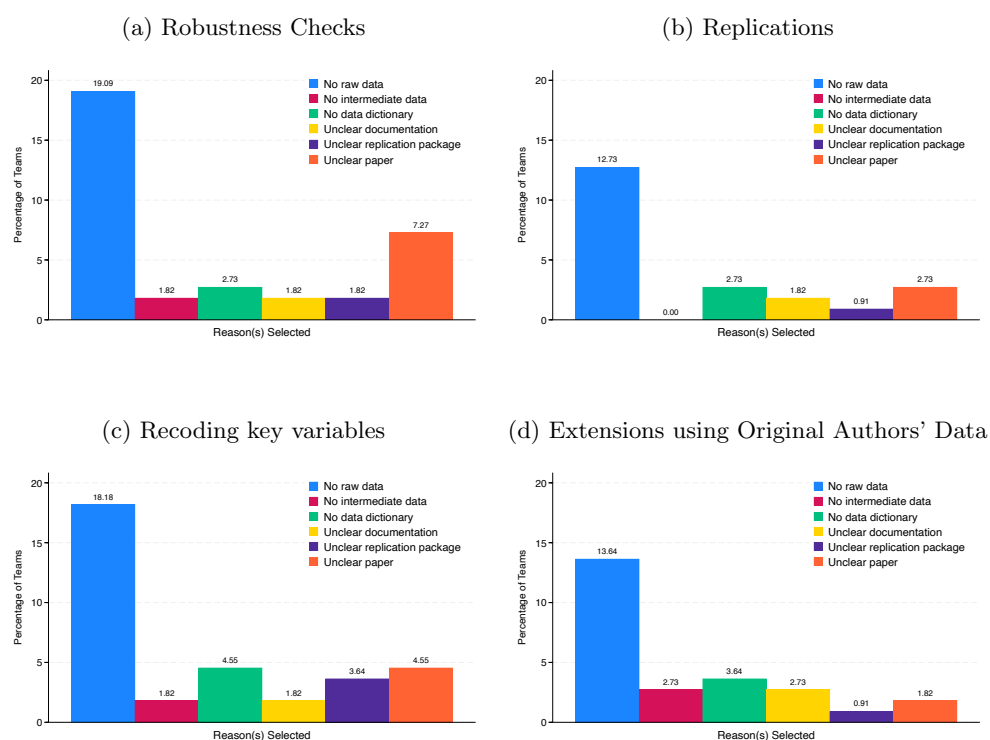


Top panels display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. We have added a dashed reference line at $t = 0$, demarcating the areas where the reproducers' and original estimates agree in sign. For both sides of the zero line, vertical reference lines are displayed at 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.

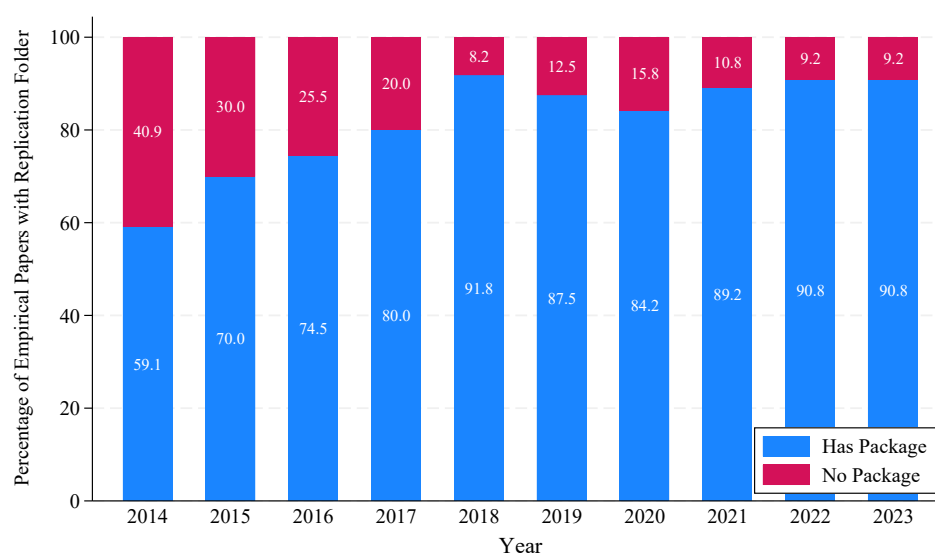
Fig. 15: Applying [46]’s Tests

Notes: This figure present p-curves and results for the battery of p-hacking tests proposed in [46] 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.

Fig. 16: 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)



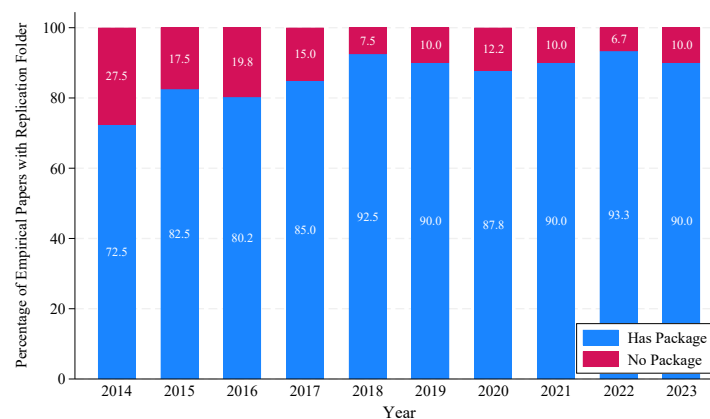
Notes: This Figure illustrates the share of teams who were unable to perform robustness checks (top-left), replications (top-right), key variable recodes (bottom-right) or extensions (bottom-left) for various reasons represented by the different coloured bars.

Fig. 17: Percentage of Papers with a Replication Folder

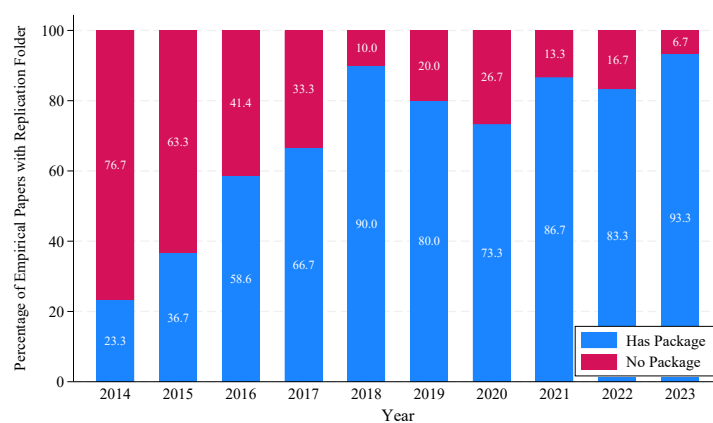
The total sample is 1150 papers with 120 papers per year from 2019 to 2023 and 110 papers per year from 2018 to 2014. Each journal has 10 papers per year except *American Economic Review: Insights* which only formally became a journal in 2019 (and are omitted in earlier years). The journals sampled over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

Fig. 18: Percentage of Papers with a Replication Folder by Discipline

(a) Economics

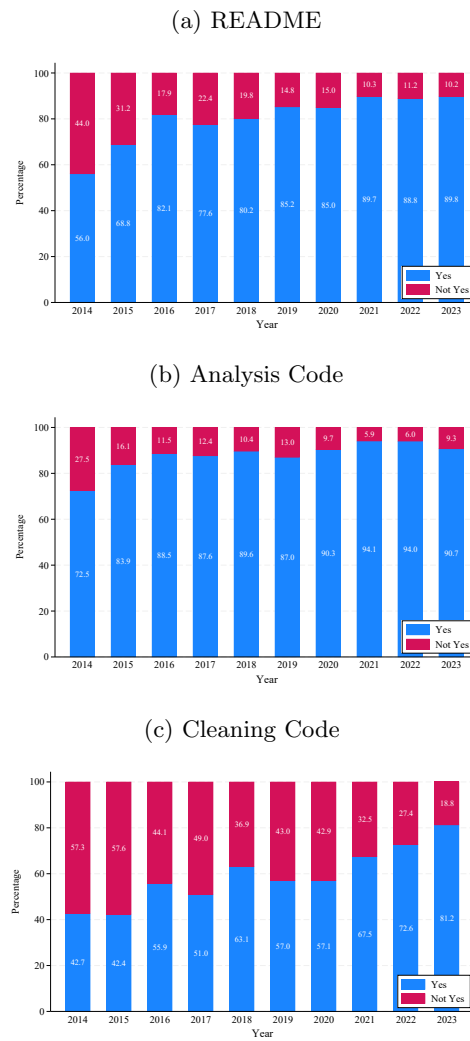


(b) Political Science



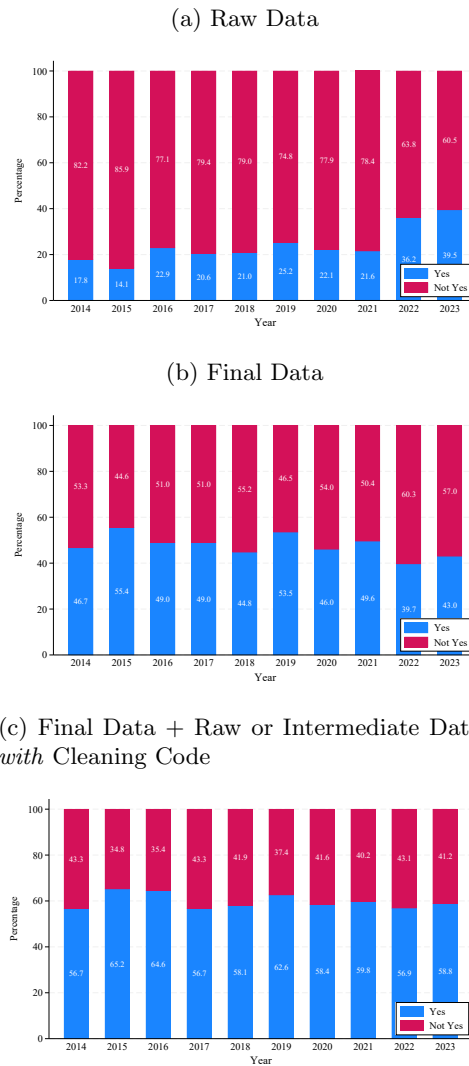
Panel (a) is for papers published in economics journals where Panel (b) is for papers published in political science. The total sample is the same as Figure 17 is 1150 papers, where 850 papers are in the economics sample and 300 papers are in the political science sample.

Fig. 19: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder



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

Fig. 20: Percentage Replication Folders' with Contents Conditional on they Should Have a Replication Folder



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

1873

1874

SUPPLEMENTARY MATERIALS APPENDIX
TABLES

Table 1: Shifts in Statistical Significance Regions

Original Significance Level	Re-Analysis Significance Level			Total
	Sign Change	Not Sig.	Sig. at 5%	
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

Notes: 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

Notes: 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

Notes: 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%

Notes: 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

Top 10 JEL Codes in our Sample	Our Sample (All)		Representative 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

Notes: This table compares the JEL Codes in our sample and in a representative sample of economics papers ([42]). 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

RQ	Category				Total
	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	
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

Notes: 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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

Notes: 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

RQ	Category				Total
	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	
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

Notes: 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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

Notes: 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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

RQ	Neg. & Sig.	Neg. & Not Sig.	Category		Total
			Pos. & Not Sig.	Pos. & Sig.	
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

Notes: 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	

Notes: 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

Notes: 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*** (0.029)	-0.094** (0.045)	-0.073 (0.051)	-0.136** (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

Notes: 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 (0.053)	-0.097 (0.063)	-0.134* (0.072)	-0.169* (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

Notes: 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 [45]

	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

Notes: An application of [45]. 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, ∞).

Table 16: Randomization Tests, Significance at 5% Level

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

Notes: Following [31], 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

Notes: 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		Difference	P-value
	Has Data Editor	No Data Editor		
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

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*.

12.17 List of Articles and Reproduction

12.17.1 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?persistentId=doi%3A10.7910%2F2FDVN%2FGLT4FX&version=&q=&fileTypeGroupFacet=%22Text%22&fileAccess=&fileSortField=size>

12.17.2 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>

12.17.3 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."

1966 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/140921/version/V1/view)
1967 [140921/version/V1/view](https://www.openicpsr.org/openicpsr/project/140921/version/V1/view)

12.17.4 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%2F2FDVN%2FFMOR6K&version=&q=&fileTypeGroupFacet=&fileAccess=&fileSortField=type>

1997 **12.17.5 Reproduction Report**

1998 **Title Original Study:** Brahmin Left Versus Merchant Right: Changing Political
 1999 Cleavages in 21 Western Democracies, 1948–2020

2000 **doi:** <https://doi.org/10.1093/qje/qjab036>, Quarterly Journal of Economics

2001 **Report's Abstract:** Gethin, Martínez-Toledano and Piketty (2022) analyze the
 2002 long-run evolution of political cleavages using a new database on socioeconomic
 2003 determinants of voting from approximately 300 elections in 21 Western democracies
 2004 between 1948 and 2020. They find that, in the 1950s and 1960s, voting for the
 2005 "left" was associated with lower-educated and low-income voters. After that, voting
 2006 for the "left" has gradually become associated with higher-educated voters, while
 2007 high income voters have continued to vote for the "right". In the 2010s, there is
 2008 a disconnection between the effects of income and education on voting. In this
 2009 replication, we first conduct a computational reproduction, using the replication
 2010 package provided by the authors. Second, we do a robustness replication testing to
 2011 what extent the original results are robust to i) restricting the sample to "core" left
 2012 and right parties, ii) analyzing the top 80% versus bottom 20%, iii) weighting by
 2013 population, iv) dropping control variables, and v) using country fixed effects. The
 2014 main results of the paper are found to be largely replicable and robust.

2015 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/19.htm>

2016 **Link to Replicators' Package:** <https://osf.io/2hpeq/>

2017 **Original Authors' Response:** "Thank you for your mail and for your interesting
 2018 report! We are happy to see that you were able to easily replicate our results and
 2019 that our main conclusions were found to be largely robust. In this context, we do
 2020 not think that an answer from our side would be particularly useful: we are happy
 2021 with the report as it is.

2022 Thank you for the very valuable work that your institute is producing in testing
 2023 the replicability and robustness of published studies!"

2024 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XUSWG6)
 2025 [persistentId=doi:10.7910/DVN/XUSWG6](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XUSWG6)

2026 12.17.6 Reproduction Report

2027 **Title Original Study:** Bubbles, Crashes, and Economic Growth: Theory and
2028 Evidence

2029 **doi:** <https://doi.org/10.1257/mac.20220015>, American Economic Journal: Macroeconomics
2030

2031 **Report's Abstract:** Guerron-Quintana, Hirano, and Jinnai (2023) explore the
2032 short-, medium-, and long-run effects of financial bubbles on economic growth by
2033 way of a macroeconomic general equilibrium framework. In their model, a key
2034 theoretical result is that, in net terms, the “crowding in” of capital investment
2035 during a bubble ushers the economy onto a higher balanced growth path post-
2036 bubble than it was on pre-bubble (Figure 10), thus (seemingly) suggesting that
2037 economic bubbles are growth-enhancing. In turn, the main result of the paper is
2038 that this positive view of bubbles is a fallacy so long as the latter are recurrent,
2039 namely because a counterfactual economy in which bubbles never occur in the first
2040 place grows at a significantly faster pace (Figure 10). The reason for this is that
2041 the expectation of future bubbles stifles capital investment and, as such, reduces
2042 economic growth in the long run.

2043 We successfully reproduce the paper's main figures using the original code pro-
2044 vided in the replication package. Given the hard-coded nature of all empirical data
2045 used in the paper, most of our efforts are devoted to reproducing the employed
2046 empirical data itself and, in turn, conducting a direct replication with our own mea-
2047 sures. Using various specifications of the HP filter, we are successful in qualitatively,
2048 but not quantitatively reproducing the paper's main time series (stock-market-to-
2049 GDP ratio). Nevertheless, even without updating the model's parameterization,
2050 the paper's main empirical findings (i.e. Figures 8-10) are largely robust to our own
2051 measure. In turn, we are successful in quantitatively reproducing the second key
2052 time series (credit-to-GDP ratio), albeit only with a highly unusual specification
2053 of the HP filter's smoothing parameter (10^{10} instead of 1600 for quarterly data).
2054 We find that, unlike in the case of the stock-to-GDP ratio, the paper's (auxiliary)
2055 findings are not robust to our own credit-to-GDP series

2056 **Link to Full Report:** <https://osf.io/d76tn/>

2057 **Link to Replicators' Package:** <https://osf.io/d76tn/>

2058 **Original Authors' Response:** Provided a short response and answered a
2059 question. Did not provide a final response as of November 2025.

2060 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/173441/version/V1/view)
2061 [173441/version/V1/view](https://www.openicpsr.org/openicpsr/project/173441/version/V1/view)

2062 **12.17.7 Reproduction Report**

2063 **Title Original Study:** Campaign Contributions and Roll-Call Voting in the U.S.
2064 House of Representatives: The Case of the Sugar Industry

2065 **doi:** <https://doi.org/10.1017/S0003055422000466>, American Political Science
2066 Review

2067 **Report's Abstract:** In their study, Grier et al. (2023) explore the causal relation-
2068 ship between campaign contributions and roll-call voting. Their analysis focuses on
2069 the influence of campaign contributions on two specific anti-sugar votes conducted
2070 in 2013 and 2018. The authors identify a substantial increase in inflationadjusted
2071 sugar contributions from the sugar industry to incumbent politicians between these
2072 two voting events. The aim of our research is to replicate and validate the authors'
2073 main models. In addition to cross-platform replication, we conduct several robust-
2074 ness checks to further examine the reliability of their findings. These include (1)
2075 clustering the standard errors, (2) utilizing an Ordinary Least Squares (OLS) model
2076 instead of the authors' logistic regression, and (3) altering the dependent variable
2077 to represent the change in the vote from 2013 to 2018. Our results largely confirm
2078 the authors' findings and reveal additional insights regarding the money buys vote
2079 hypothesis.

2080 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/57.htm>

2081 **Link to Replicators' Package:** <https://osf.io/4hjb9/>

2082 **Original Authors' Final Response:** "We thank the Institute for Replication for
2083 their diligent work replicating and performing some extensions to our 2023 APSR
2084 paper. Replication is an important and often undervalued work in the scientific
2085 process. Of course we are quite pleased to see that our results do replicate and that
2086 the extensions performed largely support the results and ideas we advanced in our
2087 paper. Keep up the good work!"

2088 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2IFZR9)
2089 [persistentId=doi:10.7910/DVN/2IFZR9](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/2IFZR9)

2090 12.17.8 Reproduction Report

2091 **Title Original Study:** Can Information Reduce Ethnic Discrimination? Evidence
2092 from Airbnb

2093 **doi:** <https://doi.org/10.1257/app.20190188>, American Economic Journal: Applied
2094 Economics

2095 **Report's Abstract:** Laouénan & Rathelot (2022) investigate the mechanism
2096 underlying ethnic discrimination using self-collected panel data from Airbnb
2097 between 2014 and 2017. They find that hosts from minority groups charge 3.2%
2098 less than those from the majority group within the same neighbourhood. Using a
2099 theoretical framework, they estimate that the ethnic price gap vanishes as more
2100 information (reviews) become available conditional on observables. The point esti-
2101 mates for their main results are statistically significant at the 1% level. This finding
2102 suggests that ethnic discrimination is due to statistical discrimination rather than
2103 taste-based discrimination. First, we reproduce the original article's main findings
2104 using R, whereby the authors of the original article use STATA. We can repro-
2105 duce the main findings in R except for a few marginal discrepancies at the second
2106 or third decimal place. Second, we extend two robustness analyses reported in the
2107 original article. These robustness analyses impose restrictions on the sample and
2108 these restrictions are not justified in the article. Once these restrictions are not
2109 imposed, the picture becomes more complex and the robustness analysis warrants
2110 more discussion. However, only a small fraction of the observations causes some
2111 ambiguity and there might be good reasons to impose restrictions. Transparently
2112 presenting the robustness analyses with and without restrictions, motivating the
2113 restrictions and discussing its implications for the main findings would have been
2114 desirable. Generally, the original article does a great job with regard to repro-
2115 ducibility by providing data, code and documentation that ease the reproduction
2116 of a complex analysis. We conclude that our reproduction and replication support
2117 the main findings of the original article.

2118 **Link to Full Report:** <https://osf.io/zn98a/>

2119 **Link to Replicators' Package:** [https://github.com/TuanNguyen04/Replication-](https://github.com/TuanNguyen04/Replication-Airbnb)
2120 [Airbnb](https://github.com/TuanNguyen04/Replication-Airbnb)

2121 **Original Authors' Response:** The authors provided initial feedback which the
2122 replicators took it into account.

2123 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/120078/version/V1/view)
2124 [120078/version/V1/view](https://www.openicpsr.org/openicpsr/project/120078/version/V1/view)

2125 12.17.9 Reproduction Report

2126 **Title Original Study:** Can Technology Solve the Principal-Agent Problem?
 2127 Evidence from China's War on Air Pollution

2128 **doi:** <https://doi.org/10.1257/aeri.20200373>, American Economic Review: Insights

2129 **Report's Abstract:** Greenstone et al. examine the effect of the introduction of
 2130 automatic air pollution monitoring on the reporting of local air pollution in China.
 2131 Using 654 regression discontinuity designs (RDDs) based on city-level variation
 2132 in the day that monitoring was automated, they find an immediate and lasting
 2133 increase of 35 percent in reported PM10 concentrations post-automation. More-
 2134 over, they find that automation's introduction increases online searches for face
 2135 masks and air filters by 200 percent and 28 percent, respectively, using an RDD.
 2136 Results are consistent when using an event study design. First, we were able to
 2137 computationally replicate the results. Second, we find that results are robust to
 2138 more flexible specifications of the weather variables, to re-constructed weather vari-
 2139 ables using the same matching procedure as the authors (i.e., closest station) and
 2140 meteorological data with additional weather stations, to alternative construction of
 2141 the weather variables using an inverse distance weighted approach of the surround-
 2142 ing weather stations, and to more flexible choices of fixed effects (up to the city
 2143 level). Finally, we find limited evidence of discontinuity in objective measures of
 2144 ground pollution (i.e., AOD) for a sub-sample using alternative weather variables.
 2145 The estimate, however, is economically insignificant. Moreover, no discontinuity is
 2146 observed in the full sample. Therefore, we believe this result does not invalidate
 2147 the original study's findings.

2148 **Link to Full Report:** <https://osf.io/b7dn2/>

2149 **Link to Replicators' Package:** [https://osf.io/m8hfr/?view_only=](https://osf.io/m8hfr/?view_only=9f6632ec96c0451daf0f8889b9ad2b25)
 2150 [9f6632ec96c0451daf0f8889b9ad2b25](https://osf.io/m8hfr/?view_only=9f6632ec96c0451daf0f8889b9ad2b25)

2151 **Original Authors' Response:** <https://osf.io/b7dn2/>

2152 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/125321/version/V1/view)
 2153 [125321/version/V1/view](https://www.openicpsr.org/openicpsr/project/125321/version/V1/view)

2154 12.17.10 Reproduction Report

2155 **Title Original Study:** Can't We All Just Get Along? How Women MPs Can
2156 Ameliorate Affective Polarization in Western Publics

2157 **doi:** <https://doi.org/10.1017/S0003055422000491>, American Political Science
2158 Review

2159 **Report's Abstract:** We present a replication and extension of Adams et al.
2160 (2023), examining the influence of women Members of Parliament (MPs) on affective
2161 polarization. Conducted during the 2023 Montreal Replication Games, our
2162 analysis reaffirms the original findings through the authors' base R code and a tidyverse
2163 simplification. Our results highlight that the mitigating effect on polarization
2164 is predominantly observed among left-wing respondents, with null effects noted
2165 for centrist and right-wing parties. This discrepancy is attributed to left-wing parties'
2166 explicit commitment to gender equality. Further analysis reveals the study's
2167 robustness across different countries and years (1996-2007) while addressing data
2168 structure and imputation methods to ensure reliability. Our findings underscore
2169 the nuanced role of women MPs in political dynamics, particularly among left-wing
2170 voters, against democratic backsliding concerns.

2171 **Link to Full Report:** <https://osf.io/69px3/>

2172 **Link to Replicators' Package:** <https://osf.io/69px3/>

2173 **Original Authors' Response:** Thank you for replicating our paper Can't We
2174 All Just Get Along? How Women MPs Can Ameliorate Affective Polarization in
2175 Western Publics (APSR 2023) as part of the Montreal Replication Games. We
2176 appreciate the attention to detail and rigor applied to the replication project. We are
2177 pleased that our initial results replicate well. We appreciate your robust approach
2178 to testing the stability of our findings using a country and year 'leave-one-out' cross-validation
2179 strategy. We also thank you for catching the coding error which dropped
2180 a handful of cases from the original analysis; we are glad that the results remain
2181 substantively the same when this error is corrected. We also are interested in the
2182 results from the extension you undertook, finding that our results are primarily
2183 driven by left-wing parties' supporters, in particular parties from the green, radical
2184 left and social Democratic parties. On the other hand, the point estimates are
2185 positive for all parties excepting the conservative and radical right parties, which
2186 can be expected to have the most conservative views on gender roles. We note that
2187 the authors' interpretation, that "the portion of women MPs affects the attitudes
2188 of left-wing voters and not the attitudes of the voters most likely to undermine
2189 democracy" is true, but that the results also suggest that far-right parties, who
2190 most aggressively challenge liberal democratic norms, may be able to "soften" their
2191 image among left-wing voters by running female candidates. This is consistent
2192 with the argument made by Catalano Week et al (2023), that radical right parties
2193 strategically run women to broaden their appeal. Again, we deeply appreciate your
2194 replication and insightful extension of our research.

2195 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AHQVR)
2196 [persistentId=doi:10.7910/DVN/AHQVR](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AHQVR)

2197 **12.17.11 Reproduction Report**

2198 **Title Original Study:** Changing Hearts and Minds? Why Media Messages
2199 Designed to Foster Empathy Often Fail

2200 **doi:** <https://doi.org/10.1086/719416>, Journal of Politics

2201 **Report's Abstract:** This paper focuses on computational reproducibility and
2202 robustness replicability of Gubler et al.'s(2022) studies which examine the effect of
2203 media messages on empathic concern, dissonance, and out-group policy attitudes.
2204 The original paper tests four hypotheses using two online experiments with large
2205 samples from one US state ($N1 = 5,800$; $N2 = 2,200$). Regarding the first experi-
2206 ment, we successfully reproduced the effect that initial antipathy weakens the effect
2207 of humanizing treatment on empathic concern (H1). However, we show that the
2208 moderating effect is negligible and has little practical significance. Moreover, the
2209 individual effect estimates in our analyses slightly differed from the original paper
2210 due to different procedure of data cleaning and minor coding errors in the original
2211 paper. The most relevant difference was the opposite effect of gender than reported
2212 in the original paper. We also show that empathic concern might mediate the effect
2213 of humanizing treatment on attitudes toward immigrants (H3). The original study
2214 rejected the mediation hypothesis due to not finding a total effect of humanizing
2215 treatment on attitudes. In contrast, we found that humanization treatment has a
2216 positive indirect effect on attitudes through empathic concern. At the same time, it
2217 also has a direct negative effect on attitudes. For the second experiment (H1, H2a,
2218 H2b, H3), we attempted to reproduce the results using a different software. We
2219 partially succeeded once receiving support from the authors of the original study.
2220 We note throughout the report issues we have encountered.

2221 **Link to Full Report:** <https://osf.io/zes6g/>

2222 **Link to Replicators' Package:** See Report's Online Appendix for the codes.

2223 **Original Authors' Response:** <https://osf.io/zes6g/>

2224 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FUCDTT)
2225 [persistentId=doi:10.7910/DVN/FUCDTT](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FUCDTT)

2226 **12.17.12 Reproduction Report**2227 **Title Original Study:** Changing Tides: Public Attitudes on Climate Migration2228 **doi:** <https://doi.org/10.1086/715163>, Journal of Politics2229 **Report's Abstract:** See entry below.2230 **Link to Full Report:** [https://www.socialsciencereproduction.org/reproductions/](https://www.socialsciencereproduction.org/reproductions/791/published/index)
2231 [791/published/index](https://www.socialsciencereproduction.org/reproductions/791/published/index)2232 **Link to Replicators' Package:** [https://github.com/alexkustov/](https://github.com/alexkustov/Replication-of-Arias-and-Blair-2021)
2233 [Replication-of-Arias-and-Blair-2021](https://github.com/alexkustov/Replication-of-Arias-and-Blair-2021)2234 **Original Authors' Response:** "Thank you very much for reaching out! We are
2235 very pleased to hear that the results of our study were replicated, and do not need
2236 to provide an answer."2237 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml;](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=name&fileSortOrder=desc)
2238 [jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=name&fileSortOrder=desc)
2239 [2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=name&fileSortOrder=desc)
2240 [name&fileSortOrder=desc](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=e19065118cc12d43f1b412109d41?persistentId=doi%3A10.7910%2FDVN%2FFDML2N&version=&q=&fileAccess=&fileTag=&fileSortField=name&fileSortOrder=desc)

2241 **12.17.13 Reproduction Report**2242 **Title Original Study:** Checking and Sharing Alt-Facts2243 **doi:** <https://doi.org/10.1257/pol.20210037>, American Economic Journal: Economic
2244 Policy

2245 **Report's Abstract:** Henry, Zhuravskaya, and Guriev (2022) examine whether
2246 people are willing to share "alternative facts" espoused by right-wing populist parties
2247 before the 2019 European elections in France and how this interacted with
2248 the availability of fact-checking information. They find that both imposed and
2249 voluntary fact-checking reduce the likelihood of sharing false statements by approximately 45%,
2250 and that imposed and voluntary fact-checking have similar effect sizes. We reproduce these findings
2251 and introduce several alternative estimates to assess the robustness of the original results,
2252 including resolving an inconsistency in the handling of pre-treatment controls. Overall, our results align
2253 with the results of the original paper. The differences we find are small in absolute magnitude but,
2254 since many effects were small, not always trivial in terms of relative differences. This replication
2255 supports the conclusions of the original paper.

2256 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/34.htm>2257 **Link to Replicators' Package:** <https://doi.org/10.5281/zenodo.7858829>2258 **Link to Original Authors' Response:** "Many thanks! No, we won't be writing
2259 a response."2260 **Original Authors' Package:** <https://www.openicpsr.org/openicpsr/project/140161/version/V1/view>
2261
2262

2263 **12.17.14 Reproduction Report**

2264 **Title Original Study:** Child Marriage Bans and Female Schooling and Labor
2265 Market Outcomes: Evidence from Natural Experiments in 17 Low- and Middle-
2266 Income Countries

2267 **doi:** <https://doi.org/10.1257/pol.20200008>, American Economic Journal: Economic
2268 Policy

2269 **Report's Abstract:** By studying child marriage bans in 17 developing countries,
2270 Wilson (2022) finds that raising the minimum legal age of marriage to 18 success-
2271 fully increased the age at first marriage, the age at first birth, and the likelihood of
2272 employment. Additionally, the bans reduced child marriage and increased educa-
2273 tional attainment in urban areas. We replicate these findings by collecting the raw
2274 data from the same sources as the paper and analysing the data following the pro-
2275 cedures described in the paper, without referring to the data and codes provided
2276 by the author. Our findings are consistent with the results of the paper in terms of
2277 the statistical significance of point estimates and differ in magnitude by a negligible
2278 amount.

2279 **Link to Full Report:** <https://osf.io/5yhxc/>

2280 **Link to Replicators' Package:** <https://osf.io/5yhxc/>

2281 **Original Authors' Response:** We could not reach out the author.

2282 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/130784/version/V1/view)
2283 [130784/version/V1/view](https://www.openicpsr.org/openicpsr/project/130784/version/V1/view)

2284 **12.17.15 Reproduction Report**2285 **Title Original Study:** Concentration Bias in Intertemporal Choice2286 **doi:** <https://doi.org/10.1093/restud/rdab043>, Review of Economic Studies

2287 **Report's Abstract:** Dertwinkel-Kalt et al. (2022) examine the effect of concen-
2288 tration bias - the tendency to overweight advantages that are concentrated in time
2289 relative to costs that are spread over multiple time periods - on intertemporal choice
2290 in a laboratory experiment. In their preferred empirical specification, the authors
2291 report that concentration bias leads to a 22.4% higher willingness to work than
2292 explained by a standard model of intertemporal discounting. We conduct a compu-
2293 tational replication of the main results of the paper using the same procedures and
2294 original data. Our results confirm the sign, magnitude and statistical significance
2295 of the author's reported estimates across each of their five main findings.

2296 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/42.htm>2297 **Link to Replicators' Package:** <https://osf.io/d42xr/>

2298 **Original Authors' Response:** "We thank Deer, Ellingsrud, Heuer, and Kordt
2299 (2023) for conducting the Reproduction Report and appreciate that their "results
2300 confirm the sign, magnitude and statistical significance of [our] reported estimates
2301 across each of [our] five main findings" (p. 1). We don't have anything substantive
2302 to add to this. "

2303 **Original Authors' Package:** <https://zenodo.org/records/5091975>

2304 **12.17.16 Reproduction Report**

2305 **Title Original Study:** Cooperative Property Rights and Development: Evidence
2306 from Land Reform in El Salvador

2307 **doi:** <https://doi.org/10.1086/717042>, Journal of Political Economy

2308 **Report's Abstract:** Montero (2022) explores a discontinuity in a land reform in
2309 El Salvador and reports two main findings. First, relative to outside-owned hacien-
2310 das operated by contract workers, the productivity of worker-owned cooperatives is
2311 higher for staple crops and lower for cash-crop. Second, cooperative property rights
2312 increase workers' incomes and compress wage distributions. In this comment, we
2313 show that the latter result rests on two mistakes: three-quarters of the observations
2314 are duplicates and income inequality is calculated over too few workers to be mean-
2315 ingful. When corrected, the data sources and research design provide no credible
2316 evidence regarding the causal effects of ownership structure on income levels and
2317 inequality.

2318 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/20.htm>

2319 **Link to Replicators' Package:** <https://doi.org/10.7910/DVN/AMD3NO>

2320 **Link to Original Authors' Response:** [https://www.journals.uchicago.edu/doi/](https://www.journals.uchicago.edu/doi/10.1086/725234)
2321 [10.1086/725234](https://www.journals.uchicago.edu/doi/10.1086/725234)

2322 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/717042/suppl_file/20190161data.zip)
2323 [1086/717042/suppl_file/20190161data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/717042/suppl_file/20190161data.zip)

2324 **12.17.17 Reproduction Report**

2325 **Title Original Study:** Decentralization Can Increase Cooperation among Public
2326 Officials

2327 **doi:** <https://doi.org/10.1111/ajps.12606>, American Journal of Political Science

2328 **Report's Abstract:** Molina-Garzón, Grillos, Zarychta, and Andersson (2022)
2329 examine how health sector decentralization affects cooperation between public offi-
2330 cials. Using a public goods game conducted in Honduras, they find that officials
2331 who work under decentralized regimes contributed 0.8 more lempiras per round to
2332 a group solidarity fund, compared to officials who work under centralized regimes.
2333 They also find that most of this increase in investment under decentralized regimes
2334 occurred during rounds of the game in which the participants were able to commu-
2335 nicate with each other. Finally, they find that decentralization was associated with
2336 a 14 percentage point increase in the proportion of potential cross-level network
2337 ties between participants that were realized. In this paper, I examine whether these
2338 results are robust to (1) the omission of some individual-level controls that may
2339 have been affected by the decentralization treatment, and (2) the use of a linear
2340 regression model instead of a Poisson regression model for the network analysis. I
2341 find that omitting the individual-level controls leads to similar conclusions about
2342 the effect of decentralization on individual contributions in the public goods game,
2343 but the interaction effect between decentralization and communication becomes
2344 statistically insignificant at the 0.05 level. For the network analysis, I find that using
2345 a linear regression instead of a Poisson regression has little bearing on the magni-
2346 tude of the effect of decentralization on the proportion of ties realized, though the
2347 effect of decentralization becomes statistically insignificant for one version of the
2348 network model.

2349 **Link to Full Report:** <https://osf.io/q3dpt/>

2350 **Link to Replicators' Package:** <https://osf.io/q3dpt/>

2351 **Link to Original Authors' Response:** <https://osf.io/q3dpt/>

2352 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZLHYSZ)
2353 [persistentId=doi:10.7910/DVN/ZLHYSZ](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZLHYSZ)

12.17.18 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>

2381 **12.17.19 Reproduction Report**2382 **Title Original Study:** Digital Addiction2383 **doi:** <https://doi.org/10.1257/aer.20210867>, American Economic Review

2384 **Report's Abstract:** Using an original economic model of digital addiction and a
2385 randomized experiment, Hunt Allcott, Matthew Gentzkow, and Lena Song (2022)
2386 isolate the effect of habit formation and self-control problems on how people use
2387 their smartphones. They find a persistent effect of temporary incentives on reduc-
2388 ing social media usage. With the model-free results, the study shows that (after the
2389 incentive was in effect), participants in the bonus group reduced use by 56, 19 and
2390 12 minutes in periods 3, 4 and 5, respectively, suggesting a persistent effect. But
2391 before the incentive was in effect in period 2, social media use reduced use by 5.1
2392 minutes per day. Participants who used the limit functionality reduced FITSBY use
2393 by over 20 minutes per day, suggesting an impact of self-control problems on social
2394 media use. All these estimates are statistically significant. We perform a direct
2395 replication of the paper. Upon re-calculating the core dependent variable (FITSBY
2396 use by period), we find a small but concerning discrepancy: For a small number
2397 of observations, the aggregated dependent variable does not equal the sum of the
2398 disaggregated categories. Thankfully, this discrepancy does not have a major effect
2399 on the results. Using the provided data, we re-coded the core figures from scratch
2400 and found that we could replicate them all. We also compare the pre-analysis plan
2401 (PAP) with the main study to identify gaps and perform computational repro-
2402 duction/replication of the structural model and model-free analysis. We only find
2403 minor differences between the PAP and the main paper, almost all of which are
2404 acknowledged in the paper.

2405 **Link to Full Report:** <https://osf.io/8kvdf/>2406 **Link to Replicators' Package:** <https://osf.io/8kvdf/>2407 **Link to Original Authors' Response:** <https://osf.io/8kvdf/>2408 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/163822/version/V2/view)
2409 [163822/version/V2/view](https://www.openicpsr.org/openicpsr/project/163822/version/V2/view)

2410 **12.17.20 Reproduction Report**

2411 **Title Original Study:** Do Thank-You Calls Increase Charitable Giving? Expert
2412 Forecasts and Field Experimental Evidence

2413 **doi:** <https://doi.org/10.1257/app.20210068>, American Economic Journal: Applied
2414 Economics

2415 **Report's Abstract:** Samek and Longfield estimate the effect of 'thank you calls'
2416 on the extensive and intensive margins of subsequent donations. Based on a series
2417 of experimental interventions, the authors find no statistically discernable effect
2418 of thank-you calls on either the likelihood of donating again, or on the size of
2419 any subsequent donations made within the period of the study. In a companion
2420 exercise the researchers quantify the ability of experts in charitable fundraising and
2421 non-experts (using the Understanding America Survey) to predict the behaviours
2422 elicited by the experiment. Experts and non-experts (incorrectly) make the same
2423 predictions of an increase to the extensive margin of donation behaviour induced
2424 by the thank you call, and while both groups overestimate the intensive margin,
2425 the non-experts overestimated by a smaller magnitude. We were able to reproduce
2426 the papers findings completely, discovering only one difference in an appendix table
2427 related to the average gift amount — treatment for experiment 1 where only the
2428 constant term of the regression was affected. Upon careful examination of the code
2429 we found a few small errors that did not affect the results (one of the errors in the
2430 code did not seem to be carried through and used anywhere). Finally, we conducted
2431 several extensions of the original analysis which demonstrated that the findings
2432 are robust to heterogeneity of treatment effect by initial donation size, as well as
2433 different specifications of the regression analysis.

2434 **Link to Full Report:** <https://osf.io/fe2tr/>

2435 **Link to Replicators' Package:** https://gitlab.com/c3754/replication-games/-/tree/main/replication%20games%20MTL%202023%20charity?ref_type=heads

2437 **Link to Original Authors' Response:** Waiting for the authors' response.

2438 **Original Authors' Package:** <https://www.openicpsr.org/openicpsr/project/149481/version/V1/view>
2439

2440 **12.17.21 Reproduction Report**

2441 **Title Original Study:** Do Transitional Justice Museums Persuade Visitors?
 2442 Evidence from a Field Experiment

2443 **doi:** <https://doi.org/10.1086/714765>, Journal of Politics

2444 **Report's Abstract:** Balcells et al. (2022) explore the effect of transitional justice
 2445 museums through a field experiment in Santiago, Chile, and attendance at the
 2446 government's remembrance museum, the Museum of Memory and Human Rights
 2447 which looks at the time of Pinochet's dictatorship. The authors want to understand
 2448 how such experiences shape an individual's perceptions of trust in government
 2449 institutions, and transitional justice policies, and how they are affected emotionally.
 2450 Additionally, they seek to measure how long they last over time. They do this by
 2451 creating treatment (museum attendance) and control (non-attendance) groups and
 2452 administering pre-and post-treatment surveys and estimating the 'complier average
 2453 causal effect' (CACE). They find that satisfaction with the current government
 2454 significantly increases for the treatment group, looking over the entire population
 2455 ($= 0.15, p = .04$) as measured with a 4-point Likert scale and support for a military
 2456 government significantly drops by 11% ($= 0.11, p = .002$) across ideological stances.
 2457 We first reproduce their results and find no major coding errors. Second, we test
 2458 the robustness of the effects by 1) testing for heterogeneous effects by gender, 2) we
 2459 combine the emotion variables into two indices, a mobilization and demobilization
 2460 index, and 3) conduct a causal mediation analysis to see how confidence in the
 2461 church may mediate effects found in the study.

2462 **Link to Full Report:** <https://osf.io/m3hwg/>

2463 **Link to Replicators' Package:** <https://osf.io/m3hwg/>

2464 **Original Authors' Response:** "We thank all involved for their interest in our
 2465 work. We are happy to hear that the results from our paper successfully replicated.
 2466 We are intrigued by the additional analyses performed by the replicators. We hope
 2467 their insights and results can inform future theorizing and empirical studies of the
 2468 impact of Transitional Justice."

2469 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml;](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=type)
 2470 [jsessionid=a651e0dbc9a5c140152aa84be2e0?persistentId=doi%3A10.7910%](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=type)
 2471 [2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=type)
 2472 [Public&fileSortField=type](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FTNFDDX&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=type)

2473 **12.17.22 Reproduction Report**

2474 **Title Original Study:** Does Competence Make Citizens Tolerate Undemocratic
2475 Behavior?

2476 **doi:** <https://doi.org/10.1017/S0003055422000119>, American Political Science
2477 Review

2478 **Report's Abstract:** We replicate the analysis conducted by Frederiksen, 2022a.
2479 We focus on assessing the computational and robustness replicability of their work.
2480 We find that their main exhibits and supplementary analysis are replicable, both
2481 when running their original Stata replication package, and when we attempt to
2482 replicate their findings from scratch in R. We also conduct additional robustness
2483 checks by estimating additional specifications and by subsetting the dataset by the
2484 time taken by the respondent to complete the survey. We again find that their work
2485 is robust to our battery of alternative specifications.

2486 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/28.htm>

2487 **Link to Replicators' Package:** [https://github.com/tjbrailey/nottingham_](https://github.com/tjbrailey/nottingham_replication.2023)
2488 [replication.2023](https://github.com/tjbrailey/nottingham_replication.2023)

2489 **Link to Original Authors' Final Response:** "Thanks a lot for this initiative
2490 and not least for replicating my results."

2491 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NGFLRO)
2492 [persistentId=doi:10.7910/DVN/NGFLRO](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NGFLRO)

2493 **12.17.23 Reproduction Report**

2494 **Title Original Study:** Does Patient Demand Contribute to the Overuse of
2495 Prescription Drugs?

2496 **doi:** <https://doi.org/10.1257/app.20190722>, American Economic Journal: Applied
2497 Economics

2498 **Report's Abstract:** We replicate Lopez et al.'s (2022) study on gatekeeping costs
2499 and the potential evidence for patient-driven and doctor-driven demand. Using
2500 their publicly available source materials, we first re-run their analysis "as is" to see
2501 if their results can be exactly replicated. We then expand the analysis to include
2502 patients previously excluded for not being acutely ill, offering a broader perspective
2503 on medication demand among all patient types. The findings confirm Lopez et al.'s
2504 results.

2505 **Link to Full Report:** <https://osf.io/x7g9z/>

2506 **Link to Replicators' Package:** <https://osf.io/x7g9z/>

2507 **Link to Original Authors' Response:** Provided feedback to an initial report.
2508 Final response: "Thank you very much for sharing the updated report. We appre-
2509 ciate that the authors of the replication reworked the paper and have no further
2510 response or comments."

2511 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/126722/version/V1/view)
2512 [126722/version/V1/view](https://www.openicpsr.org/openicpsr/project/126722/version/V1/view)

2513 **12.17.24 Reproduction Report**

2514 **Title Original Study:** Does Public Opinion Affect the Preferences of Foreign
 2515 Policy Leaders? Experimental Evidence from the UK Parliament

2516 **doi:** <https://doi.org/10.1086/719007>, Journal of Politics

2517 **Report's Abstract:** The study by Chu and Recchia (2022) tests the hypothesis
 2518 that providing public opinion information can shift policymakers' opinions in the
 2519 direction of what the public favors. They surveyed 101 British Members of Par-
 2520 liament (MPs) about their views regarding the United Kingdom's presence in the
 2521 South China Sea. Their results demonstrated that MPs who received information
 2522 about the public opinion poll expressed viewpoints closer to that of public opinion.
 2523 The authors reported an effect that is "substantively meaningful and statistically
 2524 significant at the .10 level." Our computational replication of the original study
 2525 found that the paper is fully computationally reproducible. We successfully repli-
 2526 cated the authors' results but found that the main findings are no longer significant
 2527 when analyzed using unweighted data (see Table 1). We also conducted several
 2528 robustness checks on sub-samples of the data to examine the key analyses both
 2529 with and without weights. Here, we found that the results are once again robust
 2530 and significant when weights are used, but no longer significant when weights are
 2531 not used. As a further robustness check, we found no moderating effect of gender.
 2532 Overall, our replication efforts suggest that the main finding of the original study
 2533 may be sensitive to the use of survey weights.

2534 **Link to Full Report:** <https://osf.io/bqz6w/>

2535 **Link to Replicators' Package:** [https://osf.io/vwt2n/?view_only=](https://osf.io/vwt2n/?view_only=84e52a7c684942a4880410b3c89ff4c6)
 2536 [84e52a7c684942a4880410b3c89ff4c6](https://osf.io/vwt2n/?view_only=84e52a7c684942a4880410b3c89ff4c6)

2537 **Original Authors' Response:** "Thank you for your note and engaging with our
 2538 work. We don't have a formal reply, though this is an honest question: isn't it stan-
 2539 dard practice to use weights when using YouGov's data, since making valid claims
 2540 about representativeness depends on using their weights? YouGov's MP panels
 2541 operate similarly to their public opinion poll, in that their claims to representa-
 2542 tiveness rely on using weights, provided by YouGov. I [Chu] think your write-up
 2543 mentioned that there's a debate about using weights, and cited MTurk data, but I
 2544 think that MTurk is quite different, and yes, I agree I do not use weights for MTurk
 2545 data except unless requested by a reviewer for robustness checks, etc.. But I don't
 2546 think MTurk and the MP representative poll we used is a good comparison in the
 2547 context of evaluating the validity of weighting. In any case, happy to adapt if there
 2548 is a clear consensus on this. Thanks again."

2549 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BNINNL)
 2550 [persistentId=doi:10.7910/DVN/BNINNL](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BNINNL)

2551 **12.17.25 Reproduction Report**

2552 **Title Original Study:** Effective for Whom? Ethnic Identity and Nonviolent
2553 Resistance

2554 **doi:** <https://doi.org/10.1017/S0003055421000940>, American Political Science
2555 Review

2556 **Report's Abstract:** Manekin and Mitts (2022) investigate the success chances of
2557 minority ethnic groups when engaging in non-violent protests demanding political
2558 change. First, using observational data, the authors find that the success rate for
2559 nonviolent campaign tactics is lower for excluded/minority ethnic groups than for
2560 non-excluded/majority ethnic groups. Second, the authors use two original survey
2561 experiments to show that non-violent protest by ethnic minorities is perceived as
2562 more violent and requiring more policing than identical protest by majorities. This
2563 report reproduces the paper computationally and conducts several sensitivity anal-
2564 yses for both the observational and the experimental parts of the paper. We can
2565 confirm the general direction of the postulated effects, but evidence becomes less
2566 consistent (effect magnitudes and significance levels are not robust to some of the
2567 changes).

2568 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/86.htm>

2569 **Link to Replicators' Package:** <https://zenodo.org/records/10193470>

2570 **Original Authors' Response:** Cannot provide a response.

2571 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SHHVCA)
2572 [persistentId=doi:10.7910/DVN/SHHVCA](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SHHVCA)

2573 **12.17.26 Reproduction Report**

2574 **Title Original Study:** Enabling or Limiting Cognitive Flexibility? Evidence of
2575 Demand for Moral Commitment

2576 **doi:** <https://doi.org/10.1257/aer.20201333>, American Economic Review

2577 **Report's Abstract:** We computationally reproduce Saccardo and Serra-Garcia
2578 (2023) where subjects exploit cognitive flexibility by viewing their incentives first
2579 and partially ignoring product quality information, and hence, recommend the
2580 incentivized product. We find one major coding error for the variable Selfishness.
2581 Additionally, two of the “moral cost” questions more likely capture spitefulness.
2582 After correcting the erroneous coding or dropping the two questions, we find
2583 stronger support for the authors’ main conclusion regarding Selfishness driving
2584 incentive information avoidance with double effect size. Finally, we find weak evi-
2585 dence that subjects update their posterior beliefs differently depending on the
2586 product they are incentivized to recommend.

2587 **Link to Full Report:** <https://osf.io/nwds7>

2588 **Link to Replicators’ Package:** <https://osf.io/yfdet/>

2589 **Link to Original Authors’ Response:** [https://www.aeaweb.org/doi/10.1257/](https://www.aeaweb.org/doi/10.1257/aer.20201333.appx)
2590 [aer.20201333.appx](https://www.aeaweb.org/doi/10.1257/aer.20201333.appx)

2591 **Original Authors’ Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/180741/version/V1/view)
2592 [180741/version/V1/view](https://www.openicpsr.org/openicpsr/project/180741/version/V1/view)

2593 **12.17.27 Reproduction Report**2594 **Title Original Study:** Entertaining Beliefs in Economic Mobility2595 **doi:** <https://doi.org/10.1111/ajps.12702>, American Journal of Political Science2596 **Report's Abstract:** In Entertaining Beliefs in Economic Mobility (AJPS 2023)

2597 Kim finds that watching “rags-to-riches” style reality TV programs strengthens
2598 Americans’ belief in the American dream. Through thoughtful and clever exper-
2599 imental and observational analysis, she demonstrates that exposure to television
2600 programs containing everyday people working hard to earn large prizes increases
2601 Americans’ belief that success can be internally attributed and that economic mobil-
2602 ity is possible. We computationally replicate Kim’s results, finding no major errors
2603 in her coding or statistical procedure. We also include several robustness checks.
2604 First, we merge her two experimental samples, which increases the precision of her
2605 main quantity of interest such that it attains conventional levels of statistical signif-
2606 icance. Second, we recreate tables and visualizations for alternative specifications
2607 of her main observational results. The original results are robust to these alterna-
2608 tive models, but we do find that if sports programming is operationalized in the
2609 same manner as “rags-to-riches” programming, the sign, magnitude, and signifi-
2610 cance of watching either programming type are similar. We also uncover a partisan
2611 interaction effect, as only Democrats change their beliefs in economic mobility with
2612 increased TV viewing.

2613 **Link to Full Report:** <https://osf.io/xf5w2/>2614 **Link to Replicators’ Package:** [https://github.com/jacobawinter/rep_games_](https://github.com/jacobawinter/rep_games_2023)
2615 [2023](https://github.com/jacobawinter/rep_games_2023)

2616 **Original Author’s Response:** “Thanks for this! I have no particular response
2617 per se. I’m grateful for your collective efforts to make social science much more
2618 transparent.”

2619 **Original Authors’ Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FVRZJU)
2620 [persistentId=doi:10.7910/DVN/FVRZJU](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FVRZJU)

2621 **12.17.28 Reproduction Report**

2622 **Title Original Study:** Evaluating Deliberative Competence: A Simple Method
2623 with an Application to Financial Choice

2624 **doi:** <https://doi.org/10.1257/aer.20210290>, American Economic Review

2625 **Report's Abstract:** Ambuehl et al. (2022) explore ways to evaluate interven-
2626 tions designed to enhance decision-making quality when individuals misjudge the
2627 outcomes of their choices. The authors propose a novel outcome metric that can
2628 distinguish between interventions better than conventional metrics such as financial
2629 literacy and directional behavioral responses. The proposed metric, which trans-
2630 forms price-metric bias into interpretable welfare loss measures, can be applied
2631 to evaluate various training programs on financial products. Table 4 of the paper
2632 reports the authors' significant main point estimates at the 1% level. In this repli-
2633 cation exercise, we first replicate the main findings of the original paper. Then,
2634 we modify the clustering method by using k-means with demographic variables as
2635 inputs, then we re-calculate standard errors with jackknife estimators. Finally, we
2636 include subjects who were excluded by the authors due to multiple switching in
2637 the multiple price lists. We find that all of these replications result in robust find-
2638 ings. Additionally, we successfully replicate Figure 4 from the paper. Notably, this
2639 replication demonstrates the insensitivity of the results to the choice of distance
2640 metric.

2641 **Link to Full Report:** <https://osf.io/scgbt/>

2642 **Link to Replicators' Package:** <https://osf.io/scgbt/>

2643 **Link to Original Authors' Response:** Authors provided feedback.

2644 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/171681/version/V1/view)
2645 [171681/version/V1/view](https://www.openicpsr.org/openicpsr/project/171681/version/V1/view)

2646 **12.17.29 Reproduction Report**

2647 **Title Original Study:** Exposure and Preferences: Evidence from Indian Slums
2648 **doi:** <https://doi.org/10.1111/ajps.12570>, American Journal of Political Science
2649 **Report's Abstract:** Successful computational reproducibility. The replicators
2650 could not conduct the robustness checks without the help of the author.
2651 **Link to Full Report:** No report.
2652 **Original Author's Response:** "Thanks for your email. I am not interested in
2653 participating."
2654 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AV8PLT)
2655 [persistentId=doi:10.7910/DVN/AV8PLT](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AV8PLT)

2656 12.17.30 Reproduction Report

2657 **Title Original Study:** Finance and Green Growth

2658 **doi:** <https://doi.org/10.1093/ej/ueac081>, Economic Journal

2659 **Report's Abstract:** De Haas and Popov (2023) estimate the effect of country-level
2660 financial sector size and structure on decarbonization to show that countries with
2661 relatively more equity versus debt financing have more emission-efficient economies.
2662 We uncover multiple coding errors that change the magnitude and the precision
2663 of the coefficients of interest. These coding errors include misreporting of standard
2664 errors, and misspecifying generalized method of moments (GMM) estimators. We
2665 further provide robustness tests of the results to (1) restricting the sample to con-
2666 sistent sets of countries across the country and country-by industry samples, and
2667 (2) using a limited information maximum likelihood (LIML) estimator to address a
2668 weak-instrument problem. We find that the results from the robustness checks are
2669 qualitatively different from the original results but similar to the corrected results.

2670 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/95.htm>

2671 **Link to Replicators' Package:** <https://osf.io/h8ct2/>

2672 **Link to Original Authors' Response:** [https://econpapers.repec.org/paper/](https://econpapers.repec.org/paper/zbwi4rdps/96.htm)
2673 [zbwi4rdps/96.htm](https://econpapers.repec.org/paper/zbwi4rdps/96.htm)

2674 **Original Authors' Package:** <https://zenodo.org/records/7220094>

2675 **12.17.31 Reproduction Report**

2676 **Title Original Study:** Flight to Safety: COVID-Induced Changes in the Intensity
 2677 of Status Quo Preference and Voting Behavior

2678 **doi:** <https://doi.org/10.1017/S0003055421000691>, American Political Science
 2679 Review

2680 **Report's Abstract:** Bisbee and Honig (2022) examine the effect of the COVID-
 2681 19 pandemic on voting for Bernie Sanders in the 2020 Democratic Party primary
 2682 using a difference-in-differences design, finding evidence that exposure to COVID-
 2683 19 resulted in a 7-15 percentage point increase in voting for Biden. The study also
 2684 uses a regression design with district-level fixed effects to estimate the effect of
 2685 the COVID-19 pandemic on voting for anti-establishment candidates during the
 2686 US 2020 House primaries. It finds evidence that an increase in COVID cases was
 2687 associated with a decline in voting for anti-establishment candidates in general,
 2688 and for those endorsed by the Tea Party. We re-run the code for all tests in this
 2689 paper, successfully reproducing its results in a preliminary replication. We then
 2690 use the De Chaisemartin and D'Haultfoeuille difference-in-differences estimator to
 2691 replicate their main results, finding that though the coefficient remains negative,
 2692 the results are not statistically significant. We also replicate their tests regard-
 2693 ing US House primary candidates using a different measure of anti-establishment
 2694 candidates. Here, we find that the interaction term between anti-establishment can-
 2695 didates and COVID-19 remain statistically significant, with the same sign. Finally,
 2696 we employ an expanded dataset that includes Congressional primary candidates
 2697 that were omitted in the initial dataset, as well as a re-coded extremism variable
 2698 that also includes candidates endorsed by Donald Trump. These updated find-
 2699 ings corroborate the paper's initial results. However, due to a restrictive number
 2700 of observations that interfered with our application of the De Chaisemartin and
 2701 D'Haultfoeuille estimator, we believe that the expanded U.S. House primary results
 2702 constitute the more robust half of our replication.

2703 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/36.htm>

2704 **Link to Replicators' Package:** [https://github.com/Dmscates/](https://github.com/Dmscates/Bisbee-and-Honig-2022-Flight-to-Safety-Replication)
 2705 [Bisbee-and-Honig-2022-Flight-to-Safety-Replication](https://github.com/Dmscates/Bisbee-and-Honig-2022-Flight-to-Safety-Replication)

2706 **Original Authors' Response:** "You guys are amazing. Thank you for doing this!
 2707 We are impressed by your rigor and grateful for the introduction to DCD'H DiD
 2708 estimator that we'll have to add to the repertoire. We were working on the condi-
 2709 tional accept when the flurry of generalized DiD work (Goodman-Bacon, Callaway,
 2710 etc.) was blowing up [...] We also appreciate the manner in which you communi-
 2711 cated with us during the course of your re-analysis, and the thoughtfulness of your
 2712 report. [...] Although I'm sure it is a logistical nightmare and likely would add even
 2713 more delays to the publication pipeline, it would be very pro-science if this type
 2714 of replication were part of a journal's own pre-publication process. (This is what I
 2715 naively thought replication meant back when I got my first publication, and have
 2716 been disappointed in the process ever since.)"

2717 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S5YMS7)
 2718 [persistentId=doi:10.7910/DVN/S5YMS7](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S5YMS7)

2719 **12.17.32 Reproduction Report**

2720 **Title Original Study:** Gender Differences in Cooperative Environments? Evi-
2721 dence from The U.S. Congress

2722 **doi:** <https://doi.org/10.1093/ej/ueab069>, Economic Journal

2723 **Report's Abstract:** Gagliarducci and Paserman (2022) study gender differences
2724 in cooperative behavior among politicians using information from the U.S. House
2725 of Representatives between 1988 and 2010 on (i) the number of co-sponsors on bills
2726 and (ii) the share of co-sponsors from the rival party. Through different empirical
2727 strategies, they show that women-sponsored bills tend to have more co-sponsors,
2728 but the gap is only statistically significant among Republicans. Moreover, Repub-
2729 lican women recruit a significantly larger share of co-sponsors from the rival party
2730 than Republican men, whereas the opposite is true among Democrats. GP argue
2731 that the observed pattern is consistent with a commonality of interest driving coop-
2732 eration, rather than gender per se, since during this period Republican women
2733 were ideologically closer to the rival party than their male colleagues, while female
2734 Democrats were further away. We examine the robustness of these findings to (i)
2735 the correction of some errors in two control variables of the dataset used by GP and
2736 (ii) clustering the standard errors at the individual level, instead of individual-term.
2737 These changes have a relatively minor impact on results: most coefficients are still
2738 statistically significant and the main conclusions from the analysis are confirmed.
2739 Furthermore, we extend the analysis to the 2011-2020 period. The analysis of gen-
2740 der differences in bipartisan cooperation confirms GP's hypothesis that ideological
2741 distance plays an important role. However, results are slightly different when we
2742 analyze overall cooperation. The gender gap in favor of women is larger in magni-
2743 tude than in GP and it is statistically significant in several specifications, providing
2744 support for the hypothesis that gender also matters for cooperation.

2745 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/75.htm>

2746 **Link to Replicators' Package:** <https://www.dropbox.com/scl/fo/dmvgx5wlgql3tz98dx47u/h?rlkey=af83xiqrkw70rbjicdunoach9&dl=0>

2747 **Link to Original Authors' Response:** <https://osf.io/w48tf/>

2748 **Original Authors' Package:** <https://zenodo.org/records/5111360>

2750 **12.17.33 Reproduction Report**

2751 **Title Original Study:** Good Reverberations? Teacher Influence in Music Com-
2752 position since 1450

2753 **doi:** <https://doi.org/10.1086/718370>, Journal of Political Economy

2754 **Report's Abstract:** Borowiecki (2022) studies the influence of teachers on the
2755 style of their students in the domain of musical composition. The author finds that
2756 realized student-teacher pairs are on average 0.2-0.3 standard deviations more sim-
2757 ilar to unrealized, but possible, student-teacher pairs. In this report we provide the
2758 results of our replication of Borowiecki (2022). We direct our attention to the fol-
2759 lowing tasks: 1) Replicating the outcome variables used in the paper, starting from
2760 the raw data, and generating alternative measures of similarity between students
2761 and teachers 2) Testing the validity of the random teacher-student pairing, a key
2762 assumption for the validity of the estimation strategy employed in the paper. We
2763 can replicate most of the outcome variables, but not all of them, due to incom-
2764 plete raw data. Our alternative measures of similarity confirm the robustness of
2765 the original results. We find significantly different characteristics between paired
2766 and unpaired students, suggesting that matching between students and teachers
2767 does not occur randomly. However, controlling for these characteristics in the main
2768 regressions leads to quantitatively similar results to the ones reported in the original
2769 paper.

2770 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/27.htm>

2771 **Link to Replicators' Package:** [https://www.dropbox.com/scl/fo/](https://www.dropbox.com/scl/fo/6hecmgjsq3mjo5ekkv8lm/h?rlkey=ftuoe4mf5f9jon0hiabb4brtn&dl=0)
2772 [6hecmgjsq3mjo5ekkv8lm/h?rlkey=ftuoe4mf5f9jon0hiabb4brtn&dl=0](https://www.dropbox.com/scl/fo/6hecmgjsq3mjo5ekkv8lm/h?rlkey=ftuoe4mf5f9jon0hiabb4brtn&dl=0)

2773 **Link to Original Authors' Response:** <https://osf.io/79e2z/>

2774 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/718370/suppl_file/20210405data.zip)
2775 [1086/718370/suppl_file/20210405data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/718370/suppl_file/20210405data.zip)

2776 **12.17.34 Reproduction Report**

2777 **Title Original Study:** Hate Crimes and Gender Imbalances: Fears over Mate
2778 Competition and Violence against Refugees

2779 **doi:** <https://doi.org/10.1111/ajps.12595>, American Journal of Political Science

2780 **Report's Abstract:** Dancygier et al. (2022) ascribe anti-refugee hate crime in
2781 Germany from 2015 to 2017 to the fear of mate competition felt by native Ger-
2782 man men, amplified by growing refugee populations and existing gender gaps. In
2783 a replication of this article, we discovered that the substantively and statistically
2784 significant relationship between perceptions of mate competition and support for
2785 anti-refugee violence found in a 2016–17 survey of adults in Germany were robust
2786 when analyzed with ensembles of regression trees permitting arbitrary interactions
2787 in a large design matrix. However, statistically significant pairwise comparisons
2788 between survey respondents' perceptions of mate competition across strata of the
2789 municipality-level gender gap as recorded by German censuses were not robust to
2790 controlling the family-wise Type I error rate. Moreover, statistically significant rela-
2791 tionships between the gender gap and the incidence of hate crime in Germany in the
2792 authors' panel regressions vanished in a wide range of specifications with munic-
2793 ipality fixed effects—in certain cases, being replaced with statistically significant
2794 estimates of the opposite sign.

2795 **Link to Full Report (and Initial Version of the Report):** [https://osf.io/](https://osf.io/5n3ds/)
2796 [5n3ds/](https://osf.io/5n3ds/)

2797 **Link to Replicators' Package:** <https://osf.io/5n3ds/>

2798 **Link to Original Authors' Response:** <https://osf.io/5n3ds/>

2799 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QXJDJ5)
2800 [persistentId=doi:10.7910/DVN/QXJDJ5](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QXJDJ5)

2801 **12.17.35 Reproduction Report**

2802 **Title Original Study:** Historical Lynchings and the Contemporary Voting
2803 Behavior of Blacks

2804 **doi:** <https://doi.org/10.1257/app.20190549>, American Economic Journal: Applied
2805 Economics

2806 **Report's Abstract:** Williams (2022) ties the political participation of Blacks to
2807 historical lynchings that occurred in the United States. Her findings document lower
2808 Black voter registration rates in southern counties with greater number of historical
2809 lynchings. We show that this effect is driven by four outlier counties with relatively
2810 high Black lynching rates. Excluding these counties from the analysis yields a point
2811 estimate that is no longer statistically significant. Dropping the ninety-fifth per-
2812 centile lynching rates and correcting the errors in voter registration rates rule out
2813 the effect size reported by Williams (2022), which now becomes close to zero and
2814 statistically insignificant. We also show that the main results are highly sensitive
2815 to the way lynching and voter registration rates are measured.

2816 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/32.htm>

2817 **Link to Replicators' Package:** <https://osf.io/hv7wp/>

2818 **Original Author's Response:** No response to our emails.

2819 **Original Author's Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/136741/version/V1/view)
2820 [136741/version/V1/view](https://www.openicpsr.org/openicpsr/project/136741/version/V1/view)

2821 **12.17.36 Reproduction Report**2822 **Title Original Study:** Hobo Economicus2823 **doi:** <https://doi.org/10.1093/ej/ueab103>, Economic Journal

2824 **Report's Abstract:** Peter Leeson, August Hardy and Paola Suarez (2022) test
 2825 maximizing behaviour of panhandlers at several Metrorail stations in Washington,
 2826 D.C. Their main findings are that "stations with more panhandling opportunities
 2827 attract more panhandlers" (the first statement) and that "cross-station differences
 2828 in hourly panhandling receipts are statistically indistinguishable from zero" (the
 2829 second statement). We test computational reproducibility and robustness replica-
 2830 bility of their results. We can reproduce both statements, in Stata and R. Our
 2831 robustness replications for the first statement confirm the authors' results in the
 2832 vast majority of cases (replication was successful in 91% of the cases). Our robust-
 2833 ness replications for the second statement might raise doubts on this finding. We
 2834 run weighted ANOVA tests, we change the bounds in minutes used by authors by
 2835 5 minutes in their robustness checks, we run Bartlett's tests of equality of vari-
 2836 ances of means, and run pair-wise tests of equality of means. In three out of four
 2837 cases we cannot replicate the results, and the differences (of either means, medi-
 2838 ans or variances of donations) across Metrorail stations are statistically different
 2839 from zero. We hypothesize that panhandlers have a general idea about which sta-
 2840 tions have more passers-by, and will rationally go more often there. However, they
 2841 are unlikely to have information about smaller variations in the number of passers-
 2842 by (e.g., variations in passers-by at the same station over time due to non-public
 2843 events), and therefore might find it difficult to perfectly maximize donations.

2844 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/55.htm>2845 **Link to Replicators' Package:** <https://osf.io/s4bca/>2846 **Original Authors' Response:** Declined to respond.2847 **Original Authors' Package:** <https://zenodo.org/records/5719541>

2848 **12.17.37 Reproduction Report**

2849 **Title Original Study:** How Do Beliefs about the Gender Wage Gap Affect the
2850 Demand for Public Policy?

2851 **doi:** <https://doi.org/10.1257/pol.20200559>, American Economic Journal: Economic
2852 Policy

2853 **Report's Abstract:** We conduct a replication of Settele (2022), a online survey
2854 experiment designed to find out how individual's beliefs about the gender wage
2855 gap affect their policy preferences. We reproduce Results 1 and 2 of the study: how
2856 prior beliefs around the wage gap are distributed among individuals and how a
2857 information treatment causally affects the policy demand. Our re-coded replication
2858 shows that the reported results are robust.

2859 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/12.htm>

2860 **Link to Replicators' Package:** <https://osf.io/j2ubt/>

2861 **Original Authors' Response:** "I very much appreciate the effort of you and your
2862 team to replicate not just my paper, but many others too. It is quite impressive to
2863 see the scope of your project and I am curious about your future plans with this
2864 initiative.

2865 I just read the Reproduction Report for my paper and I think it is great. In
2866 particular, Figures 2 and 3 are really cool. (They are actually new, and offer a really
2867 insightful way of looking at the data. I should have come up with them myself!)

2868 Regarding your question, I don't think the Reproduction Report requires a
2869 formal response from my side. I fully agree with the authors' interpretation of the
2870 results, and just want to say thank you for their great work. Please go ahead and
2871 publish the report whenever it suits you."

2872 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/134041/version/V1/view)
2873 [134041/version/V1/view](https://www.openicpsr.org/openicpsr/project/134041/version/V1/view)

2874 **12.17.38 Reproduction Report**

2875 **Title Original Study:** How Effective Are Monetary Incentives to Vote? Evidence
2876 from a Nationwide Policy?

2877 **doi:** <https://doi.org/10.1257/app.20200482>, American Economic Journal: Applied
2878 Economics

2879 **Report's Abstract:** Successful computational reproducibility. No re-analyses
2880 conducted.

2881 **Link to Original Authors' Response:** Not contacted.

2882 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip)
2883 [1086/720458/suppl_file/20190733data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip)

2884 **12.17.39 Reproduction Report**

2885 **Title Original Study:** How Much Should We Trust the Dictator's GDP Growth
2886 Estimates?

2887 **doi:** <https://doi.org/10.1086/720458>, Journal of Political Economy

2888 **Report's Abstract:** In this brief commentary, we have conducted a robustness
2889 reproducibility and replicability of Martinez's 2022 paper entitled "How much
2890 should we trust the dictator's GDP growth estimates?" by selecting different clus-
2891 ters and omitting fixed effect terms. Concurrently, we conduct sub-sample analyses
2892 and employ alternative measurements for the sake of robustness and direct replica-
2893 bility. Our results are generally robust, yet they also raise some intriguing questions.
2894 First, we attempt to remove the year fixed effect in the model specifications, but
2895 the elimination of the year fixed effect from the baseline equation did not account
2896 for unobserved variables across year, suggesting the variable bias by Oster (2019).
2897 Second, the entirety of the baseline results is influenced by the periods 2007-2013
2898 (for a five-year interval) and 2010-2013 (for a three-year interval). Third, when uti-
2899 lizing a more varied dataset for the autocracy measurement, the effect vanished for
2900 countries that are partially unfree.

2901 **Link to Full Report:** <https://osf.io/4sk52/>

2902 **Link to Replicators' Package:** <https://osf.io/4sk52/>

2903 **Original Author's Final Response:** "As before, please extend my gratitude to
2904 the replicators for their thoughtful work and for taking into consideration my pre-
2905 vious comments. I am reassured by the fact that they were able to replicate all the
2906 original results in the paper. I also find it reassuring that the results prove robust
2907 to additional robustness tests concerning alternative clustering structures for the
2908 standard errors or alternative data sources on political regimes (albeit with some
2909 loss of precision). The heterogeneous effects by subperiod are also quite intrigu-
2910 ing and potentially reflect changes in the geopolitical incentives to overstate GDP
2911 growth in non-democracies. My paper is certainly not the final word on this topic
2912 and these results could be the first step towards new and exciting research."

2913 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip)
2914 [1086/720458/suppl_file/20190733data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/720458/suppl_file/20190733data.zip)

2915 **12.17.40 Reproduction Report**

2916 **Title Original Study:** Ideological Asymmetries and the Determinants of Politi-
2917 cally Motivated Reasoning

2918 **doi:** <https://doi.org/10.1111/ajps.12624>, American Journal of Political Science

2919 **Report's Abstract:** Guay and Johnston (2022) examine asymmetric politically
2920 motivated reasoning on the part of liberals and conservatives. In our replication
2921 of the paper we examine four potential issues with the analysis: confounding in
2922 the numeracy task, heterogeneity across ideological constraints, the use of control
2923 variables, and heterogeneity in the moderator index items. None of these potential
2924 issues are in fact issues. The results are quite robust. We found only one minor
2925 issue with the codebook, which does not affect the results.

2926 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/79.htm>

2927 **Link to Replicators' Package:** <https://osf.io/mh5sk/>

2928 **Link to Original Authors' Response:** "Thank you again for examining our
2929 paper so closely [...] we changed the codebook and appreciate this replication effort."

2930 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CGHTPZ)
2931 [persistentId=doi:10.7910/DVN/CGHTPZ](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CGHTPZ)

2932 **12.17.41 Reproduction Report**2933 **Title Original Study:** Immigration and Redistribution2934 **doi:** <https://doi.org/10.1093/restud/rdac011>, Review of Economic Studies

2935 **Report's Abstract:** Alesina et al. (2023) examine how people perceive the num-
2936 ber and characteristics of migrants and how those perceptions affect their support
2937 for redistribution. They find that respondents from the United States, United
2938 Kingdom, Sweden, Italy, Germany and France markedly overestimate the share of
2939 immigrants in each country, with the average respondent in all countries except
2940 Sweden overestimating by more than a factor of two. We reproduce these results
2941 using the original code and data and test the robustness by (i) including participants
2942 excluded for time to complete the survey, (ii) extending the analysis of mispercep-
2943 tions to all survey respondents, and (iii) using alternative authoritative estimates
2944 of the proportion of immigrants. We find that these checks marginally change the
2945 estimates of the size of the misperception but do not change the conclusions to be
2946 drawn from the analysis. Alesina et al. (2023) also test the effect on support for
2947 redistribution of showing videos on immigrant characteristics. We computationally
2948 reproduced the treatment effects on support for redistribution.

2949 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/40.htm>2950 **Link to Replicators' Package:** <https://osf.io/ajm9g/>2951 **Original Authors' Response:** "Dear Institute for Replication team,

2952 Thank you very much for taking the time to replicate our paper. We appreciate
2953 the important work you do. We are happy to see that our results replicated well
2954 and that our robustness checks were confirmed.

2955 With best wishes, Armando Miano and Stefanie Stantcheva"

2956 **Original Authors' Package:** <https://zenodo.org/records/5997521>

2957 **12.17.42 Reproduction Report**

2958 **Title Original Study:** Indecent Disclosures: Anticorruption Reforms and Political
2959 Selection

2960 **doi:** <https://doi.org/10.1111/ajps.12646>, American Journal of Political Science

2961 **Report's Abstract:** This short report summarises a replication exercise performed
2962 on data from Szakonyi (2021). The original work applies a difference-in-differences
2963 design to the case of an anti-corruption reform implemented in Russia for local
2964 election candidates, mandating financial disclosures. The author applies this design
2965 by comparing the electoral outcomes of municipalities that happened to hold elec-
2966 tions right after the reform with those that held elections right before the reform.
2967 For both groups, the design uses information from the previous electoral cycle as a
2968 pre-treatment benchmark. Using only data provided by the author in the original
2969 dataset, I first verified that results are reproducible when using alternative soft-
2970 ware. Second, I performed two simple placebo tests to obtain evidence on violations
2971 of the design's identifying assumptions. These placebo tests return null results,
2972 reassuring on the reproducibility of the original findings.

2973 **Link to Full Report:** <https://osf.io/gx4d6/>

2974 **Link to Replicators' Package:** <https://osf.io/gx4d6/>

2975 **Original Author's Response:** "Many thanks to the replicator for taking the time
2976 to replicate and extend the paper. The placebo tests are very helpful in illuminating
2977 whether the identifying assumptions hold. I will make sure to run versions of these
2978 in future analyses."

2979 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KDUMRM)
2980 [persistentId=doi:10.7910/DVN/KDUMRM](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KDUMRM)

2981 **12.17.43 Reproduction Report**

2982 **Title Original Study:** Inflammatory Political Campaigns and Racial Bias in
2983 Policing

2984 **doi:** <https://doi.org/10.1093/qje/qjac037>, Quarterly Journal of Economics

2985 **Report's Abstract:** Grosjean et al. (2023) (GMV2023) estimate the causal effect
2986 of a Trump rally on the number of black drivers stopped by police officers, using a
2987 difference-in-difference approach. In their preferred specification, the authors find
2988 that after a Trump rally, the probability that a stopped driver is black increases by
2989 5.74%. This effect is significant at the 1% level. In this report we focus on repro-
2990 ducing the main claim of the paper. First, we reproduce the paper's main findings
2991 and uncover an issue with counties that experience multiple Trump rally treat-
2992 ments, given the original modelling choices taken in GMV2023. When we remove
2993 counties that experience multiple rallies, the estimated effect size drops to 2.46%
2994 and loses statistical significance. Second, we attempt to conduct a direct replica-
2995 bility check, by employing a new data set as a source for the dependent variable.
2996 We use data from the National Incident Based Reporting System (NIBRS). We
2997 observe no effect of Trump rallies both on the original data, covered by NIBRS
2998 and on the NIBRS data. Third, we conduct a robustness replicability exercise by
2999 coding an event-study difference-in-difference design at the day level. We estimate
3000 the event-study in a $[7; +7]$ window. We do not discover any systematic effect of
3001 Trump rallies on the dependent variable from GMV2023.

3002 **Link to Full Report:** <https://osf.io/xadb6/>

3003 **Link to Replicators' Package:** <https://osf.io/c7j58/>

3004 **Original Authors' Response:** <https://osf.io/xadb6/>

3005 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/A3B9HE)
3006 [persistentId=doi:10.7910/DVN/A3B9HE](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/A3B9HE)

3007 **12.17.44 Reproduction Report**

3008 **Title Original Study:** Interaction, Stereotypes, and Performance: Evidence from
3009 South Africa

3010 **doi:** <https://doi.org/10.1257/aer.20181805>, American Economic Review

3011 **Report's Abstract:** Corno, La Ferrara and Burns (2022) exploit the random allo-
3012 cation of freshman roommates in a large South African university to gauge the
3013 impact of a roommate's race on racial attitudes as measured by an implicit associ-
3014 ation test, and on school performance. They notably find that (a) white students
3015 randomly assigned to black roommates have less negative racial stereotypes, and
3016 (b) black students randomly assigned to live with white students have higher GPAs.
3017 We first reproduce all regression tables in Corno et al. (Corno et al. (2022)), and
3018 then test for robustness by varying the controls and conducting influential analysis.
3019 Overall, we find the results for finding (a) and (b) and robust in 15% and 40% of
3020 the robustness checks we ran, and the t/z scores are on average 78% and 85% as
3021 large as the original study.

3022 **Link to Full Report:** <https://osf.io/w7vpu/files>

3023 **Link to Replicators' Package:** <https://osf.io/w7vpu/files>

3024 **Link to Original Authors' Response:** Did not receive formal response as of
3025 November 2025.

3026 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/174501/version/V1/view)
3027 [174501/version/V1/view](https://www.openicpsr.org/openicpsr/project/174501/version/V1/view)

3028 **12.17.45 Reproduction Report**3029 **Title Original Study:** Interventions and Cognitive Spillovers3030 **doi:** <https://doi.org/10.1093/restud/rdab087>, Review of Economic Studies

3031 **Report's Abstract:** In the paper of, Altmann et al. (2022) the authors investigate
3032 whether positive effects which are due to behavioral policy interventions in policy-
3033 targeted domains come along with negative effects in policy non-targeted domains.
3034 Using lab and online experiments where subjects have to solve one policy-focused
3035 decision task and one non-focused background task, the authors show that increas-
3036 ing incentives or steering attention to the former led to higher attention spans,
3037 lower default adherence rates, and a higher choice quality in the decision task.
3038 However, because of steering participants focus to the decision task, lower choice
3039 quality and lower attention spans in the background task emerged as a consequence,
3040 which was particularly pronounced among individuals with lower cognitive capa-
3041 bilities and complex decision tasks. Essentially, the authors also describe that the
3042 negative effects in the background tasks offset the positive effects in the decision
3043 task, ultimately yielding a net-zero effect overall. Therefore, the authors emphasize
3044 policymakers to also consider the potential negative cognitive spillovers in order to
3045 not overestimate the benefits of behavioral policy interventions. All the results the
3046 authors in the main text report are significant on 5% and 1% significance levels.
3047 All findings presented in the main text of the paper can be replicated using the
3048 original Stata code and verified thoroughly using R. Additionally, we performed
3049 two robustness tests to ensure the reliability of the paper's main results, and they
3050 remained consistent. Hence, the reported findings in the paper appear to be robust.

3051 **Link to Full Report:** [https://www.econstor.eu/bitstream/10419/272845/1/](https://www.econstor.eu/bitstream/10419/272845/1/I4R-DP043.pdf)
3052 [I4R-DP043.pdf](https://www.econstor.eu/bitstream/10419/272845/1/I4R-DP043.pdf)

3053 **Link to Replicators' Package:** <https://osf.io/kugbs/>

3054 **Original Authors' Response:** "We do not have any comments regarding the
3055 replication. We just want to briefly express our gratitude for the thorough and
3056 excellent work of the authors of the replication study."

3057 **Original Authors' Package:** <https://zenodo.org/records/5652808>

3058 **12.17.46 Reproduction Report**

3059 **Title Original Study:** Jumping the Gun: How Dictators Got Ahead of Their
3060 Subjects

3061 **doi:** <https://doi.org/10.1093/ej/ueac073>, Economic Journal

3062 **Report's Abstract:** Hariri and Wingender add new nuance to the traditional
3063 wisdom that economic modernisation is a path to democracy. They show that the
3064 diffusion of repressive, military technologies, causes a decline in the number of
3065 democratisations in the following years, and argue that this is because of a greater
3066 ability to forcefully oppress popular dissent. We conduct a robustness replication
3067 exercise, focussed on three tests: i) Are findings robust to alternative weightings of
3068 individual technologies in the instrument for country-aggregate military technol-
3069 ogy? ii) Is high leverage in individual countries, regions or time periods driving the
3070 global findings? iii) Are the strength of the IV and its independence of important
3071 macroeconomic indicators a chance occurrence? The main findings of the paper are
3072 largely robust to these tests.

3073 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/50.htm>

3074 **Link to Replicators' Package:** <https://osf.io/4cx86/>

3075 **Original Authors' Response:** "we do not have any comments to the Reproduc-
3076 tion Report, so I'm just sending you this email to applaud the initiative. You and
3077 the Institute for Replication is doing a great service to the profession."

3078 **Original Authors' Package:** <https://zenodo.org/records/7077694>

3079 **12.17.47 Reproduction Report**

3080 **Title Original Study:** Liquidity Constraints in the U.S. Housing Market

3081 **doi:** <https://doi.org/10.1093/restud/rdab063>, Review of Economic Studies

3082 **Report's Abstract:** Successful computational reproducibility. No re-analyses
3083 conducted.

3084 **Link to Original Authors' Response:** Authors provided feedback and sugges-
3085 tions.

3086 **Original Authors' Package:** <http://doi.org/10.5281/zenodo.5112964>

3087 **12.17.48 Reproduction Report**

3088 **Title Original Study:** Local Elites as State Capacity: How City Chiefs Use Local
3089 Information to Increase Tax Compliance in the Democratic Republic of the Congo
3090 **doi:** <https://doi.org/10.1257/aer.20201159>, American Economic Review

3091 **Report's Abstract:** Balán et al. (2022) evaluate the impact of “local elites”
3092 involvement in local tax collection in a large city in the Democratic Republic of
3093 Congo. Using a randomized controlled trial to vary the identities of tax collectors,
3094 they find that local elites’ involvement raises tax compliance and total revenue by 50
3095 and 44 percent, respectively. The paper argues that the primary mechanism behind
3096 the results is better targeting made possible by local elites’ superior information
3097 about property holders’ willingness and ability to pay. In this replication comment,
3098 we first reproduce the paper’s main results. Then, we assess the robustness of the
3099 results by (1) employing randomization inference for statistical tests; (2) control-
3100 ling for baseline characteristics that are not balanced; and (3) using an alternative
3101 method to examine the claims supporting the preferred mechanism of better tar-
3102 geting. We find robust estimates in (1). However, the results are less robust both
3103 in terms of statistical significance and magnitude for (2) and (3). We conclude that
3104 the average treatment effect is robust, while the main claim about mechanisms,
3105 the information channel, is less robust to alternative estimation approaches. We
3106 contextualize and discuss the significance of these results, including the negligible
3107 revenue potential even under full compliance.

3108 **Link to Full Report:** <https://ideas.repec.org/p/zbw/i4rdps/191.html>

3109 **Link to Replicators’ Package:** [https://github.com/SossouAdjisse/](https://github.com/SossouAdjisse/LocalTaxReplicationProject.git)
3110 [LocalTaxReplicationProject.git](https://github.com/SossouAdjisse/LocalTaxReplicationProject.git)

3111 **Link to Original Authors’ Response:** [https://ideas.repec.org/p/zbw/i4rdps/](https://ideas.repec.org/p/zbw/i4rdps/192.html)
3112 [192.html](https://ideas.repec.org/p/zbw/i4rdps/192.html)

3113 **Original Authors’ Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/147561/version/V1/view)
3114 [147561/version/V1/view](https://www.openicpsr.org/openicpsr/project/147561/version/V1/view)

3115 **12.17.49 Reproduction Report**

3116 **Title Original Study:** Major Reforms in Electricity Pricing: Evidence from a
3117 Quasi-Experiment

3118 **doi:** <https://doi.org/10.1093/ej/ueab076>, Economic Journal

3119 **Report's Abstract:** Labandeira et al. (2022) examine the effect of a policy in
3120 Spain that modified the electricity bill structure for all Spanish households. The pol-
3121 icy simultaneously increased fixed costs and decreased marginal costs on household
3122 electricity bills. Using fixed effects and instrumental variables (IV) specifications,
3123 the main causal finding in the paper is that the reform reduced house- hold electric-
3124 ity consumption for Spanish households by 15%. Their point estimate is statistically
3125 significant at the 1% level. In a similar specification, they find the reform reduced
3126 household expenditures on electricity by 9.8%, statistically significant at the 1%
3127 level. The code provided by the authors is computationally reproducible. We found
3128 two coding errors in different IV specifications, which had served as robustness
3129 checks to their main results. Correcting the errors removes statistical significance in
3130 two of four IV results, but increases the point estimates and statistical significance
3131 in the other two IV results. We also perform robustness checks. The IV estimates
3132 lose statistical significance in two of four robustness checks (with point estimates
3133 changing 1.1% to -39%). However, the OLS regressions are robust to changing
3134 covariates (sign and significance remained for 12 of 14 tests of the OLS specification,
3135 with changes in the estimates ranging from -157% to 64%, but averaging -3.3%).

3136 **Link to Full Report:** <https://osf.io/bysa7/>

3137 **Link to Replicators' Package:** <https://osf.io/bysa7/>

3138 **Original Authors' Response:** Original authors provided feedback. Multiple
3139 rounds of back and forth with replicators.

3140 **Original Authors' Package:** <https://zenodo.org/records/5423782>

3141 **12.17.50 Reproduction Report**

3142 **Title Original Study:** Market Access and Quality Upgrading: Evidence from
3143 Four Field Experiments

3144 **doi:** <https://doi.org/10.1257/aer.20210122>, American Economic Review

3145 **Report's Abstract:** Bold et al. (2022b) investigate the effect of providing access
3146 to a market (i.e. a buyer) which rewards quality with a premium on farm productiv-
3147 ity and farming incomes from smallholder maize farmers in western Uganda, using
3148 a series of randomized experiments and a difference-in-differences approach. We
3149 successfully reproduce the results of this study using the publicly provided replica-
3150 tion packet. Then test the robustness of these results by re-defining treatment and
3151 outcome variables, testing for model misspecification and the leverage of outliers,
3152 and testing for non-random selection in the Fisher-permutation process. Our results
3153 show that the findings in Bold et al. (2022b) are robust to a variety of decisions in
3154 the research process. This evokes confidence in the internal validity of the findings.

3155 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/72.htm>

3156 **Link to Replicators' Package:** [https://journaldata.zbw.eu/dataset/](https://journaldata.zbw.eu/dataset/bold-et-al-american-economic-review-2022)
3157 [bold-et-al-american-economic-review-2022](https://journaldata.zbw.eu/dataset/bold-et-al-american-economic-review-2022)

3158 **Original Authors' Response:** "Thank you very much for sharing the report (and
3159 taking the time to replicate the study). We have no comments."

3160 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/158401/version/V1/view)
3161 [158401/version/V1/view](https://www.openicpsr.org/openicpsr/project/158401/version/V1/view)

3162 **12.17.51 Reproduction Report**3163 **Title Original Study:** Market-Based Monetary Policy Uncertainty3164 **doi:** <https://doi.org/10.1093/ej/ueab086>, Economic Journal

3165 **Report's Abstract:** Bauer et al. (2022) derive market-based monetary policy
3166 uncertainty and uncover an 'FOMC uncertainty cycle' characterized by a fall of
3167 uncertainty after FOMC announcements and its subsequent built-up. Then, the
3168 authors show that the financial markets' response to monetary policy announce-
3169 ments depends on the level of short-rate uncertainty on the day before the FOMC
3170 announcement. First, we reproduced the paper's findings, though with Matlab
3171 version-specific issues. Second, we tested the robustness of the two main results of
3172 the paper. We show that the uncertainty cycle in the monetary policy uncertainty
3173 is confirmed when the crisis period is included in the sample or when the median
3174 instead of the average of changes in the monetary policy uncertainty is considered.
3175 However, the FOMC uncertainty cycle does not appear when the monetary pol-
3176 icy uncertainty index (Husted et al. 2020) or the daily economic policy uncertainty
3177 index (Baker et al. 2016) are used as uncertainty proxies.

3178 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/77.htm>3179 **Link to Replicators' Package:** <https://osf.io/qx8aw/>3180 **Original Authors' Response:** "Thank you, glad to see that this work found our
3181 results to be rock solid!3182 We won't write a response. Do let us know if you have any other questions
3183 about our work."3184 **Original Authors' Package:** <https://zenodo.org/records/5566246>

3185 **12.17.52 Reproduction Report**3186 **Title Original Study:** Market-Based Monetary Policy Uncertainty3187 **doi:** <https://doi.org/10.1093/ej/ueab086>, Economic Journal

3188 **Report's Abstract:** This report replicates and examines Bauer et al.'s (2021)
3189 paper on monetary policy transmission to financial markets. The paper introduces
3190 novel measures of monetary policy uncertainty and analyses its drivers. It also
3191 investigates the impact of uncertainty changes on interest rates and financial asset
3192 prices. We assess reproducibility, consolidate market uncertainty measures using
3193 PCA and Factor Analysis, and rigorously test the reduction of uncertainty after
3194 Federal Market Open Committee (FOMC) announcements. Our findings support
3195 the paper's claim of reduced uncertainty on meeting days. Additionally, we explore
3196 the implications of the uncertainty channel on various financial assets, such as Gold,
3197 the Swiss Franc, European stock indexes, and Bitcoin.

3198 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/76.htm>3199 **Link to Replicators' Package:** https://github.com/YaolangZhong/Nottingham_Replication_Game/tree/main/replication_code3200
3201 **Original Authors' Response:** "Thank you, glad to see that this work found our
3202 results to be rock solid!"3203 We won't write a response. Do let us know if you have any other questions
3204 about our work."3205 **Original Authors' Package:** <https://zenodo.org/records/5566246>

3206 **12.17.53 Reproduction Report**3207 **Title Original Study:** Measuring the Welfare Effects of Shame and Pride3208 **doi:** <https://doi.org/10.1257/aer.20190433>, American Economic Review

3209 **Report's Abstract:** This Reproduction Report examines and extends the research
3210 conducted by Butera, Metcalfe, Morrison, and Taubinsky (2022) on "The Welfare
3211 Effects of Pride and Shame." The original paper explores the welfare implications of
3212 public recognition as a motivator for desirable behavior and introduces an empirical
3213 methodology to measure Public Recognition Utility (PRU), which quantifies the
3214 utility individuals experience when their actions are publicly recognized. This report
3215 focuses on the real effort experiment reported in the paper that was conducted
3216 using a classroom sample, a lab sample, and an online sample. I computationally
3217 reproduce the original results and verify their robustness. While reproducing the
3218 results, I found two minor coding errors in the replication package. Correcting
3219 these errors slightly changes some estimates reported in the paper but does not
3220 turn over any results. The main treatment effect findings are further robust to
3221 using different sets of controls and sample selection criteria. Moreover, I conduct a
3222 heterogeneity analysis which reveals significant variations in how participants value
3223 public recognition. Overall, the replication study confirms the original conclusions
3224 while providing additional insights into the heterogeneity of PRU shapes on an
3225 individual level.

3226 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/64.htm>3227 **Link to Replicators' Package:** [https://github.com/tilmanfries/](https://github.com/tilmanfries/welfare-shame-pride-replication-report)
3228 [welfare-shame-pride-replication-report](https://github.com/tilmanfries/welfare-shame-pride-replication-report)3229 **Original Authors' Final Response:** "Thanks again for all your hard work on
3230 this."3231 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/145141/version/V1/view)
3232 [145141/version/V1/view](https://www.openicpsr.org/openicpsr/project/145141/version/V1/view)

3233 **12.17.54 Reproduction Report**

3234 **Title Original Study:** Mental Health Costs of Lockdowns: Evidence from Age-
 3235 Specific Curfews in Turkey

3236 **doi:** <https://doi.org/10.1257/app.20200811>, American Economic Journal: Applied
 3237 Economics

3238 **Report's Abstract:** This report presents a replication of Altindag et al. (2022)
 3239 performed at the Oslo Replication Games in 2022. Altindag et al. (2022) estimate
 3240 the effects of an age-specific lockdown on mental health outcomes and mobility
 3241 among adults aged 65 and older in Turkey, using a regression discontinuity design.
 3242 The authors find a decline in mobility with a one-day decrease in the number of
 3243 days being outside and an increase in the probability of never going out by 30 per-
 3244 centage points. These point estimates are statistically significant at the 1% level.
 3245 The mobility restrictions lead to a worsening in mental health outcomes of approx-
 3246 imately 0.2 standard deviations, statistically significant at the 10% level in their
 3247 preferred specification. In this paper we accomplish two things. First, we success-
 3248 fully reproduce Altindag et al.'s main findings. Second, we test the ro-bustness
 3249 of the results to a small number of changes to their preferred estimations by (1)
 3250 not clustering the standard errors on the running variable, (2) not including con-
 3251 trol variables, and (3) calculating the optimal bandwidth using another technique.
 3252 Point estimates for mobility outcomes are stable to all three manipulations, and
 3253 standard errors only change marginally. Point estimates and standard errors for the
 3254 mental health outcomes are somewhat more sensitive, especially to changing the
 3255 optimal bandwidth selection method. However, the observed changes are reason-
 3256 ably expected when applying data-driven model selection methods to noisy data
 3257 (to avoid over-fitting, it is likely preferable to apply a less data-driven approach
 3258 like the original authors did). Our general impression is that the original analyses
 3259 and results are both theoretically plausible and credible, despite some defensible
 3260 model dependencies.

3261 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/16.htm>

3262 **Link to Replicators' Package:** <https://osf.io/25u7b/>

3263 **Link to Original Authors' Response:** [https://econpapers.repec.org/paper/](https://econpapers.repec.org/paper/zbwi4rdps/17.htm)
 3264 [zbwi4rdps/17.htm](https://econpapers.repec.org/paper/zbwi4rdps/17.htm)

3265 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/131981/version/V1/view)
 3266 [131981/version/V1/view](https://www.openicpsr.org/openicpsr/project/131981/version/V1/view)

3267 **12.17.55 Reproduction Report**

3268 **Title Original Study:** Mortality, Temperature, and Public Health Provision:
3269 Evidence from Mexico

3270 **doi:** <https://doi.org/10.1257/pol.20180594>, American Economic Journal: Economic
3271 Policy

3272 **Report's Abstract:** Cohen and Dechezleprêtre (2022) investigate the hetero-
3273 geneous impact of temperature on mortality across Mexico, and how affordable
3274 healthcare services that target the low-income population attenuate the mortal-
3275 ity effects of weather events. They find that while extreme temperatures are more
3276 dangerous than less extreme temperatures, the increased frequency of non-extreme
3277 temperatures mean these temperatures cause more deaths. First, we reproduce the
3278 paper's main findings, uncovering a minor coding error that has a trivial effect
3279 on the main results. Second, we test the robustness of the results to clustering at
3280 the state level, omitting precipitation, and using a different weighting scheme. The
3281 original results are robust to all of these changes.

3282 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/90.htm>

3283 **Link to Replicators' Package:** <https://osf.io/q52e4/>

3284 **Original Authors' Response:** Cohen: "We thank The Institute for Replication.
3285 Next time, I will make sure I do not forget Feb. 29th in the code!"

3286 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/125201/version/V1/view)
3287 [125201/version/V1/view](https://www.openicpsr.org/openicpsr/project/125201/version/V1/view)

3288 12.17.56 Reproduction Report

3289 **Title Original Study:** Motivated Beliefs and Anticipation of Uncertainty Reso-
3290 lution

3291 **doi:** <https://doi.org/10.1257/aeri.20200829>, American Economic Review: Insights

3292 **Report's Abstract:** Drobner (2022) examines the effect of manipulating experi-
3293 mental subjects' expectations about uncertainty resolution in learning about their
3294 performance on their belief updating patterns in an ego-relevant domain. In their
3295 preferred empirical specification, the author finds that individuals update their
3296 beliefs optimistically as they exhibit a higher belief adjustment in response to good
3297 compared to bad news only when they do not expect resolution of underlying uncer-
3298 tainty about their performance in an IQ test and neutrally when they know they will
3299 find out their relative performance at the end of the experiment. First, we reproduce
3300 the all of the paper's findings without identifying any coding errors. Second, we test
3301 the robustness of the results to (1) adding individual covariates and (2) excluding
3302 subjects who exhibit a fundamental error in their belief updating from the analysis.
3303 We find no substantial changes in the main coefficients of interest with the inclu-
3304 sion of demographic variables in the analysis, consistent with demonstrated balance
3305 in covariates between the two experimental groups. Yet, several of the main esti-
3306 mates lose statistical significance and change from conservatism (under-updating)
3307 to over-inference (over-updating) in some conditions on the subset of participants
3308 excluding those who exhibit fundamental errors in belief updating.

3309 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/65.htm>

3310 **Link to Replicators' Package:** <https://osf.io/evt3a/>

3311 **Original Authors' Response:** "Thanks for sharing the report. I think it's a great
3312 initiative and feel free to publish this report on your webpage. I will not be able to
3313 provide an "answer"."

3314 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/139262/version/V1/view)
3315 [139262/version/V1/view](https://www.openicpsr.org/openicpsr/project/139262/version/V1/view)

3316 **12.17.57 Reproduction Report**

3317 **Title Original Study:** Multinationals' Sales and Profit Shifting in Tax Havens
3318 **doi:** <https://doi.org/10.1257/pol.20200203>, American Economic Journal: Economic
3319 Policy

3320 **Report's Abstract:** We perform a robustness replication analysis of Laffitte and
3321 Toubal (2022), which considers how multinational corporations shift profit to "tax
3322 havens", jurisdictions where they face lower tax burdens. We find that the main
3323 results of Laffitte and Toubal (2022), are fairly robust to alternative versions of
3324 three important researcher choices: i) the definition of tax havens; ii) the use of
3325 a continuous measure of tax-friendliness rather than a binary classification of tax
3326 havens; and iii) a sample that omits two small but "extreme" tax havens: Bermuda
3327 and Barbados. In all cases, results remain of the same sign and retain statistical
3328 significance, though the magnitudes are somewhat attenuated in our robustness
3329 exercises.

3330 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/37.htm>

3331 **Link to Replicators' Package:** <https://osf.io/3sbmr/>

3332 **Original Authors' Response:** "Thanks for your email and for replicating our
3333 exercise. Your work is useful. We recognize that the results remain consistent
3334 even when considering different interpretations of the haven concept and a smaller
3335 sample of observations.

3336 We are also pleased to hear that the replication file we shared with the AEJ:
3337 Policy has proven helpful."

3338 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/148301/version/V1/view)
3339 [148301/version/V1/view](https://www.openicpsr.org/openicpsr/project/148301/version/V1/view)

3340 12.17.58 Reproduction Report

3341 **Title Original Study:** Multiracial identity and political preferences

3342 **doi:** <https://doi.org/10.1086/714760>, Journal of Politics

3343 **Report's Abstract:** The growing concern regarding reproducibility and replica-
3344 bility of social science re- sults has powered the adoption of open data and code
3345 requirements at journals and norms among researchers. However, even when these
3346 norms and requirements are fol- lowed, changes to the software used in data cleaning
3347 and analysis can render papers non-reproducible. This paper details the challenges
3348 of reproducibility in the face of software updates. We present a case study of a pub-
3349 lished article whose results are no longer reproducible due to changes in the software
3350 used. We then discuss the tools and techniques researchers can use to ensure that
3351 their research remains reproducible despite changes in the software used.

3352 **Link to Full Report:** <https://osf.io/ecymu/>

3353 **Link to Replicators' Package:** [https://github.com/taylorjwright/r_and_p_](https://github.com/taylorjwright/r_and_p_versioning)
3354 [versioning](https://github.com/taylorjwright/r_and_p_versioning)

3355 **Original Authors' Response:** Back and forth between authors and replicators,
3356 but did not obtain a final response as of November 2025.

3357 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BLVJJH)
3358 [persistentId=doi:10.7910/DVN/BLVJJH](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BLVJJH)

3359 12.17.59 Reproduction Report

3360 **Title Original Study:** News Shocks Under Financial Frictions

3361 **doi:** <https://doi.org/10.1257/mac.20170066>, American Economic Journal: Macroeconomics

3362
3363 **Report's Abstract:** Görtz et al. (2022) estimate the effects of innovations to
3364 future total factor productivity (TFP) on financial markets. In a Bayesian vector
3365 autoregression, they identify a TFP news shock as one that explains the largest
3366 share of 40- quarter ahead forecast error variance (FEV) of TFP. Their estimated
3367 impulse responses functions show that a positive news shock significantly decreases
3368 credit market spreads and increases credit market supply. They also find that a
3369 shock that explains the maximum of the FEV of the "excess bond premium" (EBP)
3370 (Gilchrist and Zakrajsek 2012) causes similar responses. These results are consistent
3371 with an estimated DSGE model with financial frictions. We estimate the main IRFs
3372 of the study using the original data and a frequentist estimation approach. We
3373 obtain similar point estimates for the dynamic responses to TFP news and EBP
3374 max-share shocks. We also update their macroeconomic and financial time series,
3375 as some of the data has been revised substantially since their original estimate.
3376 We use the updated data to re-estimate the above-mentioned IRFs, and we find
3377 that the results are robust to this change in the data. Finally, we investigate the
3378 computational reproducibility of their DSGE results, and find that their provided
3379 code (consistent with warnings in their README file) does not execute in the most
3380 recent version of Dynare or Matlab. Using the version indicated in their replication
3381 files, we encounter issues estimating the posterior mode.

3382 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/51.htm>

3383 **Link to Replicators' Package:** [https://github.com/gionikola/](https://github.com/gionikola/replication-game-ucsd)
3384 [replication-game-ucsd](https://github.com/gionikola/replication-game-ucsd)

3385 **Original Authors' Final Response:** "Thank you for the update and considering
3386 our work for replication."

3387 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/130141/version/V1/view)
3388 [130141/version/V1/view](https://www.openicpsr.org/openicpsr/project/130141/version/V1/view)

3389 12.17.60 Reproduction Report

3390 **Title Original Study:** Non-Linearities, State-Dependent Prices and the Trans-
 3391 mission Mechanism of Monetary Policy

3392 **doi:** <https://doi.org/10.1093/ej/ueab049>, Economic Journal

3393 **Report's Abstract:** Ascari and Haber (2022) fill the gaps in the literature by
 3394 showing evidence in favor of the state-dependent sticky price model's predictions
 3395 using the macro-aggregates. We report a replication and robustness check of the
 3396 study. We employ several additional macroeconomic control variables and different
 3397 alternative measurements for monetary policy shocks and find that the original
 3398 results remain qualitatively robust. Our analysis further shows that the turbulent
 3399 periods of inflation in the 1970s and 1980s have an important role in claiming the
 3400 robustness of the original results.

3401 **Link to Full Report:** <https://osf.io/kbwap/>

3402 **Link to Replicators' Package:** <https://osf.io/kbwap/>

3403 **Original Authors' Response:** No response.

3404 **Original Authors' Package:** [https://oup.silverchair-cdn.com/](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~PFKA4t_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA)
 3405 [oup/backfile/Content_public/Journal/ej/132/641/10.1093_ej_ueab049/](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~PFKA4t_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA)

3406 [1/ueab049_replication_package.zip?Expires=1765387359&Signature=](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~PFKA4t_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA)

3407 [Ysj-YfzVPuVPtgGm1ZBFtFNI1APv5x1Rgajkm2orCIsJXt2dZG3Amp92XbuA6m0iP-4LwOFv~PFKA4t_](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~PFKA4t_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA)
 3408 [_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA](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~PFKA4t_&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA)

3409 **12.17.61 Reproduction Report**

3410 **Title Original Study:** Not All Elections Are Created Equal: Election Quality
3411 and Civil Conflict

3412 **doi:** <https://doi.org/10.1086/714778>, Journal of Politics

3413 **Report's Abstract:** Utilizing a time-series cross-sectional dataset on developing
3414 countries, Donno et al. (2022) examine how variation in election quality shapes
3415 opportunities and incentives for civil conflict. Across a number of models in their
3416 analysis, they find that civil conflict is more likely when elections are not free and
3417 fair. They also find that for countries with low integrity elections, the probability of
3418 conflict occurring is higher if it has experienced conflict before. We begin by repro-
3419 ducing Donno et al.'s (2022) main models and findings, which yielded no coding
3420 errors or differences in effect estimates. Afterwards, for replication purposes we run
3421 a series of robustness and conceptual replication tests. For our first replication, we
3422 examine the heterogeneous effect between electoral integrity and ethnic fractional-
3423 ization on conflict. Our second test examines whether a subsample of authoritarian
3424 regimes should have been included in the authors' original analysis.

3425 **Link to Full Report:** <https://osf.io/unhkr/>

3426 **Link to Replicators' Package:** [https://drive.google.com/drive/folders/1Vlwfr3_](https://drive.google.com/drive/folders/1Vlwfr3_Q7c56XFPsQuKUNSnEU_lMFUQz?usp=sharing)
3427 [Q7c56XFPsQuKUNSnEU_lMFUQz?usp=sharing](https://drive.google.com/drive/folders/1Vlwfr3_Q7c56XFPsQuKUNSnEU_lMFUQz?usp=sharing)

3428 **Link to Original Authors' Response:** <https://osf.io/unhkr/>

3429 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8B31FG)
3430 [persistentId=doi:10.7910/DVN/8B31FG](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8B31FG)

3431 12.17.62 Reproduction Report

3432 **Title Original Study:** Parties as Disciplinarians: Charisma and Commitment
 3433 Problems in Programmatic Campaigning

3434 **doi:** <https://doi.org/10.1111/ajps.12638>, American Journal of Political Science

3435 **Report's Abstract:** Hollyer, Klačnja, and Titunik (2022) analyse the trade-off
 3436 that political parties face between running programmatic campaigns and fielding
 3437 charismatic candidates, whose electoral appeal may come at the cost of undermin-
 3438 ing the party brand. They argue that higher electoral volatility prompts parties
 3439 to rely on charismatic candidates, even though they might not be as loyal to
 3440 the party's programmatic stance. They substantiate their argument with a cross-
 3441 national dataset and a quantitative case study in Brazil. We computationally
 3442 reproduced and conducted further robustness tests for their cross-national study by
 3443 translating the Stata code to R. Next, we conducted a computational reproduction
 3444 and some additional robustness tests for the quantitative case study. We find that
 3445 their cross-national analysis is reproducible, albeit with some minor discrepancies.
 3446 The quantitative case study is also largely reproducible and both are robust in sev-
 3447 eral ways. We conclude by making some suggestions about data dissemination and
 3448 robustness checks for authors of regression discontinuity designs.

3449 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/54.htm>

3450 **Link to Replicators' Package:** [https://osf.io/93gfx/?view_only=](https://osf.io/93gfx/?view_only=7063353244d646ffaf7bfd53013e3143)
 3451 [7063353244d646ffaf7bfd53013e3143](https://osf.io/93gfx/?view_only=7063353244d646ffaf7bfd53013e3143)

3452 **Original Authors' Response:** "Thanks for your note and for all the work of
 3453 Kelly, Odermatt, and Metson in replicating our paper. [...] Our read of the Repro-
 3454 duction Reports that the findings in our paper hold in the Kelly et al replication.
 3455 [...] Our sense is that the discrepancies between the replication and original paper
 3456 are sufficiently small, and the task of comparing the replication R code to the orig-
 3457 inal Stata code is likely to be sufficiently demanding of time, that the opportunity
 3458 cost of a thorough response is high. So, I think we'll forgo the opportunity to draft
 3459 a response, and just let the replication stand without reply.

3460 We'll leave it to any sufficiently interested parties with expertise in both Stata
 3461 and R to iron out the discrepancies between the replication and original paper."

3462 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AWSQTW)
 3463 [persistentId=doi:10.7910/DVN/AWSQTW](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/AWSQTW)

3464 **12.17.63 Reproduction Report**

3465 **Title Original Study:** Patience, Risk-Taking, and Human Capital Investment
3466 Across Countries

3467 **doi:** <https://doi.org/10.1093/ej/ueab105>, Economical Journal

3468 **Report's Abstract:** Hanushek et al. (2021) test how country-level measures of
3469 patience and risk-taking from the Global Preference Survey predict student per-
3470 formance on the Programme for International Student Assessment (PISA) math
3471 test. They find that country-level patience positively predicts math test scores and
3472 country-level risk-taking negatively predicts math test scores. They find similar
3473 results when holding country of residence characteristics constant and focusing on
3474 the preferences of the country of origin of migrants. We have checked the com-
3475 putational reproducibility and find that the data and analysis script provided by
3476 the authors allowed us to exactly reproduce the main tables in the paper. We
3477 also checked the robustness replicability by testing how robust the results are to
3478 decisions about imputation, weighting, operationalization of dependent variables,
3479 choice of control variables, and the inclusion of highly leveraged observations. We
3480 see that results are generally robust, though statistical significance of the risk-
3481 taking coefficient in the migrant analysis hinges on whether a control for OECD
3482 country of residence is included. Finally, we check the conceptual replicability of
3483 the results by using data from the Trends in International Mathematics and Science
3484 Study (TIMSS) instead of PISA - a different dataset with a different standardized
3485 test. This exercise shows that their results are robust to expanding the analysis to
3486 countries participating in both PISA and TIMSS.

3487 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/48.htm>

3488 **Link to Replicators' Package:** <https://osf.io/kgt8z/>

3489 **Link to Original Authors' Response:** "We would like to thank the replicators
3490 and compliment them for their thoughtful replication and extension of our paper.
3491 We are particularly impressed by the extension to the TIMSS data, which is actually
3492 great support for the underlying idea. We do not see a reason to formulate a formal
3493 response for your website.

3494 Thank you all for your valuable work!"

3495 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/153101/version/V2/view)
3496 [153101/version/V2/view](https://www.openicpsr.org/openicpsr/project/153101/version/V2/view)

3497 **12.17.64 Reproduction Report**

3498 **Title Original Study:** Peer Effects in Academic Research: Senders and Receives
3499 **doi:** <https://doi.org/10.1093/ej/ueac031>, Economical Journal

3500 **Report's Abstract:** In this report, we provide an overview from reproducing and
3501 replicating Bosquet et al. (2022). As a first step, we successfully reproduce all the
3502 results in the paper, as well as figure A1. All results were fully reproducible and
3503 match the published version of the paper. Next, we carry out three sensitivity
3504 analysis. We examine how the main results change from the weights used, additional
3505 controls, and author-university pairs. The main results are robust to these checks.

3506 **Link to Full Report:** <https://osf.io/czkgw/>

3507 **Link to Replicators' Package:** <https://osf.io/czkgw/>

3508 **Link to Original Authors' Response:** The authors responded to the replicators'
3509 questions. Bosquet then responded to the final report: "I would simply thank the
3510 team of replicators and I am happy to see that the tested results are robust to the
3511 tested alternatives. As written in my previous email, I think those kinds of efforts
3512 are very useful for the community and the credibility of published results so thanks
3513 as well for that."

3514 **Original Authors' Package:** <https://zenodo.org/records/6457037>

3515 **12.17.65 Reproduction Report**

3516 **Title Original Study:** Playing Politics with Environmental Protection: The
3517 Political Economy of Designating Protected Areas

3518 **doi:** <https://doi.org/10.1086/718978>, Journal of Politics

3519 **Report's Abstract:** Mangonnet et al. (2022) examine whether political alignment
3520 at the national and sub-national levels explain the spatial designation of Protected
3521 Areas (PAs) in Brazil. Their identification relies on spatial discontinuities in polit-
3522 ical alignment across municipalities. They find that a president-mayor coalition
3523 alignment reduces the incidence of PAs by about one percentage point, whereas
3524 they find no party alignment effects. We were able to reproduce the paper's findings
3525 using the same code and software. Alternative software routines reproduce their
3526 results with small and inconsequential numerical differences. Moreover, robustness
3527 replications find consistent results for one out the two treatments. Finally, we find
3528 no evidence of fabrication of data.

3529 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/73.htm>

3530 **Link to Replicators' Package:** <https://osf.io/t76jd/>

3531 **Original Authors' Response:** "We are grateful to Laura Villalobos, Jill Caviglia-
3532 Harris, Tharaka Jayalath, and the team at the Institute for Replication for
3533 generously replicating our work. We encourage readers to follow their alternative
3534 software routines for faster estimations."

3535 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N6LIMH)
3536 [persistentId=doi:10.7910/DVN/N6LIMH](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N6LIMH)

3537 **12.17.66 Reproduction Report**

3538 **Title Original Study:** Policy Deliberation and Voter Persuasion: Experimental
3539 Evidence from an Election in the Philippines

3540 **doi:** <https://doi.org/10.1111/ajps.12566>, American Journal of Political Science

3541 **Report's Abstract:** I would characterize my robustness replication as almost
3542 entirely successful. The design checks I report all support a straightforward under-
3543 standing of the design. My effect and uncertainty estimates barely differ from the
3544 original estimates (when compared with like estimation procedures), with any dis-
3545 crepancies attributable to simulation error. One small area of difference was the
3546 weighting scheme employed by the authors to correct for “over-representation” of
3547 meeting attendees in the treatment group. As discussed below, I do not understand
3548 the design reason for this choice and when I simulate its properties, I can obtain
3549 small amounts of bias. The net consequence of their approach was usually to make
3550 coefficient estimates smaller, so we don't have a major difference in conclusion
3551 except perhaps in a secondary analysis of mechanisms.

3552 **Link to Full Report:** <https://osf.io/y8ubt/>

3553 **Link to Replicators' Package:** <https://osf.io/y8ubt/>

3554 **Link to Original Authors' Response:** <https://osf.io/y8ubt/>

3555 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3HACJ)
3556 [persistentId=doi:10.7910/DVN/S3HACJ](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3HACJ)

3557 **12.17.67 Reproduction Report**

3558 **Title Original Study:** Political Turnover, Bureaucratic Turnover, and the Quality
3559 of Public Services

3560 **doi:** <https://doi.org/10.1257/aer.20171867>, American Economic Review

3561 **Report's Abstract:** The politically motivated replacement in local governments
3562 is a pervasive fact in our modern democracies. Whether it has causal effects on
3563 the quality of public services, such as education, is a critical question and yet
3564 understudied. This paper uses a regression discontinuity design (RDD) for close
3565 elections to replicate Akthari, Moreira and Trucco (2022) who find negative effects
3566 on the quality of public education in Brazil (.05-.08 standard deviations of lower test
3567 scores). I first reproduce these main results, finding minor computational differences
3568 that have no effect on the conclusions. I also show that the estimates for Brazil
3569 are in general robust to different specifications following Brodeur, Cook and Heyes
3570 (2020). Finally, I implement the same RDD framework now applied to Chilean
3571 administrative records to find null effects on test scores. Taken together, these
3572 results suggest that political turnover has weakly negative effects on service quality.

3573 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/39.htm>

3574 **Link to Replicators' Package:** <https://osf.io/q43vz/>

3575 **Link to Original Authors' Response:** <https://osf.io/kv4pj/>

3576 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/150323/version/V1/view)
3577 [150323/version/V1/view](https://www.openicpsr.org/openicpsr/project/150323/version/V1/view)

3578 **12.17.68 Reproduction Report**

3579 **Title Original Study:** Pre-Colonial Warfare and Long-Run Development in India
3580 **doi:** <https://doi.org/10.1093/ej/ueab089>, Economic Journal

3581 **Report's Abstract:** We test the reproducibility and replicability of Dincecco et
3582 al. (2022), which reports a positive relationship between pre-colonial interstate
3583 warfare and long-run development patterns across India. Overall, we confirm that
3584 all of the study's estimates are computationally reproducible by using both the
3585 provided replication package in Stata and code written by the present authors in
3586 R. We test for and find no evidence of data manipulation in the final datasets.
3587 Concerning direct replicability, we consider different ways of measuring distance to
3588 conflicts and also alternative proxies for both the dependent variable and variables
3589 which capture channels by which the main effects operate. We are able to replicate
3590 the magnitude and significance of the estimated coefficient on conflict exposure in
3591 most of the tests, noting that while most estimates are substantively in line with
3592 the original study, some alternative measures of distance to conflict imply different
3593 magnitudes for estimates, and proxy estimates are sensitive to both the time period
3594 and type of conflict considered.

3595 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/35.htm>

3596 **Link to Replicators' Package:** <https://osf.io/af6m2/>

3597 **Link to Original Authors' Response:** <https://osf.io/af6m2/>

3598 **Original Authors' Package:** <https://zenodo.org/records/5583263>

3599 **12.17.69 Reproduction Report**

3600 **Title Original Study:** Public Infrastructure and Economic Development: Evi-
3601 dence from Postal Systems

3602 **doi:** <https://doi.org/10.1111/ajps.12594>, American Journal of Political Science

3603 **Report's Abstract:** Rogowski et al. (2022) use secondary data to study the impact
3604 of historic postal infrastructure on economic development, both cross-country and
3605 within the US. Their results suggest a large positive effect of post offices on economic
3606 development that is robust across various sensitivity checks. We successfully com-
3607 putationally reproduce all results. In a robustness assessment, we find the results
3608 to be robust to simple changes in the analysis but observe some sensitivity to
3609 accounting for spatial trends in the cross-country analysis. Additionally, we correct
3610 a coding inconsistency, showing that in the corrected version, one main robustness
3611 check for the US-analysis is no longer supporting the result. Despite this, we find
3612 the results to be overall robust given the numerous analyses and robustness checks
3613 in the original paper.

3614 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/92.htm>

3615 **Link to Replicators' Package:** [https://osf.io/j3ydr/?view_only=](https://osf.io/j3ydr/?view_only=ad14a07cb3a741ca9bbfab391ad7c6fe)
3616 [ad14a07cb3a741ca9bbfab391ad7c6fe](https://osf.io/j3ydr/?view_only=ad14a07cb3a741ca9bbfab391ad7c6fe)

3617 **Original Authors' Response:** "Thanks so much for reproducing the findings in
3618 our paper and exploring extensions of our results. We also appreciate your sharing
3619 the report with us. [...] I [Rogowski] confirm that we are comfortable letting your
3620 report stand and that we will not write a response to it. "

3621 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/33K3EF)
3622 [persistentId=doi:10.7910/DVN/33K3EF](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/33K3EF)

3623 **12.17.70 Reproduction Report**

3624 **Title Original Study:** Re-Assessing Elite-Public Gaps in Political Behavior
3625 **doi:** <https://doi.org/10.1111/ajps.12583>, American Journal of Political Science
3626 **Report's Abstract:** Kertzer (2022) conducts a meta-analysis of parallel experi-
3627 ments on samples of political elites and ordinary citizens. He examines whether the
3628 average treatment effect for elites is significantly different from the average treat-
3629 ment effect for citizens, finding that only 19 of 162 (11.7%) difference-in-difference
3630 estimates are statistically significant after adjusting for the false discovery rate. He
3631 also finds that elites and masses hold similar foreign policy attitudes after control-
3632 ling for their demographic characteristics. In this Reproduction Report, we begin
3633 by running robustness and heterogeneity tests for the first claim. We find that the
3634 results survive many robustness tests. We also find, however, that only a small num-
3635 ber of the these treatments significantly affected masses (N=28) or elites (N=30).
3636 This low rate suggests the possibility that almost all of these experiments failed to
3637 successfully manipulate either masses or elites. If so, we may not be able to con-
3638 clude that masses and elites respond similarly to experiments with confidence until
3639 political scientists produce more experiments with actual treatment effects or with
3640 successful manipulation checks in cases of null effects. In the second part of this
3641 Reproduction Report, we conceptually replicate the second Kertzer analysis, find-
3642 ing a strong correlation between elite and mass political decisions and attitudes,
3643 thus confirming Kertzer's analysis.
3644 **Link to Full Report:** [https://www.econstor.eu/bitstream/10419/266385/1/](https://www.econstor.eu/bitstream/10419/266385/1/I4R-DP010.pdf)
3645 [I4R-DP010.pdf](https://www.econstor.eu/bitstream/10419/266385/1/I4R-DP010.pdf)
3646 **Link to Replicators' Package:** <https://osf.io/93urk/>
3647 **Original Authors' Response:** "Thank you for your email and for the invitation.
3648 [...] please send my appreciation to the authors for their interest in the manuscript;
3649 I find their analysis very interesting."
3650 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LHOTOK)
3651 [persistentId=doi:10.7910/DVN/LHOTOK](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LHOTOK)

3652 **12.17.71 Reproduction Report**

3653 **Title Original Study:** Rebel on the Canal: Disrupted Trade Access and Social
3654 Conflict in China, 1650–1911

3655 **doi:** <https://doi.org/10.1257/aer.20201283>, American Economic Review

3656 **Report's Abstract:** Cao and Chen (2022a) study the role of disruption of trade
3657 on social conflict in China in the 19th century. Identification builds on the closure
3658 of China's Grand Canal in 1826 in a difference-in-differences framework. In their
3659 preferred analytical specification, the authors find that counties along the canal
3660 experienced a 117 percent increase in rebelliousness after the initial closure of the
3661 canal in 1826 relative to their non-canal counterparts. First, we reproduce the
3662 paper's main findings using the official replication package. Second, we examine
3663 whether a sub-sample of counties/prefectures/provinces drives the result. Third,
3664 we test the robustness of the results to alternative treatment periods.

3665 **Link to Full Report:** <https://osf.io/dhn6e/>

3666 **Link to Replicators' Package:** <https://osf.io/dhn6e/>

3667 **Link to Original Authors' Response:** <https://osf.io/dhn6e/>

3668 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/157781/version/V1/view)
3669 [157781/version/V1/view](https://www.openicpsr.org/openicpsr/project/157781/version/V1/view)

3670 **12.17.72 Reproduction Report**

3671 **Title Original Study:** Recessions, Mortality, and Migration Bias: Evidence from
3672 the Lancashire Cotton Famine

3673 **doi:** <https://doi.org/10.1257/app.20190131>, American Economic Journal: Applied
3674 Economics

3675 **Report's Abstract:** Vellore Arthi, Brian Beach and W. Walker Hanlon (2022)
3676 investigate the effect of the Lancashire Cotton Famine on mortality, accounting
3677 for the migration response to the downturn. They use difference-in-differences to
3678 estimate the effect of the cotton famine on mortality. To account for the migration
3679 response to the cotton famine, they construct a linked dataset giving mortality
3680 rates by district of residence during the cotton famine, rather than by district of
3681 residence at the time of death. They find that the cotton famine increased mortality
3682 in cotton-textile producing districts, and that accounting for migration matters,
3683 in the sense that their estimates would have been markedly different had they
3684 not accounted for it. I check that ABH results are fully reproducible using their
3685 data and code, and that their claims are robust to (1) decreasing the age window
3686 for building the linked dataset, (2) modifying the specification and (3) computing
3687 different standard errors. The only significant discrepancy in results is that I find
3688 stronger effects of the cotton famine when I decrease the age window for building
3689 the linked dataset, likely because this reduces measurement errors.

3690 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/25.htm>

3691 **Link to Replicators' Package:** [https://www.openicpsr.org/openicpsr/project/
3692 192272/version/V1/view](https://www.openicpsr.org/openicpsr/project/192272/version/V1/view)

3693 **Original Authors' Response:** "Thanks for the interest in our work. We've had
3694 a chance to review the report and it looks like everything replicated. Since there
3695 are no outstanding queries, we are happy to sign off on this."

3696 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/
3697 128521/version/V1/view](https://www.openicpsr.org/openicpsr/project/128521/version/V1/view)

3698 **12.17.73 Reproduction Report**

3699 **Title Original Study:** Reshaping Adolescents' Gender Attitudes: Evidence from
3700 a School-Based Experiment in India

3701 **doi:** <https://doi.org/10.1257/aer.20201112>, American Economic Review

3702 **Report's Abstract:** Dhar et al. (2022) examine the effect of a gender attitude
3703 change program in secondary schools in India. In their preferred specification, the
3704 authors show that the program made the students report more gender-egalitarian
3705 attitudes by 0.18 of a standard deviation, and shifted self-reported behaviors to
3706 be more aligned with gender-progressive norms by 0.20 standard deviations (both
3707 significant at 1% level). In contrast, they found no effect on girls' aspirations,
3708 as these were already high before the intervention. The effects did not attenuate
3709 between the first end-line (right after the programme was completed) and the second
3710 (two years later). To put the paper's results in perspective, we first comment on
3711 the authors' deviations from their pre-registration and pre-analysis plans, provide
3712 detailed power calculations, and add multiple-hypothesis-testing-adjusted standard
3713 errors. Second, we show that the paper's results are perfectly reproducible. Third,
3714 we show that the results are robust to excluding control variables, and alternative
3715 ways of constructing indices and dealing with non-response.

3716 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/24.htm>

3717 **Link to Replicators' Package:** <https://osf.io/r5jfe/>

3718 **Final Original Authors' Response:** "Thanks. the revision looks good. I actually
3719 don't think we need to have a formal response any more. [...] Thus, I don't think
3720 there is anything substantive for us to include in a discussion paper/response. That
3721 reflects the fact that the Reproduction Reports fair and there is nothing major to
3722 respond to, so it's good news, from both the perspective of the integrity of our
3723 original paper and the professionalism of the replication."

3724 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/149882/version/V1/view)
3725 [149882/version/V1/view](https://www.openicpsr.org/openicpsr/project/149882/version/V1/view)

3726 **12.17.74 Reproduction Report**3727 **Title Original Study:** Run-off Elections in the Laboratory3728 **doi:** <https://doi.org/10.1093/ej/ueab051>, Economic Journal

3729 **Report's Abstract:** Bouton et al. (2022) make a causal claim by manipulating
3730 the voting system under which participants vote (runoff or plurality) and exam-
3731 ining whether this manipulation affects the proportion of strategic voting. They
3732 estimate the effect of the voting system on the proportion of strategic voting for
3733 the participant population, using random effect regression where standard errors
3734 are clustered on group level. Regarding replication results, we reproduced the orig-
3735 inal study's main findings. Our analysis confirms that there are minor and mostly
3736 non-significant disparities in electoral outcomes and voters' welfare between the
3737 two voting systems, consistent with the original study's conclusions. Specifically,
3738 we conducted tests to assess the study's computational reproducibility and direct
3739 replicability. While the authors provided the raw data, they did not include a script
3740 for cleaning it or a codebook describing its contents. Consequently, we developed a
3741 data cleaning script to ensure accurate and consistent data processing.

3742 **Link to Full Report:** <https://osf.io/a8cev/>3743 **Link to Replicators' Package:** [https://github.com/carinahausladen/](https://github.com/carinahausladen/runoff-elections)
3744 [runoff-elections](https://github.com/carinahausladen/runoff-elections)3745 **Original Authors' Response:** The authors provided feedback which was taken
3746 into account.3747 **Original Authors' Package:** <https://doi.org/10.1093/ej/ueab051>

3748 **12.17.75 Reproduction Report**

3749 **Title Original Study:** School Spending and Student Outcomes: Evidence from
 3750 Revenue Limit Elections in Wisconsin

3751 **doi:** <https://doi.org/10.1257/pol.20200226>, American Economic Journal: Economic
 3752 Policy

3753 **Report's Abstract:** Baron (2022) explores the independent effects of operational
 3754 expenditure and capital expenditure on student outcomes in school districts across
 3755 Wisconsin from the outcomes of close referendum approvals. By utilizing a dynamic
 3756 regression discontinuity framework and cubic specification, the author finds that
 3757 narrowly passing an operational referendum, increases operational expenditure per
 3758 pupil by \$298 each year on average, following the referendum over a ten year period.
 3759 From this \$198 are spent on instructional expenses. These point estimates are
 3760 statistically significant at the 10% and 5% level, respectively. We first reproduce
 3761 the main results from the paper without any issues arising. Secondly, we conduct
 3762 a robustness replicability to (1) dropping school districts from the top and bottom
 3763 5% of the revenue limits distribution, categorically, and (2) dividing the time frame
 3764 of the study into two periods: 1996-2005 and 2005-2014. We find that dropping the
 3765 top 5% of the school districts by revenue limits reduces the additional operational
 3766 expenditure by \$140 per pupil (lower by 50 percent) and the effects of passing an
 3767 operational referendum were nearly double in the former period compared to the
 3768 latter period. Lastly, we find that the estimated effects on student outcomes rely
 3769 heavily on recent observations.

3770 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/88.htm>

3771 **Link to Replicators' Package:** <https://osf.io/m2w4x/>

3772 **Original Author's Response:** "Thank you for sharing the Reproduction Report.
 3773 Please pass on my thanks to the replicators for their important work. First and
 3774 foremost, I'm glad to see that the results in the paper are reproducible without any
 3775 issues arising. The report explores two additional sources of heterogeneity. I have
 3776 no additional comments on these. I do briefly want to clarify the last sentence in
 3777 the report's abstract, which reads "Lastly, we find that the estimated effects on
 3778 student outcomes rely heavily on recent observations." While I am not entirely sure
 3779 what the replicators are referring to, my guess is that they refer to Table 2 in the
 3780 report. In this table, they discuss that they are unable to study heterogeneity in
 3781 the impacts of passing a referendum on test scores and postsecondary enrollment
 3782 from 1996-2005, because data on these outcomes are unavailable prior to 2005. The
 3783 availability of each dataset was discussed in the published version of the paper (see,
 3784 for example, Table 1). Perhaps a more accurate statement would be to explain
 3785 that the replicators couldn't explore the impact of passing a referendum on these
 3786 specific outcomes in the early period due to data constraints—and that this was
 3787 acknowledged in the published version."

3788 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/125821/version/V1/view)
 3789 [125821/version/V1/view](https://www.openicpsr.org/openicpsr/project/125821/version/V1/view)

3790 **12.17.76 Reproduction Report**

3791 **Title Original Study:** Social Class and (Un)Ethical Behaviour: Causal and
3792 Correlational Evidence

3793 **doi:** <https://doi.org/10.1093/ej/ueac022>, Economic Journal

3794 **Report's Abstract:** The relationship between social status and ethical behav-
3795 ior is a widely debated topic in research. In their study, Gsottbauer et al. (2022b)
3796 investigate whether higher socio-economic status is linked to lower ethical behavior,
3797 using data from two large survey experiments involving over 11,000 participants.
3798 In this replication project, we test the computational reproducibility and robust-
3799 ness to the replication of their study, using the provided data and code from the
3800 replication package (Gsottbauer et al., 2022a). Nearly all the figures and tables
3801 were reproducible-in the process of reproducing the results, some minor rounding or
3802 transcription errors were discovered. In testing the robustness replicability, we find
3803 consistent results for our extensions. The effort for the replication was manageable,
3804 even though the authors treat categorical variables as numeric, or use manually-
3805 coded interaction variables (i.e., in regression models). In summary, we applaud
3806 the transparency of Gsottbauer et al. (2022b) in facilitating replications, and make
3807 some general recommendations for further improvements for data-analysis studies.

3808 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/29.htm>

3809 **Link to Replicators' Package:** [https://github.com/ha0ye/](https://github.com/ha0ye/replication-gsottbauer-2022)
3810 [replication-gsottbauer-2022](https://github.com/ha0ye/replication-gsottbauer-2022)

3811 **Original Authors' Response:** Declined to respond.

3812 **Original Authors' Package:** <https://zenodo.org/records/6226207>

3813 **12.17.77 Reproduction Report**

3814 **Title Original Study:** Sorting or Steering: The Effects of Housing Discrimination
3815 on Neighborhood Choice

3816 **doi:** <https://doi.org/10.1086/720140>, Journal of Political Economy

3817 **Report's Abstract:** This comment revisits the analysis in Christensen and Tim-
3818 mins (2022). We identify two critical errors used in the original analysis, one with
3819 the data and the other with coding. When either error is corrected several major
3820 results in the paper change, either in statistical significance or in effect size. The
3821 data error is a result of including fixed effects for the string variable 'city'. The raw
3822 variable is case sensitive and has many spelling mistakes. The coding error involves
3823 assigning a value of zero for the variable "of color" to both individuals identified as
3824 'white' and as 'other' in the raw data. The level of clustering in the paper is also
3825 arguably too fine. Many of the results are not robust to clustering at the city level,
3826 as opposed to the subject pair level. In total, we affirm the authors' overarching
3827 claim of substantial and nuanced housing discrimination against racial minorities
3828 generally, and African Americans in particular; however, the effect sizes and sig-
3829 nificance are generally (although not always) smaller than the original authors
3830 findings. Additionally, there are several instances where the effects of discrimina-
3831 tion on African Americans are no longer statistically significant but the effect of
3832 discrimination on Hispanics becomes significant.

3833 **Link to Full Report:** <https://osf.io/vwgxd/>

3834 **Link to Replicators' Package:** <https://github.com/mattdwebb/HUDreplication>

3835 **Original Authors' Response:** Authors mentioned that they are currently writing
3836 a response.

3837 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/720140/suppl_file/20191181data.zip)
3838 [1086/720140/suppl_file/20191181data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/720140/suppl_file/20191181data.zip)

3839 **12.17.78 Reproduction Report**

3840 **Title Original Study:** Spillover Effects of Intellectual Property Protection in the
3841 Interwar Aircraft Industry

3842 **doi:** <https://doi.org/10.1093/ej/ueab091>, Economic Journal

3843 **Report's Abstract:** We are attempting to reproduce the results of Hanlon and
3844 Jaworski (2022) based on their dataset. Our work is conducted in two different ways:
3845 (i) computational reproducibility, aiming to produce the same results using different
3846 software, notably R, with the given data; and (ii) checking the robustness of the
3847 results. For (i), the estimated coefficients are consistent based on the R software.
3848 For (ii), we carefully examine the given datasets of Hanlon and Jaworski (2022)
3849 and review the economic history of the US Interwar aircraft industry between 1918
3850 and 1935 to identify potential confounding variables (apart from IPP strengthening
3851 in the year 1926) that might affect both the controls and error term, and thus the
3852 results. We identify some confounding variables that may affect the results and
3853 attempt to illustrate them before and after 1926 when IPP is strengthened. Overall,
3854 we find that the results are replicable by utilizing the codes and datasets of Hanlon
3855 and Jaworski (2022), which is encouraging.

3856 **Link to Full Report:** <https://osf.io/t4avf/>

3857 **Link to Replicators' Package:** <https://osf.io/t4avf/>

3858 **Link to Original Authors' Response:** <https://osf.io/t4avf/>

3859 **Original Authors' Package:** <https://zenodo.org/records/5627298>

3860 12.17.79 Reproduction Report

3861 **Title Original Study:** State Action to Prevent Violence against Women: The
3862 Effect of Women's Police Stations on Men's Attitudes toward Gender-Based
3863 Violence

3864 **doi:** <https://doi.org/10.1086/714931>, Journal of Politics

3865 **Report's Abstract:** Córdoba and Kras (2022) examine how the existence of a
3866 women's police station (WPS) in the place of residence influences citizens' atti-
3867 tudes toward gender-based violence in Brazil. In their analytical specification, the
3868 authors find that men are more likely to reject violence against women (VAW)
3869 and support bystander intervention in municipalities with a WPS, especially if the
3870 WPS has been operating for a long time. This paper examines the replicability
3871 and robustness of Córdoba & Kras' (2022) findings. First, we reproduce the paper's
3872 main findings and uncover one minor coding error and three estimates that have
3873 been reported with the opposite sign compared to that in our reproduction; neither
3874 is of consequence for the study's main results. Second, we test the robustness of
3875 the results by (1) recoding one of the main explanatory variables and several of the
3876 control variables to account for non-linear trends, (2) using alternative techniques
3877 to estimate clustered standard errors, (3) consistently applying a 95% confidence
3878 level in the presentation of the results, (4) altering the propensity score match-
3879 ing (PSM) procedure as well as the composition of the variables used in the PSM
3880 robustness check, (5) using an alternative technique to test for multicollinearity,
3881 (6) excluding potential endogenous control variables, and (7) using an alternative
3882 coding for computing margins. Reassuringly, the results are robust to most of these
3883 tests. However, two of the robustness checks challenge parts of the paper's main
3884 findings. First, allowing for non-linearity in the effect of time since the establish-
3885 ment of WPS shows (a) a non-linear effect on VAW and (b) no apparent changes in
3886 either male or female attitudes over time once the WPS has been established. Sec-
3887 ond, the inclusion of other variables in the PSM procedure renders part of the main
3888 estimates of interest statistically nonsignificant ($p < 0.1$). Our findings highlight
3889 the need for further re-analyses and replications for investigating the preventive
3890 effects of women's police stations.

3891 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/67.htm>

3892 **Link to Replicators' Package:** <https://osf.io/yjwr8/>

3893 **Link to Original Authors' Response:** Responded to our emails but no formal
3894 response as of February 2024.

3895 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/D2WL5I)
3896 [persistentId=doi:10.7910/DVN/D2WL5I](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/D2WL5I)

3897 **12.17.80 Reproduction Report**

3898 **Title Original Study:** Student Performance, Peer Effects, and Friend Networks:
3899 Evidence from a Randomized Peer Intervention

3900 **doi:** <https://doi.org/10.1257/pol.20200563>, American Economic Journal: Economic
3901 Policy

3902 **Report's Abstract:** Wu et al. (2023) estimate the effect of classroom seat-
3903 ing arrangements in China using a randomized control trial with two treatment
3904 schemes. The first treatment scheme involves seating high and low achieving stu-
3905 dents together, and the second treatment involves this same seating arrangement
3906 with financial incentives for the high-achieving students, if their deskmates' test
3907 scores improved. All statistically significant impacts come from the incentivized
3908 treatment scheme. Wu et al. (2023) find that low-achieving students sitting next
3909 to incentivized high-achieving students perform 0.24 SD (p-value=0.018) better
3910 on math exams. In addition, being assigned to the incentive treatment scheme
3911 increased extraversion and agreeableness for low and high achieving students.
3912 Lastly, they do not find much evidence of peer effects on test scores nor personality
3913 traits. This study is computationally reproducible using their provided replication
3914 package. We ran their code using Stata 14, 17, and 18. After running their replica-
3915 tion package, we further investigated Tables 2-5. The main conclusions are generally
3916 robust to various coding decisions. Notably, in investigating the peer effects, when
3917 we change the specification to also control for the difference in baseline scores
3918 between the student and their deskmate, we find that the more dissimilar deskmates
3919 are at baseline, the bigger the peer effects.

3920 **Link to Full Report:** <https://osf.io/9hx3b/>

3921 **Original Authors' Response:** The authors provided feedback which was taken
3922 into account.

3923 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/149262/version/V2/view)
3924 [149262/version/V2/view](https://www.openicpsr.org/openicpsr/project/149262/version/V2/view)

12.17.81 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.”

3970 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3M5AX)
3971 [persistentId=doi:10.7910/DVN/S3M5AX](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/S3M5AX)

3972 **12.17.82 Reproduction Report**

3973 **Title Original Study:** Targeting High Ability Entrepreneurs Using Community
3974 Information: Mechanism Design in the Field

3975 **doi:** <https://doi.org/10.1257/aer.20200751>, American Economic Review

3976 **Report's Abstract:** Hussam et al. (2022a) use a cash grant experiment in India
3977 to demonstrate that community knowledge can help target high-growth microen-
3978 trepreneurs. In their preferred specification, the authors find that the average
3979 marginal return to the grant is 9.4 percent per month, while estimated returns
3980 for entrepreneurs reported by peers to be in the top third of the community are
3981 between 24 percent and 30 percent. First, we reproduce the paper's main findings
3982 and uncover one minor coding error, which affects the estimates for one of the main
3983 tables but does not change the overall conclusions of the paper. Second, we test
3984 the robustness of the results to: (1) different treatment of outliers, (2) dropping
3985 surveyor and survey month fixed effects, and (3) using quartiles instead of terciles
3986 for grouping the ranking of entrepreneurs. The paper's results are robust to these
3987 robustness checks. Finally, we test heterogeneity of results by gender, which was
3988 not reported in the original study.

3989 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/49.htm>

3990 **Link to Replicators' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DI7RR9)
3991 [persistentId=doi:10.7910/DVN/DI7RR9](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DI7RR9)

3992 **Link to Original Authors' Response:** "We are very grateful to Isabella Masetto,
3993 Diego Ubfal, and to the team at I4R for their excellent work. We verified the coding
3994 error and we agree that it did not meaningfully alter the conclusion of our paper
3995 that community information is informative over and above the predictive power of
3996 observable characteristics. We will post a link to this correction on our websites
3997 and will consult the editors of the AER as to whether this error rises to the level
3998 of requiring a formal correction."

3999 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/151841/version/V1/view)
4000 [151841/version/V1/view](https://www.openicpsr.org/openicpsr/project/151841/version/V1/view)

4001 **12.17.83 Reproduction Report**

4002 **Title Original Study:** Teaching Norms: Direct Evidence of Parental Transmission
4003 doi: <https://doi.org/10.1093/ej/ueac074>, Economic Journal

4004 **Report's Abstract:** This paper is a replication study of Brouwer, T., Galeotti,
4005 F., & Villeval, M. C. (2023), using the original data. The study explores how social
4006 norms are transmitted from one generation to another, specifically from parents to
4007 children. The authors conducted a field experiment involving 601 parents of children
4008 aged 3 to 12 in Lyon, France, to examine whether parents engage more in norm
4009 enforcement in the presence of their child, and whether the nature of punishment
4010 changes in the presence of the child. The study found that parents do engage more
4011 in norm enforcement in the presence of their child, and tend to use more indirect
4012 punishment when their child is present. This study highlights the role that parents
4013 play in transmitting social norms to their children. The replication analysis was
4014 successful, with the results of the original study being robust to changes in the
4015 model specification.

4016 **Link to Full Report:** <https://osf.io/qnbfa/>

4017 **Link to Replicators' Package:** <https://zenodo.org/records/8114738>

4018 **Original Authors' Response:** The replicators took into account the authors'
4019 feedback. They wrote at the end of the back and forth: "We thank you and the
4020 replication team for the replication and the successive interactions. We created an
4021 OSF project including the data replication package enabling the reproduction of
4022 the analysis presented in our article. The package comprises a source file (in Stata
4023 format and in TXT) and a Stata do-file that allows the reconstruction of the master
4024 file used in the replication package submitted to the Economic Journal."

4025 **Original Authors' Package:** <https://zenodo.org/records/7045559>

4026 **12.17.84 Reproduction Report**4027 **Title Original Study:** Technological Change and the Consequences of Job Loss4028 **doi:** <https://doi.org/110.1257/aer.20210182>, American Economic Review

4029 **Report's Abstract:** Braxton and Taska (2023) find that technological change
4030 accounts for 45 percent of the decline in earnings after job loss. We first reproduce
4031 all regression tables in Braxton and Taska (2023), and then test for robustness by
4032 controlling for the initial level of wages, additional fixed effects, multi-way cluster-
4033 ing, and conducting influential analysis. We find that the paper's original results are
4034 sensitive to controlling for initial wages and some additional fixed effects. Overall,
4035 we find the results are robust with a coefficient in the same direction and signifi-
4036 cant at 5% in 33% of the robustness checks we ran, with average t/z scores 28% as
4037 large as the original study.

4038 **Link to Full Report:** <https://osf.io/qws2p/>4039 **Link to Replicators' Package:** <https://osf.io/qws2p/>4040 **Original Authors' Response:** Did not get a response as of November 2025.4041 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/181166/version/V1/view)
4042 [181166/version/V1/view](https://www.openicpsr.org/openicpsr/project/181166/version/V1/view)

4043 **12.17.85 Reproduction Report**4044 **Title Original Study:** The Common-Probability Auction Puzzle4045 **doi:** <https://doi.org/10.1257/aer.20191927>, American Economic Review

4046 **Report's Abstract:** Ngangoué and Schotter (2023) investigate common-
4047 probability auctions. By running an experiment, they find that, in contrast to the
4048 substantial overbidding found in common-value auctions, bidding in strategically
4049 equivalent common-probability auctions is consistent with the Nash equilibrium.
4050 We reproduce their results in R, conduct robustness checks on how their sample
4051 was constructed, and consider possible heterogeneity. We confirm their documented
4052 qualitative results.

4053 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/74.htm>4054 **Link to Replicators' Package:** <https://osf.io/7bq4s/>

4055 **Original Authors' Response:** "Thank you for putting the effort in replicating
4056 our study! Your results are also quite interesting to us as we haven't thought of
4057 all the robustness checks you've made. At this point, we do not have any major
4058 comments to make and refrain from submitting a response."

4059 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/
4060 184041/version/V1/view](https://www.openicpsr.org/openicpsr/project/184041/version/V1/view)

4061 **12.17.86 Reproduction Report**

4062 **Title Original Study:** The Curious Case of Theresa May and the Public That
4063 Did Not Rally: Gendered Reactions to Terrorist Attacks Can Cause Slumps Not
4064 Bumps

4065 **doi:** <https://doi.org/10.1017/S0003055421000861>, American Political Science
4066 Review

4067 **Report's Abstract:** Holman et al. (2022; HMZ) propose women (compared to
4068 men) political leaders experience significant drops in public approval ratings after
4069 a transnational terrorist attack. After documenting how survey-based evaluations
4070 of then-Prime Minister Theresa May suffered after the 2017 Manchester Arena
4071 attack, HMZ assemble a country-quarter level panel database to explore the gener-
4072 ality of their hypothesis. They report evidence suggesting women (compared to
4073 men) leaders systematically experience decreased public approval rates after major
4074 transnational terrorist attacks (p-value of 0.020). We find that result disappears
4075 once any of the following adjustments is implemented: (i) excluding election quarter
4076 covariates ($p = 0.104$); (ii) correcting objective coding errors in the election quarter
4077 covariates ($p = 0.058$); (iii) excluding the May-Manchester observation ($p = 0.098$);
4078 or (iv) clustering standard errors at the country level ($p = 0.558$). Exploring all 2^5
4079 combinations of the five control groups HMZ incorporate in their specification, none
4080 of them clears the 5% threshold of statistical significance once the corrected elec-
4081 tion quarter variables are employed. We conclude that the empirical evidence does
4082 not provide sufficient support for HMZ's abstract claim that "conventional theory
4083 on rally events requires revision: women leaders cannot count on rallies following
4084 major terrorist attacks."

4085 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/41.htm>

4086 **Link to Replicators' Package:** <https://doi.org/10.5683/SP3/6SYCML>

4087 **Link to Original Authors' Response:** [https://econpapers.repec.org/paper/](https://econpapers.repec.org/paper/zbwi4rdps/44.htm)
4088 [zbwi4rdps/44.htm](https://econpapers.repec.org/paper/zbwi4rdps/44.htm)

4089 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VHNPUO)
4090 [persistentId=doi:10.7910/DVN/VHNPUO](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VHNPUO)

4091 **12.17.87 Reproduction Report**

4092 **Title Original Study:** The Dynamics and Spillovers of Management Interventions: Evidence from the Training within Industry Program

4093 **doi:** <https://doi.org/10.1086/719277>, Journal of Political Economy

4094 **Report's Abstract:** Bianchi and Giorcelli (2022) study the long-term and spillover
 4095 effects of a management intervention program on firm performance in the US,
 4096 between 1940 and 1945. The authors find that the Training Within Industry (TWI)
 4097 program led to positive effects which lasted for at least 10 years. Firm sales of
 4098 treated firms increased by 5.3% in the first year after implementation, peaking at
 4099 21.7% after 8 years, before reducing to 16% gains after a decade. The authors claim
 4100 that the program generated long-lasting changes in managerial practices. Finally,
 4101 the program also led to positive spillover effects on the supply chain of treated
 4102 firms. First, we reproduce the paper's main findings. Second, we test the robustness
 4103 of the results to (1) changing the main specification sample and (2) testing other
 4104 difference-in-differences estimators, using the same data, provided by the authors.
 4105 We find that the results are robust to these changes. All point estimates in the
 4106 study remain statistically significant and of similar magnitude. While the paper's
 4107 finding reproduce and replicate, challenges in reproducing results we encountered
 4108 lead us to recommend improvements to journals' code policies.

4109 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/66.htm>

4110 **Link to Replicators' Package:** https://github.com/cwestheide/i4r_dp66_code

4111 **Original Authors' Final Response:** "Thanks a lot for sharing the updated
 4112 report with us. We don't have anything to add at this point."

4113 **Original Authors' Package:** https://www.journals.uchicago.edu/doi/suppl/10.1086/719277/suppl_file/20200781data.zip

4116 **12.17.88 Reproduction Report**

4117 **Title Original Study:** The Economic Effects of Long-Term Climate Change:
4118 Evidence from the Little Ice Age

4119 **doi:** <https://doi.org/10.1086/720393>, Journal of Political Economy

4120 **Report's Abstract:** Waldinger (2022) finds significant negative economic effects
4121 (proxied by city size) from gradual climate change which occurred during the Little
4122 Ice Age (1600-1850) and offers two potential mechanisms (agricultural productivity
4123 and mortality) and two potential adaptations (trade and land use). In this comment,
4124 we show that while Waldinger (2022)'s findings can be replicated, the main result
4125 relies on a critical author assumption: Cities with 0 inhabitants in the original data
4126 are instead assumed to have 500. This assumption affects 23.5% of observations and
4127 49.6% of cities in the sample. When these "missing data" are excluded from the
4128 analysis, the effect estimated by otherwise identical research methods is of similar
4129 magnitude and statistical significance but of opposite sign.

4130 **Link to Full Report:** <https://osf.io/tmn2j/>

4131 **Link to Replicators' Package:** <https://osf.io/tmn2j/>

4132 **Link to Original Authors' Response:** <https://osf.io/tmn2j/>

4133 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/720393/suppl_file/2015548data.zip)
4134 [1086/720393/suppl_file/2015548data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/720393/suppl_file/2015548data.zip)

4135 **12.17.89 Reproduction Report**

4136 **Title Original Study:** The Effects of Banking Competition on Growth and
4137 Financial Stability: Evidence from the National Banking Era

4138 **doi:** <https://doi.org/10.1086/717453>, Journal of Political Economy

4139 **Report's Abstract:** Carlson et al. (2022) examine the causal impact of banking
4140 competition by investigating a unique circumstance in the National Banking Era
4141 of the nineteenth century in the US, where a discontinuity in bank capital require-
4142 ments occurred. On the one hand, their findings suggest that banks operating in
4143 markets with fewer barriers to entry tend to increase their lending activities, pro-
4144 moting real economic growth. On the other hand, banks in less restricted markets
4145 also exhibit a higher propensity for risk-taking, posing risks to financial stability.
4146 First, we fully reproduce the paper's outcomes apart from a minor discrepancy in
4147 the estimate of Table 9 attributed to issues in the provided codes. Second, we test
4148 the robustness of the results by (i) changing the ranges used to select the sample
4149 of cities included in the analysis, (ii) adopting different options to address outliers'
4150 potential issues and (iii) introducing additional control variables. We observe that
4151 the estimation results remain mostly consistent when subjecting them to various
4152 robustness checks. However, it is worth highlighting that the results can be par-
4153 tially influenced by the criteria used to select the sample of cities and the inclusion
4154 of control variables.

4155 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/81.htm>

4156 **Link to Replicators' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BB864R)
4157 [persistentId=doi:10.7910/DVN/BB864R](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BB864R)

4158 **Original Authors' Final Response:** "We thank the replication team (Andrea
4159 Calef, Sya In Chzhen, Marco Mandas, and Fabio Motoki) for the detailed Reproduc-
4160 tion Report. We are glad to hear that the replicating team affirms the robustness of
4161 the paper's findings. We are also glad that the replicators were able to successfully
4162 replicate all tables and figures. We thank the replicating team for identifying vari-
4163 ous smaller issues regarding the code structure which fortunately did not affect our
4164 original findings. We believe that the report as such does not require us to respond
4165 in any further detail. We highly appreciate the effort of both the replicating team
4166 and the I4R."

4167 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/717453/suppl_file/20200610data.zip)
4168 [1086/717453/suppl_file/20200610data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/717453/suppl_file/20200610data.zip)

4169 **12.17.90 Reproduction Report**

4170 **Title Original Study:** The Geography of Repression and Opposition to Autocracy
 4171 **doi:** <https://doi.org/10.1111/ajps.12614>, American Journal of Political Science

4172 **Report's Abstract:** Analytic data sets and analysis code are available and they
 4173 produce the same results as presented in the paper (CRA). Robustness checks
 4174 involve the (i) use of matching estimators to address possible bias from misspec-
 4175 ification, based on propensity score estimated from a random forest model, (ii)
 4176 doubly robust (TMLE) estimation to address possible bias from misspecification
 4177 in either the propensity score or outcome regression stages, using a super learner
 4178 ensemble with random forest, GAM, mean, and non-parametric regression models
 4179 and averaged over repeated runs to minimize randomness, (iii) define treated comu-
 4180 nas as those within a fixed physical distance radius of the nearest military base,
 4181 rather than only those that contain it, and (iv) instead of using 2SLS to assess the
 4182 causally mediated effect of military bases on plebiscite outcomes via repression, we
 4183 propose to conduct mediation analysis (Tingley et al 2013), implemented in the R
 4184 'mediation' package.

4185 **Link to Full Report:** [https://www.socialsciencereproduction.org/reproductions/](https://www.socialsciencereproduction.org/reproductions/789/published/index?step=4)
 4186 [789/published/index?step=4](https://www.socialsciencereproduction.org/reproductions/789/published/index?step=4)

4187 **Link to Replicators' Package:** [https://github.com/pjesscarter/](https://github.com/pjesscarter/repression-replication)
 4188 [repression-replication](https://github.com/pjesscarter/repression-replication)

4189 **Link to Original Authors' Response:** We are happy that the replicators suc-
 4190 cessfully reproduced all the analysis in our published paper. Moreover, additional
 4191 robustness checks within the quantitative framework of the paper further confirm
 4192 the empirical results. Two extensions using propensity score matching give some-
 4193 what different results. Unfortunately, these additional estimators violate standard
 4194 requirements for credible matching designs, i.e., overlap in the propensity score dis-
 4195 tribution across treatment and control groups. As shown by previous research, this
 4196 lack of overlap leads to unstable estimators with variance that may explode in finite
 4197 samples such as ours (Frölich 2004, Khan and Tamer 2010). In another extension,
 4198 the replicators employ a mediation analysis to re-interpret the empirical evidence
 4199 in our paper. To support the use of our method, i.e., instrumental variables, we
 4200 rule out alternative explanations and provide a range of historical evidence. With-
 4201 out historical and contextual support for alternative assumptions, we believe that
 4202 the method used by the replicators is hard to interpret.

4203 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EYAWES)
 4204 [persistentId=doi:10.7910/DVN/EYAWES](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EYAWES)

4205 **12.17.91 Reproduction Report**

4206 **Title Original Study:** The Labor Market Impacts of Universal and Permanent
4207 Cash Transfers: Evidence from the Alaska Permanent Fund

4208 **doi:** <https://doi.org/10.1257/pol.20190299>, American Economic Journal: Economic
4209 Policy

4210 **Report's Abstract:** Jones and Marinescu (2022) study the employment effects
4211 of a universal cash transfer in Alaska. Using a synthetic control method, they find
4212 that the transfer had no negative effects on employment. We reproduce the results
4213 using their replication package and investigate if the results hold when using a
4214 different software to run the analysis. We also use different estimation techniques
4215 and perform sensitivity checks to assess robustness of the results. We find some
4216 differences in the size and significance of the average treatment effects on labor force
4217 participation and hours worked when we use a different software (R) and various
4218 extensions of the synthetic control method. We also find smaller coefficients on
4219 part-time employment when including more covariates. However, these differences
4220 do not contradict the main conclusion of the paper.

4221 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/80.htm>

4222 **Link to Replicators' Package:** <https://osf.io/6atfw/>

4223 **Original Authors' Final Response:** "Thanks for putting in all this effort to
4224 evaluate the robustness of our results! I [Marinescu] think this is really a worthwhile
4225 endeavor."

4226 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/140121/version/V1/view)
4227 [140121/version/V1/view](https://www.openicpsr.org/openicpsr/project/140121/version/V1/view)

4228 **12.17.92 Reproduction Report**

4229 **Title Original Study:** The Long-Run Effects of Sports Club Vouchers for Primary
4230 School Children

4231 **doi:** <https://doi.org/10.1257/pol.20200431>, American Economic Journal: Economic
4232 Policy

4233 **Report's Abstract:** Marcus, Siedler and Ziebarth (2022 American Economic
4234 Journal: Economic Policy) examine the long-run health effects of a universal sports-
4235 club voucher program that was introduced in Saxony for primary school children
4236 in 2009. In 2018, the authors designed a survey that targeted the affected cohorts
4237 and nearby cohorts in Saxony and two neighboring states, and use a differences-in-
4238 differences identification strategy that exploits variation across states and cohorts
4239 in policy exposure. The authors document that treated individuals have knowledge
4240 of the program and recall receiving and redeeming the vouchers at higher rates,
4241 but find no effects on any health outcomes or behaviors. We successfully reproduce
4242 the main results of the paper exactly using data available in the paper's replication
4243 package and new Stata and R code. We also verify the robustness of the results
4244 using different outcomes, different control variables, different sample restrictions
4245 and different inference methods.

4246 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/46.htm>

4247 **Link to Replicators' Package:** <https://osf.io/4bnjt/>

4248 **Original Authors' Response:** "We would like to thank the authors for their
4249 interest in our paper. We greatly appreciate their careful reading of the paper
4250 and the insightful robustness exercises they conducted. We are pleased that our
4251 results were successfully reproduced using different software packages, and that the
4252 additional robustness analyses performed by the authors further strengthen and
4253 support our conclusions."

4254 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/138922/version/V1/view)
4255 [138922/version/V1/view](https://www.openicpsr.org/openicpsr/project/138922/version/V1/view)

4256 **12.17.93 Reproduction Report**

4257 **Title Original Study:** The Long-Term Effects of Measles Vaccination on Earnings
4258 and Employment

4259 **doi:** <https://doi.org/10.1257/pol.20190509>, American Economic Journal: Economic
4260 Policy

4261 **Report's Abstract:** Atwood (2022) analyzes the effects of the 1963 U.S. measles
4262 vaccination on longrun labor market outcomes, using a generalized difference-in-
4263 differences approach. We reproduce the results of this paper and perform a battery
4264 of robustness checks. Overall, we confirm that the measles vaccination had positive
4265 labor market effects. While the negative effect on the likelihood of living in poverty
4266 and the positive effect on the probability of being employed are very robust across
4267 the different specifications, the headline estimate-the effect on earnings-is more
4268 sensitive to the exclusion of certain regions and survey years.

4269 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/33.htm>

4270 **Link to Replicators' Package:** <https://osf.io/jv7kx/>

4271 **Link to Original Authors' Response:** <https://osf.io/qxjnk/>

4272 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/
4273 138401/version/V1/view](https://www.openicpsr.org/openicpsr/project/138401/version/V1/view)

4274 **12.17.94 Reproduction Report**

4275 **Title Original Study:** The Macroeconomics of Sticky Prices with Generalized
4276 Hazard Functions

4277 **doi:** <https://doi.org/10.1093/qje/qjab042>, Quarterly Journal of Economics

4278 **Report's Abstract:** We replicate the empirical results in Section 4 of Alvarez et
4279 al. (2022). First, we were able to reproduce the original authors' major empirical
4280 results, but only after editing the program for it to run on our computing platform.
4281 There are small discrepancies in the empirical estimates, e.g. bootstrapped standard
4282 errors, that involve the use of simulations. Second, we replicated Alvarez et al.'s
4283 results by adopting the data cleaning criteria used by their original data source
4284 (Cavallo 2018) to evaluate its robustness to data handling procedures. We found
4285 noticeable changes in the empirical results that can have important implications on
4286 the effects of monetary policy. To conclude, we propose using Docker container to
4287 promote research reproducibility, and more attention is needed on data handling
4288 to improve the robustness of empirical research.

4289 **Link to Full Report:** [https://github.com/atyho/
4290 Ottawa-Replication-Games-2023/blob/main/Ho_Huynh_Rea_Replication_Report.
4291 pdf](https://github.com/atyho/Ottawa-Replication-Games-2023/blob/main/Ho_Huynh_Rea_Replication_Report.pdf)

4292 **Link to Replicators' Package:** [https://github.com/atyho/
4293 Ottawa-Replication-Games-2023/](https://github.com/atyho/Ottawa-Replication-Games-2023/)

4294 **Link to Original Authors' Response:**

4295 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?
4296 persistentId=doi:10.7910/DVN/IBM0IE](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IBM0IE)

4297 **12.17.95 Reproduction Report**

4298 **Title Original Study:** The Morning After: Cabinet Instability and the Purging
4299 of Ministers after Failed Coup Attempts in Autocracies

4300 **doi:** <https://doi.org/10.1086/716952>, Journal of Politics

4301 **Report's Abstract:** We replicate the analysis provided in Bokobza et al. (2022).
4302 They identify a causal effect of failed coup attempts on cabinet minister removals
4303 in autocracies on both the country and individual minister level and show that
4304 higher-ranking ministers and those holding strategic positions are more likely to
4305 be purged than more loyal and veteran ministers using fixed effects panel models.
4306 We focus on computational reproducibility and robustness replicability. In addition
4307 to reproducing the original results using Stata and R, we replicate analyses
4308 using random effects panel models and ordered beta regression models, reproduced
4309 analyses performed in R using different packages, replaced the main independent
4310 variable, clustered standard errors on a different level, and added independent variables
4311 related to coup-proofing. We find that the original findings were reproducible
4312 and robust.

4313 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/45.htm>

4314 **Link to Replicators' Package:** <https://doi.org/10.7910/DVN/21HZCC>

4315 **Link to Original Authors' Response:** <https://osf.io/sm526/>

4316 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GCDJ25)
4317 [persistentId=doi:10.7910/DVN/GCDJ25](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GCDJ25)

4318 12.17.96 Reproduction Report

4319 **Title Original Study:** The Origin of the State: Land Productivity or Appropri-
4320 ability?

4321 **doi:** <https://doi.org/10.1086/718372>, Journal of Political Economy

4322 **Report's Abstract:** This is a replication of Mayshar et al. (2022) (MPP). The
4323 article posits that the state (defined as societal hierarchy such as tax-levying elites)
4324 originated from cultivation of appropriable cereal grains, contrary to the conven-
4325 tional theory that the state originated from increased land productivity following
4326 the adoption of agriculture. The article uses multiple datasets to demonstrate a
4327 causal effect of cereal cultivation on hierarchy (Claim 1) without finding a similar
4328 effect for land productivity (Claim 2), and that societies based on roots or tubers
4329 display levels of hierarchy similar to nonfarming societies (Claim 3). The results of
4330 our replication in brief are: 1. The data and code provided by MMP closely repro-
4331 duce the main results presented in their Table 1 (see our Table 1). 2. Concurrently
4332 testing the cereal cultivation and land productivity claims leads to slightly less sta-
4333 tistical significance, on average, than the published article (Table 2). 3. Removing
4334 the inherited 1-5 scale of the dependent variable (hierarchy) finds that cereal pro-
4335 duction is not as effective at moving across all levels of hierarchy compared to the
4336 more general claim (Table 3 and 4). 4. Using the same procedures with an aim to
4337 confirm the conventional hypothesis (land productivity leads to increased hierarchy
4338 conditional on cereal growth) offers statistically significant evidence both for and
4339 against Claims 1 and 2 and against Claim 3 (Table 6). 5. The statistical significance
4340 of Claim 1 is sensitive to the removal of the top 3% of observations outliers (Table
4341 7). 6. Correction of mis-classified 'none or none specified' crop societies alters the
4342 interpretation of coefficients behind Claim 3. Societies that rely more on agricul-
4343 ture among farming societies experience more complex hierarchies, irrespective of
4344 being primarily cereal producing or tubers growing (Table 8 and 9). (...)

4345 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/82.htm>

4346 **Link to Replicators' Package:** <https://osf.io/ekzdg/>

4347 **Original Authors' Response:** Comments taken into account in the report.

4348 **Original Authors' Package:** [https://www.journals.uchicago.edu/doi/suppl/10.](https://www.journals.uchicago.edu/doi/suppl/10.1086/718372/suppl_file/2018030data.zip)
4349 [1086/718372/suppl_file/2018030data.zip](https://www.journals.uchicago.edu/doi/suppl/10.1086/718372/suppl_file/2018030data.zip)

4350 **12.17.97 Reproduction Report**

4351 **Title Original Study:** The Power of Hydroelectric Dams: Historical Evidence
4352 from the United States over the Twentieth Century

4353 **doi:** <https://doi.org/10.1093/ej/ueac059>, Economic Journal

4354 **Report's Abstract:** Successful computational reproducibility. No coding errors
4355 uncovered.

4356 **Original Authors' Package:** <https://doi.org/10.1093/ej/ueac059>

4357 **12.17.98 Reproduction Report**

4358 **Title Original Study:** The Relative Efficiency of Skilled Labor across Countries:
4359 Measurement and Interpretation

4360 **doi:** <https://doi.org/10.1257/aer.20191852>, American Economic Review

4361 **Report's Abstract:** Rossi (2022) examines the relative efficiency of skilled workers
4362 across countries. He finds the elasticity of skill efficiency with respect to GDP per
4363 worker is 1.4 and that the relative human capital accounts for only about 9 percent.
4364 We reproduce the paper's main findings and test the sensitivity of the results to (1)
4365 alternative samples and (2) additional controls for determining wages. We find the
4366 results remain robust to these alternative specifications, and the estimated values
4367 of the key elasticities remain nearly unchanged.

4368 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/59.htm>

4369 **Link to Replicators' Package:** <https://osf.io/fge7z/>

4370 **Original Author's Response:** "Thanks for replicating the paper. I don't have
4371 any comments to add to the report."

4372 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/146041/version/V1/view)
4373 [146041/version/V1/view](https://www.openicpsr.org/openicpsr/project/146041/version/V1/view)

4374 **12.17.99 Reproduction Report**

4375 **Title Original Study:** The Side Effects of Immunity: Malaria and African Slavery
4376 in the United States

4377 **doi:** <https://doi.org/10.1257/app.20190372>, American Economic Journal: Applied
4378 Economics

4379 **Report's Abstract:** Esposito (2022) documents the role of malaria in the diffusion
4380 of African slavery in the US. She finds that the introduction of malaria triggered a
4381 demand for malaria-resistant labour, which led to a massive expansion of African
4382 enslaved workers in more malaria-infested areas. Further results document that,
4383 among African slaves, more malaria-resistant individuals commanded significantly
4384 higher prices. We reproduce the paper's main findings, uncovering only one minor
4385 coding error that has no effect on the study's main results. We then test the robust-
4386 ness of the results to (1) varying the set of control variables used in various analyses;
4387 (2) conducting permutation tests; and (3) conducting event studies that account
4388 for time-varying controls. We generally find that the author's results are robust to
4389 all of these alternative specifications, though there are some sets of controls that
4390 cause estimates to become small and statistically insignificant.

4391 **Link to Full Report:** <https://osf.io/728ud/>

4392 **Link to Replicators' Package:** <https://osf.io/728ud/>

4393 **Original Authors' Response:** Original author provided feedback. No final
4394 response on the updated version.

4395 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/120483/version/V1/view)
4396 [120483/version/V1/view](https://www.openicpsr.org/openicpsr/project/120483/version/V1/view)

4397 **12.17.100 Reproduction Report**

4398 **Title Original Study:** The Wheels of Change: Technology Adoption, Millwrights
4399 and the Persistence in Britain'S Industrialisation

4400 **doi:** <https://doi.org/10.1093/ej/ueab102>, Economic Journal

4401 **Report's Abstract:** Mokyr et al. (2022) estimate the effects of early technol-
4402 ogy adoption on industrialization. Authors argue that human capital was the main
4403 determinant of the location of the industry in the first decades of the Industrial Rev-
4404 olution. They document that the location of mills in the eleventh century (reported
4405 in the Domesday Book) has a positive and statistically significant impact on the
4406 number of wrights in the early eighteenth century. We confirm the computational
4407 reproducibility of the paper. The estimates are not sensitive to outliers, which are
4408 common in the data. The results are also robust to changes in the control variables.
4409 The results remain robust if we adjust the estimated p-values for the low number
4410 of clusters, and if we include county fixed effects. We conduct a placebo experi-
4411 ment with a present-day outcome (the Brexit referendum) to check if the results
4412 are picking up on a more general demographic and economic correlation pattern;
4413 the experiment shows no spurious correlations.

4414 **Link to Full Report:** <https://osf.io/gdne3/>

4415 **Link to Replicators' Package:** <https://osf.io/tws8n/>

4416 **Original Authors' Response:** No response.

4417 **Original Authors' Package:** <https://zenodo.org/records/5734954>

4418 **12.17.101 Reproduction Report**

4419 **Title Original Study:** Understanding Ethnolinguistic Differences: The Roles of
 4420 Geography and Trade

4421 **doi:** <https://doi.org/10.1093/ej/ueab065>, Economic Journal

4422 **Report's Abstract:** Dickens (2022) studies the role of trade on long-run inter-
 4423 ethnic linguistic differences. He establishes that neighboring ethnolinguistic groups
 4424 have smaller (lexicostatistical) linguistic distances when there is a larger agricul-
 4425 tural productivity variation between them. Specifically, he establishes that pre-1500
 4426 land productivity variation (CSI SD) and its change due to Columbian Exchange in
 4427 the post-1500 (CSI SD CHANGE) era decrease linguistic distances between groups.
 4428 In what can be considered his main specification, which includes geographical con-
 4429 trols, spatial controls, and language family fixed effects (Table 1 column 5), he
 4430 estimates that a one standard deviation increase in the change in land productiv-
 4431 ity variation (post-1500) decreases linguistic distances by 0.11 standard deviations
 4432 (p-value ≤ 0.01) and a one standard deviation increase in land productivity varia-
 4433 tion (pre-1500) decreases linguistic distances by 0.06 standard deviations (p-value
 4434 $= 0.12$). We conduct a direct replication of the paper by (i) reconstructing the
 4435 main independent variables using the same original sources and following the proce-
 4436 dures explained in the original study, (ii) using an updated version of the linguistic
 4437 map (Ethnologue v17 instead of v16), and (iii) constructing alternative measures
 4438 of inter-ethnic potential gains from trade. Our results basically confirm the sign,
 4439 magnitude, and statistical significance of the point estimates in the original study.

4440 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/62.htm>

4441 **Link to Replicators' Package:** <https://osf.io/k3p7g/>

4442 **Link to Original Authors' Response:** [https://econpapers.repec.org/paper/](https://econpapers.repec.org/paper/zbwi4rdps/63.htm)
 4443 [zbwi4rdps/63.htm](https://econpapers.repec.org/paper/zbwi4rdps/63.htm)

4444 **Original Authors' Package:** <https://doi.org/10.1093/ej/ueab065>

4445 **12.17.102 Reproduction Report**4446 **Title Original Study:** Vulnerability and Clientelism4447 **doi:** <https://doi.org/10.1257/aer.20190565>, American Economic Review

4448 **Report's Abstract:** The paper estimates the effect that changes in household
4449 vulnerability have on citizens' participation in clientelist relationships. The authors
4450 exploit two sources of variation in household vulnerability: rainfall shocks, and a
4451 randomized intervention that provided cisterns in drought-prone areas. We repro-
4452 duce all the findings presented in the four main results tables presented in the
4453 paper. The results of our robustness replication show that the results in the origi-
4454 nal paper are robust to variations in the rainfall period used as a baseline to assess
4455 changes in household vulnerability, and to exclusions that eliminate individuals in
4456 the sample who may have been substituted with others at different survey points.
4457 However, some of the original results that explain the underlying mechanisms are
4458 sensitive to how "clientelist relationships" are defined. When more frequent inter-
4459 actions with politicians are used as the defining characteristic of households in
4460 clientelist relationships, we find that the original results suggesting clientelism as
4461 a significant mechanism are no longer statistically significant at any standard sig-
4462 nificance level. We note, however, that the authors, in a reply to questions we sent
4463 them after the Replication Games, convincingly show that their results are robust
4464 to changing the definition of the clientelist marker.

4465 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/83.htm>4466 **Link to Replicators' Package:** <https://osf.io/q2tw6/>4467 **Link to Original Authors' Response:** <https://econpapers.repec.org/paper/zbwi4rdps/84.htm>4468 **Original Authors' Package:** <https://www.openicpsr.org/openicpsr/project/173341/version/V1/view>
4469
4470

4471 **12.17.103 Reproduction Report**

4472 **Title Original Study:** Wage Cyclicalities and Labor Market Sorting
4473 **doi:** <https://doi.org/10.1257/aeri.20210161>, American Economic Review: Insights
4474 **Report's Abstract:** Figueiredo (2022) examines wage cyclicalities across the skill
4475 mismatch distribution finding large differences. Some key results include finding
4476 that wages are acyclical in good labor market matches but procyclical in poor
4477 matches. Using the public replication material provided by the authors, we were
4478 able to exactly duplicate the results of the study. Further, using several further
4479 robustness checks, such as subtracting (potentially correlated) covariates in the
4480 regressions, using different standard errors (rather than clustered ones), or different
4481 time periods of the data left the key results largely unchanged with some minor
4482 caveats.
4483 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/78.htm>
4484 **Link to Replicators' Package:** <https://osf.io/a8hcg/>
4485 **Original Authors' Response:** "I have read the report and I do not wish to write
4486 a reply.
4487 Congratulations on this initiative – it is great!"
4488 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/150581/version/V1/view)
4489 [150581/version/V1/view](https://www.openicpsr.org/openicpsr/project/150581/version/V1/view)

4490 **12.17.104 Reproduction Report**

4491 **Title Original Study:** War, Socialism, and the Rise of Fascism: an Empirical
4492 Exploration

4493 **doi:** <https://doi.org/10.1093/qje/qjac001>, Quarterly Journal of Economics

4494 **Report's Abstract:** In this report, we present the results from a replication of
4495 Acemoglu et al. (2022). The authors suggest that the emergence of the 'Red Scare'
4496 in the aftermath of World War I led to a rise of fascism in Italy in the early 1920s.
4497 Their approach uses the war casualties as an instrument for the rise in socialist
4498 voting. We performed a series of replication strategies, including pre-trend controls,
4499 applying an alternative instrument and modifying the first-stage specification, as
4500 well as investigating the long-run political alignment. Based on our findings, the
4501 original authors' results are replicable under a variety of alternative specifications.

4502 **Link to Full Report:** <https://osf.io/a672c/>

4503 **Link to Replicators' Package:** <https://osf.io/a672c/>

4504 **Link to Original Authors' Response:** No response.

4505 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml?](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CLJTSC)
4506 [persistentId=doi:10.7910/DVN/CLJTSC](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CLJTSC)

4507 **12.17.105 Reproduction Report**

4508 **Title Original Study:** What Makes Anticorruption Punishment Popular?
 4509 Individual-Level Evidence from China

4510 **doi:** <https://doi.org/10.1086/715252>, Journal of Politics

4511 **Report's Abstract:** It also indirect effects through affecting evaluations of compe-
 4512 tence and morality. Conducting a conjoint study in China where respondents were
 4513 asked to choose between two potential local officials, Tsai et al. found that 26% of
 4514 the total effect of these officials punishing corrupt subordinates was estimated to
 4515 come through indirect effects that go through evaluations of morality and compe-
 4516 tence. Using their code, I reproduced their original findings, and did not find any
 4517 notable coding errors while doing so. Then, taking advantage of the fact that Tsai et
 4518 al. included several additional covariates beyond punishment in their experiment, I
 4519 engaged in an extension of the original model, using the same method, to examine
 4520 whether economic performance characteristics have indirect effects on evaluation
 4521 through competence and morality as well. I found results that suggest that eco-
 4522 nomic performance does have an indirect effect on preferences through competence
 4523 and morality. I then tested the robustness of Tsai et al.'s original heterogeneous
 4524 sensitivity tests by varying cut points on two demographic variables and found
 4525 that their findings of a lack of heterogeneous sensitivity remain robust to different
 4526 cut-points. In all, my efforts suggest that Tsai et al.'s methods are valid and their
 4527 findings robust.

4528 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/7.htm>

4529 **Link to Replicators' Package:** <https://osf.io/czs6j/>

4530 **Original Authors' Response:** "We appreciate your efforts, both in replicating
 4531 our paper and in doing so systematically for other studies in leading political sci-
 4532 ence and economic journals. Your contribution is valuable to the entire academic
 4533 community and to us especially.

4534 We also appreciate your sharing Reproduction Reports with the original authors
 4535 prior to dissemination and are glad to see from the Reproduction Report that our
 4536 results and methods appear to be both valid and robust. Although a longer follow-
 4537 up may not be necessary, we do wish to convey our gratitude to the replicator(s)
 4538 and to the editorial team."

4539 **Original Authors' Package:** [https://dataverse.harvard.edu/dataset.xhtml;](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=date)
 4540 [jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=date)
 4541 [2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=date)
 4542 [Public&fileSortField=date](https://dataverse.harvard.edu/dataset.xhtml;jsessionid=34454d461ad29192edc557995095?persistentId=doi%3A10.7910%2FDVN%2FXTRWKG&version=&q=&fileTypeGroupFacet=&fileAccess=Public&fileSortField=date)

4543 **12.17.106 Reproduction Report**

4544 **Title Original Study:** When a Doctor Falls from the Sky: The Impact of Easing
4545 Doctor Supply Constraints on Mortality

4546 **doi:** <https://doi.org/10.1257/aer.20210701>, American Economic Review

4547 **Report's Abstract:** Okeke (2023) evaluates a policy experiment conducted in
4548 Nigeria, whereby communities were randomly allocated to receive a new doctor
4549 at the local public health center. The performance of these centers was compared
4550 to other sites which were allocated either a new midlevel health-care provider, or
4551 no additional staff. The study finds that communities assigned a new doctor were
4552 associated with a decrease in seven-day infant mortality, such a decrease was not
4553 observed in communities assigned a midlevel health-care provider. This suggests
4554 that it is the 'quality' of the additional doctor driving the effects rather than due
4555 to a quantity increase of an additional health worker. The size of the mortality
4556 reduction increased with increased exposure to the intervention. We first conduct
4557 a computational reproduction, rerunning the original code and data, finding that
4558 the results reported in the original study are reproducible. Second, we test the
4559 robustness of the results in several ways, by 1) adapting the existing controls to
4560 make the results robust to contamination bias, 2) altering and adding to the control
4561 variables included, 3) changing the specification or regression technique used, and
4562 4) testing coding grouping and changing how service use was coded. These changes
4563 cause little change to the point estimates, although we find that the original paper's
4564 standard errors were overly conservative, and thus the statistical significance of
4565 some results was understated.

4566 **Link to Full Report:** <https://econpapers.repec.org/paper/zbwi4rdps/53.htm>

4567 **Link to Replicators' Package:** [https://github.com/e-mcmanus/Okeke23_](https://github.com/e-mcmanus/Okeke23_Replication)
4568 [Replication](https://github.com/e-mcmanus/Okeke23_Replication)

4569 **Original Authors' Response:** "Thank you for sharing the Reproduction Report
4570 (and please pass on my thanks to the replicators). There does not appear to be
4571 much for me to respond to. It is gratifying to see that the results have held up well
4572 to additional scrutiny."

4573 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/181581/version/V1/view)
4574 [181581/version/V1/view](https://www.openicpsr.org/openicpsr/project/181581/version/V1/view)

4575 **12.17.107 Reproduction Report**

4576 **Title Original Study:** Who Chooses Commitment? Evidence and Welfare
4577 Implications

4578 **doi:** <https://doi.org/10.1093/restud/rdab056>, Review of Economic Studies

4579 **Report's Abstract:** We conduct a computational reproduction and a robustness
4580 replication of Carrera et al. (2022) by using the same dataset and similar procedures
4581 as specified in their paper (i.e., method and analysis). Instead of using STATA,
4582 we use R and code the results from scratch. We also replicate the MATLAB code
4583 used for simulations and test whether it produces reasonable results for different
4584 parameter values. We confirm all of the main results and do not find high sensitivity
4585 of the model to changes in parameters.

4586 **Link to Full Report:** <https://osf.io/752q9/>

4587 **Link to Replicators' Package:** <https://osf.io/752q9/>

4588 **Link to Original Authors' Response:** The authors provided feedback which
4589 was taken into account.

4590 **Original Authors' Package:** <https://zenodo.org/records/5173081>

4591 **12.17.108 Reproduction Report**

4592 **Title Original Study:** Who Sells During a Crash? Evidence from Tax Return
4593 Data on Daily Sales of Stock

4594 **doi:** <https://doi.org/10.1093/ej/ueab059>, Economic Journal

4595 **Report's Abstract:** Hoopes et al., (2021) analyze United States tax return
4596 data encompassing all individual taxable stock sales between 2008 and 2009, to
4597 investigate the individuals who increased their stock sales in response to market
4598 turbulence. Our findings reveal that such increases were notably prevalent among
4599 investors in the highest tiers of the income distribution, including the top 1% and
4600 0.1%, as well as retirees and those at the uppermost levels of the dividend income
4601 distribution. We reproduce the paper's main findings and results are very similar.

4602 **Link to Full Report:** <https://osf.io/b6s9k/>

4603 **Link to Replicators' Package:** [https://www.dropbox.com/scl/fo/](https://www.dropbox.com/scl/fo/c3ysdlenysq391mugzprm/h?rlkey=riooyohci7i5vwx475r13jaqq&dl=0)
4604 [c3ysdlenysq391mugzprm/h?rlkey=riooyohci7i5vwx475r13jaqq&dl=0](https://www.dropbox.com/scl/fo/c3ysdlenysq391mugzprm/h?rlkey=riooyohci7i5vwx475r13jaqq&dl=0)

4605 **Original Authors' Response:** The authors provided feedback which was taken
4606 into account.

4607 **Original Authors' Package:** <https://doi.org/10.1093/ej/ueab059>

4608 **12.17.109 Reproduction Report**

4609 **Title Original Study:** Why Don't Firms Hire Young Workers During Recessions?
4610 **doi:** <https://doi.org/10.1093/ej/ueab096>, Economic Journal

4611 **Report's Abstract:** We gauge the replicability of the results of Forsythe (2022)
4612 studying the cyclical transitions of individuals' labor market transitions conditional on their
4613 experience. Using Current Population Survey (CPS) data and state-level variation
4614 in cyclical unemployment, this paper shows that the hiring probability of youths
4615 is more sensitive to business-cycle conditions than that of experienced individuals.
4616 We replicate the main results in this paper by reconstructing the dataset using
4617 data from the IPUMS-CPS database (Flood et al. (2020)) and recoding the paper's
4618 main regressions. We also conduct a robustness replicability analysis and show
4619 that the paper's main results are robust in terms of statistical significance to (i)
4620 extending the sample period from 1994-2014 to 1994-2019 and (ii) using MSA-level
4621 unemployment variation instead of state-level variation. However, these extensions
4622 reduce the magnitude of the main effects of interest. The paper's key conclusions
4623 are unaffected.

4624 **Link to Full Report:** <https://osf.io/3pqbt/>

4625 **Link to Replicators' Package:** [https://github.com/jcrechet/replication_](https://github.com/jcrechet/replication_forsythe.2022_EJ)
4626 [forsythe.2022_EJ](https://github.com/jcrechet/replication_forsythe.2022_EJ)

4627 **Link to Original Authors' Response:** The author responded but did not
4628 provide a response.

4629 **Original Authors' Package:** <https://zenodo.org/records/5710784>

12.17.110 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.

4674 **Original Authors' Package:** [https://www.openicpsr.org/openicpsr/project/](https://www.openicpsr.org/openicpsr/project/128143/version/V1/view)
4675 [128143/version/V1/view](https://www.openicpsr.org/openicpsr/project/128143/version/V1/view)