Mass Reproducibility and Replicability: A New Hope

Brodeur, Mikola, Cook et al.*

September 2024

Abstract

This study pushes our understanding of research reliability by reproducing and replicating claims from 110 papers in leading economic and political science journals. The analysis involves computational reproducibility checks and robustness assessments. It reveals several patterns. First, we uncover a high rate of fully computationally reproducible results (over 85%). Second, excluding minor issues like missing packages or broken pathways, we uncover coding errors for about 25% of studies, with some studies containing multiple errors. Third, we test the robustness of the results to 5,511 re-analyses. We find a robustness reproducibility of about 70%. Robustness reproducibility rates are relatively higher for re-analyses that introduce new data and lower for re-analyses that change the sample or the definition of the dependent variable. Fourth, 52% of re-analysis effect size estimates are smaller than the original published estimates and the average statistical significance of a re-analysis is 77% of the original. Lastly, we rely on six teams of researchers working independently to answer eight additional research questions on the determinants of robustness reproducibility. Most teams find a negative relationship between replicators' experience and reproducibility, while finding no relationship between reproducibility and the provision of intermediate or even raw data combined with the necessary cleaning codes.

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

JEL Codes: B41, C10, C81

*Corresponding author: Abel Brodeur (University of Ottawa), Email: abrodeur@uottawa.ca. See Section A.1 for the full list of authors and each author's contribution. Disclaimers: We acknowledge support from Open Philanthropy and the Social Sciences and Humanities Research Council. Any views expressed therein are the authors' personal opinions and not those of PHAC. The work by Jeremy D. Gretton was not undertaken under the auspices of PHAC as part of his employment responsibilities. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. The findings, interpretations, and conclusions expressed in this work are entirely those of the authors and do not necessarily reflect the views of the World Bank or its Board of Directors. The Center for Crisis Early Warning (Kompetenzzentrum Krisenfrüherkennung) is funded by the German Federal Ministry of Defense and the German Federal Foreign Office. The views and opinions expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of any agency of the German government. All remaining errors are the authors' responsibility.

Author contribution, Section A.1.

Author list: Abel Brodeur, Derek Mikola, Nikolai Cook, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, Lenka Fiala, Jacopo Gabani, Romain Gauriot, Joanne Haddad, Goncalo Lima, Jörg Ankel-Peters, Anna Dreber, Douglas Campbell, Lamis Kattan, Diego Marino Fages, Fabian Mierisch, Pu Sun, Taylor Wright, Marie Connolly, Fernando Hoces de la Guardia, Magnus Johannesson, Edward Miguel, Lars Vilhuber, Alejandro Abarca, Mahesh Acharya, Sossou Simplice Adjisse, Ahwaz Akhtar, Eduardo Alberto Ramirez Lizardi, Sabina Albrecht, Synøve Nygaard Andersen, Zubaria Andlib, Falak Arrora, Thomas Ash, Etienne Bacher, Sebastian Bachler, Félix Bacon, Manuel Bagues, Timea Balogh, Alisher Batmanov, Mara Barschkett, B. Kaan Basdil, Jaromír Baxa, Sascha Becker, Monica Beeder Louis-Philippe Beland, Abdel-Hamid Bello, Daniel Benenson Markovits, Grant Benjamin, Thomas Bergeron, Moussa Blimpo, Marco Binetti, Carl Bonander, Joseph Bonneau, Endre Borbáth, Nicolai Topstad Borgen, Solveig Topstad Borgen, Jonathan Borowsky, Elisa Brini, Myriam Brown, Martin Brun, Stephan Bruns, Nino Buliskeria, Andrea Calef, Alistair Cameron, Pamela Campa, Santiago Campos-Rodríguez, Giulio Giacomo Cantone, Fenella Carpena, Perry Carter, Paul Castañeda Dower, Ondrej Castek, Jill Caviglia-Harris, Gabriella Chauca Strand, Shi Chen, Asya Chzhen, Jong Chung, Jason Collins, Alexander Coppock, Hugo Cordeau, Ben Couillard, Jonathan Crechet, Lorenzo Crippa, Jeanne Cui, Christian Czymara, Haley Daarstad, Danh Chi Dao, Dong Dao, Marco David Schmandt, Astrid de Linde, Lucas De Melo, Lachlan Deer, Micole De Vera, Velichka Dimitrova, Jan Fabian Dollbaum, Jan Matti Dollbaum, Michael Donnelly, Luu Duc Toan Huynh, Tsvetomira Dumbalska, Jamie Duncan, Kiet Tuan Duong, Thibaut Duprey, Christoph Dworschak, Sigmund Ellingsrud, Ali Elminejad, Yasmine Eissa, Andrea Erhart, Giulian Etingin-Frati, Elaheh Fatemi-Pour, Alexa Federice, Jan Feld, Guidon Fenig, Mojtaba Firouzjaeiangalougah, Erlend Fleisje, Alexandre Fortier-Chouinard, Julia Francesca Engel, Tilman Fries, Reid Fortier, Nadjim Fréchet, Thomas Galipeau, Sebastián Gallegos, Areez Gangji, Xiaoying Gao, Cloé Garnache, Attila Gáspár, Evelina Gavrilova, Arijit Ghosh, Garreth Gibney, Grant Gibson, Geir Godager, Leonard Goff, Da Gong, Javier González, Jeremy D. Gretton, Cristina Griffa, Idaliya Grigoryeva, Maja Grøtting, Eric Guntermann, Jiaqi Guo, Alexi Gugushvili, Hooman Habibnia, Sonja Häffner, Jonathan D. Hall, Olle Hammar, Amund Hanson Kordt, Barry Hashimoto, Jonathan S. Hartley, Carina I. Hausladen, Tomáš Havránek, Harry He, Matthew Hepplewhite, Mario Herrera-Rodriguez, Felix Heuer, Anthony Heyes, Anson T. Y. Ho, Jonathan Holmes, Armando Holzknecht, Yu-Hsiang Dexter Hsu, Shiang-Hung Hu, Yu-Shiuan Huang, Mathias Huebener, Christoph Huber, Kim P. Huynh, Zuzana Irsova, Ozan Isler, Niklas Jakobsson, Michael James Frith, Raphaël Jananji, Tharaka A. Jayalath, Michael Jetter, Jenny John, Rachel Joy Forshaw, Felipe Juan, Valon Kadriu, Sunny Karim, Edmund Kelly, Duy Khanh Hoang Dang, Tazia Khushboo, Jin Kim, Gustav Kjellsson, Anders Kjelsrud, Andreas Kotsadam, Jori Korpershoek, Lewis Krashinsky, Suranjana Kundu, Alexander Kustov, Nurlan Lalayev, Audrée Langlois, Jill Laufer, Blake Lee-Whiting, Andreas Leibing, Gabriel Lenz, Joel Levin, Peng Li, Tongzhe Li, Yuchen Lin, Goncalo Lima, Ariel Listo, Dan Liu, Xuewen Lu, Elvina Lukmanova, Alex Luscombe, Lester R. Lusher, Ke Lyu, Hai Ma, Nicolas Mäder, Clifton Makate, Alice Malmberg, Adit Maitra, Marco Mandas, Jan Marcus, Shushanik Margaryan, Lili Márk, Andres Martignano, Abigail Marsh, Isabella Masetto, Anthony McCanny, Emma McManus, Ryan McWay, Lennard Metson, Jonas Minet Kinge, Sumit Mishra, Myra Mohnen, Jakob Möller, Rosalie Montambeault, Sébastien Montpetit, Louis-Philippe Morin, Todd Morris, Scott Moser, Fabio Motoki, Lucija Muehlenbachs, Andreea Musulan, Marco Musumeci, Munirul Nabin, Karim Nchare, Florian Neubauer, Quan M. P. Nguyen, Tuan Nguyen, Viet Nguyen-Tien, Ali Niazi, Giorgi Nikolaishvili, Ardyn Nordstrom, Patrick Nüß, Angela Odermatt, Matt Olson, Henning Øien, Tim Ölkers, Miquel Oliver i Vert, Emre Oral, Christian Oswald, Ali Ousman, Ömer Özak, Shubham Pandey, Alexandre Pavlov, Martino Pelli, Romeo Penheiro, RyuGyung Park, Eva Pérez Martel, Tereza Petrovičová, Linh Phan, Alexa Prettyman, Jakub Procházka, Aqila Putri, Julian Quandt,

Kangyu Qiu, Loan Quynh Thi Nguyen, Andaleeb Rahman, Carson H. Rea, Adam Reiremo, Laëtitia Renée, Joseph Richardson, Nicholas Rivers, Bruno Rodrigues, William Roelofs, Tobias Roemer, Ole Rogeberg, Julian Rose, Andrew Roskos-Ewoldsen, Paul Rosmer, Barbara Sabada, Soodeh Saberian, Nicolas Salamanca, Georg Sator, Daniel Scates, Elmar Schlüter, Cameron Sells, Sharmi Sen, Ritika Sethi, Anna Shcherbiak, Moyosore Sogaolu, Matt Soosalu, Erik Ø. Sørensen, Manali Sovani, Noah Spencer, Stefan Staubli, Renske Stans, Anya Stewart, Felix Stips, Kieran Stockley, Stephenson Strobel, Ethan Struby, John Tang, Idil Tanrisever, Thomas Tao Yang, Ipek Tastan, Dejan Tatić, Benjamin Tatlow, Féraud Tchuisseu Seuyong, Rémi Thériault, Vincent Thivierge, Wenjie Tian, Filip-Mihai Toma, Maddalena Totarelli, Van Tran, Hung Truong, Nikita Tsoy, Kerem Tuzcuoglu, Diego Ubfal, Laura Villalobos, Julian Walterskirchen, Joseph Tao-yi Wang, Vasudha Wattal, Matthew D. Webb, Bryan Weber, Reinhard Weisser, Wei-Chien Weng, Christian Westheide, Kimberly White, Jacob Winter, Timo Wochner, Matt Woerman, Jared Wong, Ritchie Woodard, Marcin Wroński, Myra Yazbeck, Chung Yang, Luther Yap, Kareman Yassin, Hao Ye, Jin Young Yoon, Chris Yurris, Tahreen Zahra, Mirela Zaneva, Aline Zayat, Jonathan Zhang, Ziwei Zhao, Yaolang Zhong

And from nature we should learn
That all can start again
As the stars must fade away
To give a bright new day.
"Oh My Love" by Riz Ortolani - (feat. Katyna Ranieri)

1 Introduction

Reproducibility and replication efforts contribute in essential ways to the production of scientific knowledge by testing accumulated evidence.[1] Reproductions and replications assess which findings are robust, promoting self-correcting science and affecting policy-making (Vazire (2017)). Importantly, reproductions and replications emphasize that evidence is cumulative and should be assessed holistically. Active research fields appear when previous research fails to be replicated or reproduced. Yet, reproducible and replicable research increases the confidence in scientific communities and our investments and innovations relying on that knowledge. Replications and reproductions in research are also foundational to teaching, ensuring that the knowledge being passed on is accurate and reliable, and providing practical experiences for students. For all these reasons, reproductions and replications are considered to be an essential diagnostic tool (King (1995); Moonesinghe et al. (2007); Peterson and Panofsky (2021)) and there is broad agreement that they should be given more visibility (Brandon and List (2015); Freese and Peterson (2017); Maniadis and Tufano (2017); Munafò et al. (2017); Nosek et al. (2022)).

Yet a large literature has documented relatively low data availability and computational reproducibility rates. For around half of the papers published in leading economics journals, the data are not publicly available (Askarov et al. (2023); Brodeur et al. (2024b); Christensen and Miguel (2018); Pérignon et al. (2019)) because of their nature: administrative, proprietary, or copyrighted, data. For many other studies, the required computer code is unavailable or incomplete (Dafoe (2014); Gertler et al. (2018)). Even more puzzling is that some published results cannot be fully computationally reproduced even when all required resources (data, software, hardware, *etc.*) are available (Chang and Li (2022); Pérignon et al. (2023)). Reasons put forward to explain the latter case include: lack of complete documentation, versioning issues for packages, and results which are numerically fragile.[2]

There is also growing evidence on the lack of replicability—*i.e.*, when subsequent attempts to test a hypothesis using new data yield inconsistent results—in the social sciences. A few large-scale systematic replication projects have taken place recently, including one in psychology (Open Science Collaboration (2015)), one in experimental economics (Camerer et al. (2016)) and a social science replication project (Camerer et al. (2018)). Pooling the results of these large replication projects yielded a replication rate of about 50%.

[1] See Section 2.2 for definitions of reproducibility and replicability.
[2] Using varying approaches and definitions of computational reproducibility, Chang and Li (2022), Gertler et al. (2018) and Wood et al. (2018) find, respectively, that between 14% and 43% of published studies were computationally reproducible.

This paper examines reproducibility and replicability rates for a large number of papers recently published in leading economic and political science outlets. Studying more recent studies may shed a different light on the issues discussed above. Journals are increasingly complying with specific guidelines (*i.e.*, TOP Guidelines Nosek et al. (2015)) or conducting internal reproducibility checks (Vilhuber (2020)). Support for posting data and code has increased FAIRness (Ferguson et al. (2023)): easier findability of research, easier accessibility of computational artifacts, greater clarity on how to understand the underlying data and methods, and an increase in the critical re-use of data, code, and methods.

Our project involves mass reproducing and replicating the main claims from studies published from 2022 onwards in nine leading economics outlets and three leading political science outlets. We present the results from our first 110 reproductions/replications in this piece. For each study, we work in small teams and first conduct computational reproducibility checks—the extent to which results in original studies can be reproduced using both the data and code from those studies—and document coding errors and discrepancies between the codes and the article. We then conduct robustness checks, recode the original analysis, or both, using the data provided in the original study's replication folder. Some teams also replicated the original study's findings using new data.

We document a high rate of computational reproducibility using the *Social Science Reproduction Platform*'s (SSRP) 10-point scale on computational reproducibility. This scale ranges from 1 to 10, with 1 indicating an inability to reproduce results due to missing data or code, and 10 indicating the capability to faithfully reproduce results from raw data to final numerical results. Teams assigned reproducibility scores to the papers they reproduced, focusing on the claims they investigated. The results showed a majority (over 85%) of examined results were fully reproducible using the data and code provided by the authors. The remaining 15% included studies with partial availability of analytic code and data or cases where some codes failed to run or produced inconsistent results. These findings contrast with previous studies, which uncovered low rates of computational reproducibility. This is likely influenced by our approach of targeting newer studies and nine (out of 12) outlets internally conducting computational reproducibility checks. See Section 4 for more details.

We then investigate the prevalence of coding errors and discrepancies between the code and article. Except for minor discrepancies (*i.e.*, missing packages or broken pathways), we identified coding errors in approximately one-fourth of the studies, with some studies containing multiple errors. 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 did 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.

Our main analysis documents robustness reproducibility rates based on 5,511 re-analyses. Robustness reproducibility explores the extent to which results in original studies remain robust to alternative analytical decisions, utilizing the same datasets as in the original studies. Robustness re-

producibility is conducted by teams of economists and political scientists on subjects that they themselves are familiar with (i.e., within their primary field of interest), and make re-analysis choices that are theoretically informed. Our re-analyses involve specification checks such as changing the weighting scheme, changing the choice of control variables or changing estimation methods. The specification checks are theoretically informed, and vary across papers. We rely on several definitions of robustness reproducibility throughout. Our main definition is whether the effect is in the same direction and remains statistically significant at the 5% level. Using this definition, we find a robustness reproducibility rate of about 70%. Further, we find that half of original point estimates significant at the 10% level (but insignificant at the 5% level) become statistically insignificant at the 10% threshold with our robustness checks. For original estimates significant at the 5% level (but insignificant at the 1% level), more than a quarter of re-analyses become insignificant at the 10% threshold. More formally, while we document the presence of publication bias and p-hacking using methods proposed in Gerber and Malhotra (2008a) and Andrews and Kasy (2019) for the original studies, we find reduced selection issues for the re-analyses.

We then explore heterogeneity in robustness reproducibility and replicability. We group re-analyses into eight groups. We find that robustness reproducibility rates are markedly lower when replicators change the (coding of the) dependent variable (45%) and the sample (64%). In contrast, replicability rate is the highest for re-analyses that introduce new data (87%). The remaining groups (i.e., changing control variables, estimation method, inference method, main independent variable or weighting scheme) result in robustness reproducibility rates of about 75%.

Last, we use a "many-analysts" approach where six research teams use the re-analysis data to tackle eight additional research questions (in the spirit of Silberzahn et al. (2018) and Huntington-Klein et al. (2021)). We tackle questions ranging from "Does reproducibility/replicability rate depend on replicators' academic experience or experience coding?" to "Does reproducibility/replicability rate depend on the interaction of the authors' prestige and replicators' prestige?" and "Does reproducibility/replicability rate depend on the original authors providing raw data?" Each team receives the same instructions and answers each research question independently. Teams may choose to produce multiple estimates to each question, though we weight the estimates in a way that ensures all teams' results obtain equal weight. We allow full flexibility to all teams and pre-registered this exercise.

We observe a general agreement among analyst teams on the answer to some of these research questions. We provide evidence for a negative relationship between robustness reproducibility and replicators' experience (implying more experience coding leads to lower reproduction probability). A similar finding is found for replicators' academic experience, but not for original authors' experience. For their interaction, the teams find weak evidence that the reproducibility rate increases when authors have high experience relative to the replicators. Prestige (defined independently by each analyst team) has a similar pattern. The last three research questions focused on the relationship between robustness reproducibility and the original authors' degree of provision of data and code. According to the teams, the provision of raw or intermediate data, relative to the provision of only the final processed data, does not seem to affect the robustness of research.

In the course of our 110 reproductions and replications, we always engage with the original authors allowing and encouraging them to respond to our replication reports. Sharing the replication reports creates an opportunity for constructive exchange of ideas and expert feedback, which can lead to mutual learning and improvement in research practices. The vast majority of original authors engaged with our reports and 78% provided feedback to the replicators and/or wrote a formal response. We document the types of remaining disagreements in Section 2.7.

Our project differs from previous efforts in several ways. First, our focus is not solely on laboratory experiments (e.g., Camerer et al. (2016)), but rather on all data types used in economics and political science research. Second, we computationally reproduce *and* conduct robustness reproducibility or replicate research findings, in contrast to a growing literature conducting large-scale computational reproducibility. Conducting sensitivity analysis on a large scale allows us to assess the stability and reliability of empirical findings. Third, we conduct a large range of recoding and specification checks, in contrast to studies focusing on one method/robustness check (e.g., Young (2022)). Fourth, our focus is on a sample of articles recently published which have potentially relied on new open science tools such as pre-analysis plans and registered reports. This is a key difference with previous work that investigated reproducibility and replicability before the open science movement.

One of our contributions is the scale of this ongoing project.[3] We believe mass reproduction and replication of research findings offers the potential to change research norms and researchers' behavior at scale. It may encourage researchers to adopt more rigorous methodologies and perhaps even act as a deterrent to questionable research practices, while also emphasizing the collaborative nature of science. In turn, it may lead to a shift in publication norms, with a strong emphasis on the reliability of results.

The rest of this paper is organized as follows. Section 2 provides a conceptual background and describes our methodology. Section 3 describes our data and provides descriptive statistics. Section 4 documents computational reproducibility rates and the prevalence of coding errors. Section 5 documents robustness reproducibility and replicability rates. Section 6 discusses our main findings and barriers to reproducibility. Section 7 discusses benefits of our approach to replicators. In Section 8, we rely on a many-analysts approach to answer additional research questions. Section 9 concludes.

2 Conceptual Background and Methodology

2.1 Existing Literature and Incentives to Reproduce and Replicate

Concerning experimental data, several extensive replication initiatives have occurred in various fields recently. Notable examples include a project in psychology (Open Science Collaboration (2015)), one in experimental economics (Camerer et al. (2016)), and a social science replication

[3]For economics, we are currently conducting robustness reproducibility or replicability for about 250 studies per year, or about 25% of all empirical studies published in our targeted journals. We hope to soon expand the scale of our project to include more journals, fields and types of data (e.g., hard-to-access administrative data).

project (Camerer et al. (2018)). See Nosek et al. (2022) for a review. In this context, replication involves selecting the primary significant result from the original study and conducting the study anew on a fresh sample using comparable methods and tests (referred to as "direct replication"; see the following section for definitions). Combining the outcomes of these large-scale replication projects revealed an overall replication rate of approximately 50%.

The low replicability rates for experiments can be due to many factors, including low statistical power (Arel-Bundock et al. (2022); Maniadis et al. (2014)). These factors are also present for non-experimental work. Indeed, many observational studies have been performed on small sample sizes, possibly implying low statistical power. Ioannidis et al. (2017) surveyed 159 empirical economics literatures and found that the median statistical power is 18% or less. Moreover, there are typically many ways of testing a hypothesis, giving researchers many "degrees of freedom" in their analysis. Specification searching (or "p-hacking") and publication bias have also been found to be a problem (e.g., Doucouliagos and Stanley (2011); Havránek and Sokolova (2020)). Numerous studies indicate that the prevalence of p-hacking is lower in papers employing Randomized Controlled Trials (RCTs) compared to those utilizing alternative methods of causal inference (Brodeur et al. (2016), Brodeur et al. (2020), and Vivalt (2019)). These results potentially imply that prioritizing mass reproducibility and replicability might be of greater significance for non-experimental work.

In addition to the technical and logistical hurdles that prevent researchers from reproducing past evidence, the current publication incentives remain unfavorable to reproductions and replications (Clemens (2017); Coffman et al. (2017); Mueller-Langer et al. (2019)). Publication outlets tend to favor novel conceptual insights over new tests of a published idea, regardless of what these tests find. Another reason why journals potentially do not publish replications is that comments are hard to review and do not get a lot of citations (Ankel-Peters et al. (2023)).

2.2 Definitions of Reproducibility and Replicability

Several definitions of reproducibility and replicability have been posited and examined (Hamermesh (2007) and Clemens (2017)). Indeed, the authors of this study do not always rely on the same definitions in their reproduction/replication as there is no consensus in the literature.[4] Dreber and Johannesson (2023) have recently introduced definitions and indicators which we rely on throughout this paper.

Reproducibility is the examination of whether the results and conclusions of original studies can be duplicated using the original studies' data, while **replicability** is defined as using data other than what was used in the original studies.

Reproducibility is further delineated into three categories. **Computational reproducibility** gauges the extent to which results in original studies can be reproduced using both the data and code from those studies. **Recreate reproducibility** assesses the extent to which results can be reproduced using the information in the original studies without access to the processed data set

[4]For instance, "replication" as used by many authors of this study (and researchers in economics and political science) encompasses both "reproduction" and "replication" in the conceptual framework of Dreber and Johannesson (2023).

and/or the analysis code. **Robustness reproducibility** explores the extent to which results in original studies remain robust to alternative plausible analytical decisions, utilizing the same data as in the original studies.

Replicability is also classified into two types. **Direct replicability** evaluates the extent to which results in original studies can be repeated on new data using the original studies' research design and analysis. This classification is further subdivided based on whether data from the same population, a similar population, or a different population is employed. **Conceptual replicability** employs new data to assess the extent to which results in original studies can be repeated; however, this type of replication involves an alternative research design and/or alternative analysis to test the same hypothesis as in the original study. Conceptual replicability is also further subdivided into the same three categories based on populations which are the same, similar or different.

2.3 Methodology

For this paper, our focus is on the following nine economic journals: (1) *American Economic Review*, (2) *American Economic Review: Insights*, (3) *American Economic Journal: Applied Economics*, (4) *American Economic Journal: Economic Policy* and (5) *American Economic Journal: Macroeconomics*, (6) *The Economic Journal*, (7) *Journal of Political Economy*, (8) *Quarterly Journal of Economics* and (9) *Review of Economic Studies*. For political science, our focus is on three journals: (1) *American Journal of Political Science*, (2) *American Political Science Review* and (3) *Journal of Politics*. These journals were selected for multiple reasons. First, all of these journals are considered leading outlets in their respective disciplines. Second, they all have a data and code availability policy. Third, most of these journals conduct computational reproducibility checks for most accepted articles. The computational reproducibility is conducted internally by a data editor and his/her team.[5] The journals which do not conduct computational reproducibility checks are the *American Political Science Review*, the *Journal of Political Economy* and the *Quarterly Journal of Economics*. Data editors also enforce their journal data and code availability policy and enhance the completeness of the replication package. About 77% of articles in our sample were published by a journal with a data editor or a group conducting computational reproducibility such as the Odum Institute.[6]

Our sample of journals should thus be seen as highly selective. We focus on journals which enforce their data and code availability policy and are high impact. Moreover, we focus solely on studies published since 2022. Our aim is to reproduce and replicate studies as soon as they are published, as to achieve at least two goals: (i) provide a rapid assessment of the credibility of new claims and (ii) make original authors more engaged. The high response rate from original authors is perhaps indicative that focusing on more recent work make them more engaged. We come back to this point later in Section 2.7, and the representativity of our sample in Section 2.8.

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

[6]Some of the articles published in those outlets are not computationally reproduced by the data editor for a variety of reasons, including not having access to the restricted data or software. These reasons and the share of articles computationally reproduced vary across journals.

2.4 Reproduction and Replication Process

Assessments of reproducibility and replicability may unfortunately gravitate towards binary judgments that declare an entire paper as "irreplicable". For our empirical analysis, we directly compare original point estimates to the revised point estimates. This one-on-one comparison allows us to speak to the reproducibility and replicability of a specific hypothesis test, in addition to the reproducibility and replicability of our entire sample. Our strategy differs from large-scale replications such as Camerer et al. (2016) along (at least) one crucial dimension; we are looking at several claims within a study and conduct robustness reproducibility or replicability for each claim.

Replicators are economists and political scientists with an interest in the article and have some expertise of the research question. They first identify the main claims, check for computational reproducibility, and are then free to conduct any robustness or recoding exercises. This flexibility is very important as each study is different and allows for different re-analyses. For instance, some studies provide the raw data, while others only provide the final data set. Furthermore, the type of sensitivity analysis and recoding that are relevant for an applied microeconomics paper using a difference-in-differences method might be different from a political science study using a regression discontinuity. We do our best to match replicators' skills and fields of expertise with papers from similar fields. Replicators reproduce or replicate a study in their primary field of interest. We provide summary statistics by types of re-analyses and field of study in the next subsections.

This flexibility in choosing which re-analyses to conduct has advantages and disadvantages. One key advantage is that we can document reproduciblity and replicability rates for different types of re-analysis.[7] Another advantage is that replicators act as "super" reviewers. They do not make a recommendation to the editor, nor do they comment on the contribution to the literature. Instead, they focus on the reproducibility of the claims and have access to the replication package, allowing them to directly test the sensitivity of the main results. This is a crucial advantage over the traditional review process as replicators 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, replicators spend different amounts of time and effort on their respective replication. Some replicators are more experienced at coding, while others are more familiar with methods. This means that not all replication reports are of the same quality. We come back to the discussion of quality in Section 6.

2.5 Generating Reproductions and Replications

We have two streams to generate reproductions and replications. All replicators are coauthors on this paper.

[7] Once our sample size becomes larger, we will also be able to document replicability rates by field and method. One of our goals is to compare the importance of different robustness checks and recoding by method (e.g., removing outliers for instrumental variable estimation versus a difference-in-differences estimation).

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

Replication games. Our second stream to generate reproductions and replications is the replication games (RGs). RGs are one-day meet-ups open to faculty, post-docs, graduate students and other researchers. Participants join a small team of about 3–5 researchers all working in the same subfield (*e.g.*, development economics).[8,9]

Participants are offered a short list of (about 5) studies in their field of interest about three weeks before the games. They are asked to choose a paper as a team, read it and familiarize themselves with the replication package prior to the games. (See Section 2.8 for the determinants of study selection.)

Teams are asked to develop a game plan for the games; each team member should know what they are supposed to do during the games.[10] Teams then have to write a (templated - https://osf.io/8dkxc/) replication report summarizing their work and results in the following months. Of note, virtually all teams kept working on their replication after the games and some even started the re-analysis prior to the games.

Participants are offered the possibility to virtually attend RGs. In our sample of completed reports, about 68% of participants attended the games in-person, while 32% virtually attended the events.[11,12]

2.6 Replication Reports

Teams have on average worked 13 active days on their reproduction or replication (std. dev. of 24). Appendix Figure 5 shows the distribution of days across reports, trimmed at over 100 days.[13] About half the teams worked from 5 to 20 days on their replication report. Most of the remaining teams

[8] So far, teams have been as small as one individual or as large as seven.

[9] The location of RGs are chosen based on (i) local interests, (ii) geography, (iii) possibility to have the RGs as part of a major conference, and (iv) EDI considerations.

[10] A.B. assigns each participant to a team of about 3–5 participants based on research interests. A group of researchers may come as a pre-defined team, but this is not required. We do our best to team up graduate students with faculty members and senior researchers, ensuring a mix of junior and more senior economists in each team. A virtual meeting with the organizers before the games allows each team to ask questions and discuss a game plan. During the games, A.B., D.M. or one of I4R's co-directors, provide live assistance to each team.

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

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

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

worked between 25 to 85 active days.[14] Replication reports are on average 19 pages long, with a standard deviation of 14.

The goal for all replicators is clearly stated; testing whether the main claims are reproducible and robust. I4R emphasizes to replicators that the goal is NOT to show that the results are not reproducible. The goal is instead to test if the results are reproducible to recoding and/or robustness checks. This is key as some replicators might engage in reverse specification searching (i.e., selective reporting of insignificant results). Moreover, we ask replicators from I4R's Board stream to provide a pre-reanalysis plan. The game plan acts as a pre-reanalysis plan for the second stream.[15]

For both streams, I4R stresses the importance of reasonable robustness checks and recoding (Simonsohn et al. (2020)). Re-analyses are sensible tests of the research question and expected to be statistically valid and theoretically informed. This explains why replicators were asked to focus on studies in their own field and using methods they are familiar with.[16] See Brodeur et al. (2024a) (pages 133-244) for a brief description of each report.

2.7 Communication with Original Authors

Once a replication 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.[17] I4R's policy is to share the replication report with the original authors prior to publicly disseminating the report (Brodeur et al. (2023)). I4R then disseminate the replication report and the original authors' response simultaneously. Note that the replicators may change their replication 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 replicators may remain anonymous. In practice, about 11% of replicators have decided to remain anonymous.

Original authors have been incredibly fast at providing a response, perhaps since papers being reproduced and replicated have just been published. See Brodeur et al. (2024a) (pages 133-244) for a link to each authors response. Overall, about 95% of original authors that A.B. reached out to

[14] A very small fraction worked less than 5 days. This is due to the replicators not being able to conduct robustness checks. In contrast, about 8% of teams worked more than 100 days. This is typically due to uncovering major coding errors or issues with the original study and having to engage in multiple rounds of back and forth with the original authors. There is also the potential for people to have spent many days on their paper even if the number of hours were low.

[15] In practice, some teams in both streams did not write a pre-reanalysis plan and virtually all teams that did write one ended up deviating from it. The latter is because it is very unclear from only reading the original paper what is the range of re-analyses that is feasible. Replicators 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.

[16] The discussion between original authors and replicators also helped, in some instances, to resolve issues raised by the reviewers. Similarly, original authors have occasionally pointed out issues with re-analyses conducted by the replicators. See 6.2 for more information.

[17] 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.

responded to his email.[18] Of those that responded, 11% provided a very short note (e.g., thanking replicators) or mentioned they could not respond (e.g., due to personal reasons or ongoing conflict in their country), 59% provided feedback without a formal response and 30% provided a formal response.[19] See Appendix Table 6 for a breakdown by discipline.[20]

How often do replicators and original authors agree? This is a key question as replicators 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 replicators.[21] Overall, we find that there are remaining disagreements for only 23% of articles in our sample.[22] 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 replicators 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.

Overall, we observed a general lack of adversariality between original authors and replicators (Clark and Tetlock (2023)). 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 replicators is mediated by the Institute for Replication (I4R). Moreover, original authors may feel less targeted by our replicators as our aim is to mass-reproduce and replicate studies published in leading economic and political science outlets.

2.8 Study Selection

Not all studies from our targeted journals have been reproduced or replicated. This brings the questions: "Which studies are being reproduced/replicated and why?"

Our approach leads to an over-representation of studies using publicly available data, and articles using either third-party surveys and own-collected data.[23] 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.[24] Our sample should thus be viewed as

[18] This includes one author that was unreachable as he left academia.

[19] In some instances, original authors requested to see the replicators' replication package, which we provided.

[20] As a benchmark, Fišar et al. (2023) also offered original authors the possibility to provide a short formal response. Approximately 25% of authors in their sample provided a formal response.

[21] The coding was done by A.B. and three ambiguous cases were discussed at length with D.M.

[22] 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.

[23] Brodeur et al. (2024b) document for over 1,000 articles from 13 economic journals with a data availability policy that only 13% of administrative data are in articles which provide access to data and code for replication in comparison to 24% for third-party surveys and 55% for own-collected data.

[24] 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

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.

We explore in the team survey 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. Appendix Figure 6 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 (Appendix A.2).

If a large portion of replicators 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%). A small share also indicated that their choice was based on statistical power/sample size (4.6%) and/or trust of original authors (6.4%). [25]

3 Meta Database and Descriptive Statistics

In what follows, we describe the Meta Database and provide summary statistics. The main objective of this paper is to document computational reproducibility and coding errors, and robustness reproducibility/replicability in our sample. For robustness, we need to directly compare the point estimates from the original studies to the new point estimates provided in the replication reports. To do so, we build a Meta Database. The Meta Database is mainly built from three sources of raw data: (1) replication reports; (2) surveys for individual replicators; and (3) surveys for teams of replicators. We also collected information from publicly available *curricula vitae* of all original authors and replicators.

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.

[25] Appendix Table 7 explores if our sample is representative of all subfields within economics. We compare JEL Codes of economic papers that we reproduced or replicated 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 Hoces de la Guardia et al. (2024). 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.

3.1 Replication Reports

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

Only a subset of results was considered suitable for our research. We follow the following criteria. We exclude extensions of the original authors' research, effects by heterogeneity, or mediation analysis. These analyses correspond to situations where there are no "original" estimates for which we can reasonably compare the replicators' estimates. Most often, replicators included tables and figures which were the output of a computational reproduction using the original authors' replication package. These are always left out for the re-analyses.[26] After being checked, replicators would then be contacted with their subset of the Meta Database and asked to confirm our transcribing of their reports into the Meta Database.

We report some additional information in the Meta Database. We collect information on the journal, year of publication, number of Google Scholar citations at the time of entry into the Meta Database, the research field, the position of the test in the original article and the number of original authors and replicators. We also collect information from *curricula vitae* of all the original authors and replicators. We obtained information on their academic affiliation at the time of publication, their position at the main institution and year the PhD was earned. In addition, we gather for each author and replicator the following information (at the time of completing the replication): the total number of Google Scholar citations and whether they had published in a Top-5 economic journal, a leading political science journal, and/or one of the other economic journals we are reproducing/replicating.[27]

3.2 Surveys

We asked all replicators to fill out an individual survey. We also asked one author per replication report to fill out a team survey. Both surveys gave additional information on the academic and programming experience of replicators, 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 robust-

[26] Coding errors and discrepancies are also excluded from the re-analyses. We discuss coding errors and discrepancies between original authors' values in their published paper compared to what their replication package produces in Section 4.4.

[27] The Top-5 economic journals are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics* and the *Review of Economic Studies*. The leading political science journals considered here are the *American Journal of Political Science*, *American Political Science Review* and *Journal of Politics*.

ness 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 replicators participated in the RGs, whether they virtually attended, why they participated in the RGs 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?"

3.3 Descriptive Statistics

The Meta Database described above provides 6,583 re-analyzed test statistics from 103 replication reports. (Seven reports did not include robustness checks.) The other test statistics are estimates obtained by re-coding the analysis. We come back to those tests in Section 4.3.

Appendix Table 8 provides summary statistics for the full sample and by journal. In total, 83 replication reports were completed through RGs in comparison to 27 through the editorial board stream. 79 replication reports are for the field of economics against 31 for political science.

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

> **First Dependent Variable**: dummy variable indicating whether the re-analysis is statistically significant at 5% level and same sign. For this dependent variable, we only keep original estimates statistically significant at the 5% level.
>
> **Second Dependent Variable**: dummy variable indicating whether the re-analysis is statistically significant at 10% level and same sign. For this dependent variable, we only keep original estimates statistically significant at the 10% level.
>
> **Third Dependent Variable**: dummy variable indicating whether the re-analysis remains not statistically significant at 5% level. For this dependent variable, we only keep original estimates statistically insignificant at the 5% level.
>
> **Fourth Dependent Variable**: dummy variable indicating whether the re-analysis remains not statistically significant at 10% level. For this dependent variable, we only keep original estimates statistically insignificant at the 10% level.

The average number of re-analyzed test statistics per article is about 60. The standard deviation is very high (73), with a maximum of 421. This is unsurprising given that some teams, for instance, focused most of their attention to (blindly) recoding using the raw data (either provided by the authors or re-downloaded by the replicators), while other teams have focused solely on conducting

robustness checks for multiple central hypotheses.[28] As a robustness check, we deal with this issue by adjusting the weight of each test statistics by the number of such statistics in the replication report such that each replication report has the same weight.

Table 1 provides descriptive statistics. The articles in our sample are all recently published with a relatively small number of Google Scholar citations (44 on average) as of the completion of a replication report. The original authors are more experienced than replicators with 11 years of experience (*i.e.*, years since completing their Ph.D.) against 3. Original authors have on average 4,269 Google Scholar citations in comparison to 478 for replicators. Those differences are mostly driven by the larger share of graduate students among replicators than for original authors (49% against 6%). There are about 2.6 original authors per article in comparison to 3.2 for replicators. About 15% of replicators have recently published in a Top 5 or one of the three leading political science journals in our sample. Approximately 30% have published in those journals or in one of the other journals in our sample.

While replicators have less academic experience than original authors on average, their level of expertise as a programmer is quite advanced. About 10%, 48% and 33% of replicators report that their level of expertise is "Expert", "Proficient" and "Competent," respectively. Moreover, about 55% of replicators had already produced a replication package for their own work or journal publication.

3.4 Types of Re-Analyses

One of our main objectives is to document the relative importance of several robustness checks and re-analyses in impacting the magnitude and significance of the original point estimates.[29] We group the robustness checks and coding exercises conducted by the replicators into eight groups: (i) alternative control variables, (ii) changing the sample, (iii) changing the dependent variable (e.g., rescaling or using an alternative), (iv) changing the main independent variable (e.g., scaling or introducing an alternative definition), (v) changing the estimation method/model (e.g., from ordinary least squares to a probit when the outcome was a binary variable), (vi) changing the method of inference (most commonly the level of clustering but also randomization inference), (vii) change weighting scheme and (viii) replication using new data. Appendix A.7 provides a description and examples for each group. Replicators often make coding decisions which involve multiple categories simultaneously. For instance, a team of replicators may change the dependent variable, which also leads to a change in the sample size as the new dependent variable might have missing or additional values.

In practice, many replicator 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 replicators made when comparing to an original estimate

[28] As an illustrative example, imagine that an original article has three main outcome variables and relies on two main specifications. If the replicators conduct five different robustness checks for each outcome variable and specification, then this would lead to 30 re-analyzed test statistics.

[29] A medium run objective will be to document the impacts of each of those robustness checks by field and method. Our sample is currently too small to investigate these patterns.

and coded accordingly. In our sample, about 809 re-analyses fall into at least two categories of simultaneous robustness checks.

Table 2 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 replication 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 replicators are more likely to change the method of inference for economic articles than in political science.

4 Computational Reproducibility and Coding Errors

4.1 Replication Packages and Expectations

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

4.2 Computational Reproducibility

We first evaluate computational reproducibility in our sample. We rely on the Social Science Reproduction Platform (SSRP)'s 10-point scale to document computational reproducibility. This scale is useful as it is standardized and offers more details than a simple indicator for whether the results are computationally reproducible (Visit here 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. Appendix Table 9 and Appendix A.3 provide a concise overview of this assessment approach, including a detailed description of each level of reproducibility.

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 Figure 1. This figure shows the variation across papers, with the highest concentration of scores concentrated at levels 10 and 5. Indeed, over 85% 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 replicators successfully (computationally) reproduced the numerical results using the analytical data, but the raw data were not provided, while L10 indicates that the replicators 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) fails to run or produces 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 (Chang and Li (2022); Gertler et al. (2018); Wood et al. (2018)). 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 replicators. A more comparable (and recent) study is Fišar et al. (2023) 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 replicators.

Our approach also involves interacting with original authors who often help replicator teams computationally reproduce their results. We come back to this interaction between authors and replicators in Section 6.2.

4.3 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 replicators engaged in computationally reproducing and checking for coding errors using a different statistical software than the original authors. This may be due to replicators being more comfortable in another software language or the availability of specific commands (to run a robustness check).[30] Five teams also recoded the empirical analysis without looking at the authors' code/programs.

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

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

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

Additionally, one team who computationally reproduced the results using a different *version* of the software used by the authors uncovered noteworthy differences in the magnitude and significance of the estimates. About half the main claims were no longer reproducible (i.e., same sign and statistically insignificant or different sign) due to a change in the defaults used by base R when generating random numbers starting in version 3.6.0.[31] This is the only instance where using a different version of the software led to major differences in the size and significance of the estimates.

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

4.4 Coding Errors and Discrepancies

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

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

We uncover minor or major coding errors in 26 of the 110 studies in our sample, with some studies containing multiple errors. The errors can be broadly categorized into errors of the dependent variable (4 articles), main independent variable (5), control variables (10), estimation (2), inference (2), sample/observations (8) and other (5).[32] While not all coding errors lead to changes in the conclusions of the original study, we uncovered several major coding errors worth discussing. Some examples of major errors include: a very large number of duplicated observations, failing to fully interact a difference-in-differences regression specification, miscoding the treatment variable for a large number of (or all) observations, and clear model misspecification.

[31] This change is described in R version 3.6.0 release notes: https://stat.ethz.ch/pipermail/r-announce/2019/000641.html.

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

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

5 Changes in Statistical Significance, Effect Size, and the Reproducibility and Replicability Rate

In this section, we first compare statistical significance both visually and with a suite of state-of-the-art tests of publication bias. Second, we compare the relative effect size of re-analysis estimates. Third, we detail how originally published estimates 'move' from statistical significance to insignificance (and vice versa) during the re-analysis process. We then identify which types of re-analysis have the best (and worst) replication rates.

5.1 Statistical Significance

Before visually examining a distribution of the statistical significance of re-analysis estimates, it is worth thinking about what we might expect the distribution to look like absent any distortions (such as publication bias or p-hacking). There are two common ways to present the distribution: a histogram of the associated t-statistics or a histogram of p-values.[33] The formal tests of publication bias and p-hacking (discussed later) make continuity and differentiability assumptions of the t-statistics distribution (e.g., the calipers of Gerber and Malhotra (2008b)) and the p-curve (Elliott et al., 2022). These assumptions provide the rationale behind the discontinuity or caliper tests, where the absence of publication bias implies the absence of specific clusters of significance tests just above (in the case of t-statistics) or just below (in the case of p-values) arbitrarily defined statistical significance thresholds.

We present both t-statistics and p-curves in Figure 2. The top left panel provides the distribution of t-statistics from the *originally* published estimates. We restrict the visualization to $t \in [0, 5]$, present bins of width 0.1, and present an Epanechnikov kernel (with standard errors in blue) which softens valleys and peaks. We provide reference lines at the conventional two-tailed significance levels. Roughly 60%, 50%, and 25% of test statistics are significant at the 10%, 5% and 1% levels, respectively. We note especially that the distribution exhibits a peak (global maximum) just above the two-star statistical significance threshold of $t = 1.96$ and a valley before the one-star statistical significance threshold between $t = 1.0$ and $t = 1.65$. We take this as our first piece of evidence that the original studies in our sample suffer from (marginal) p-hacking and publication bias. The bottom left panel provides the equivalent p-curve for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025. We have removed $p < 0.0025$ (for a two-tailed test this is roughly $t = 3$) for illustrative purposes only, as inclusion of that mass in the left-most bar of the p-curve leads the resolution of the remaining bars to be quite low. We note that, much like the peak after $t = 1.96$ and the valley

[33] Several authors provide predictions based on statistical theory. For instance, Elliott et al. (2022) demonstrate that, irrespective of the distribution of true effects, the p-curve should exhibit a non-increasing and continuous pattern under the assumption of no p-hacking or publication bias (or both) across a wide range of circumstances.

just before, the p-curve exhibits a too-tall bar just to the left of the $p = 0.05$ threshold. Whether interpreted through the t-statistic or p-curve, we consider this to be our first piece of evidence that the sample of original studies suffers from some form of p-hacking and publication bias. We formally document the extent of p-hacking and publication bias in the original articles in Appendix A.4 which applies a suite of state-of-the-art methods for detecting p-hacking and publication bias in the presence of either. For instance, using Andrews and Kasy (2019)'s method, we document that a not statistically significant test statistic is only 17% as likely as a (very) statistically significant test statistic to be observed (published).

We present t-and-p-curves using data from Brodeur et al. (2020) in the right panels to serve as a benchmark with which to compare the original studies. The top right panel presents the distribution of t-statistics associated with hypothesis tests from articles published in 25 leading economics journals in 2015 and 2018. These articles rely on one of four popular identification methods (i.e., difference-in-differences, instrumental variable, randomized controlled trials, and regression discontinuity design). Overall, the distribution from our original studies sample is similar to that in Brodeur et al. (2020), although with visually markedly more bunching around the 5% significance threshold.[34]

Figure 3 directly compares the distribution of test statistics for original studies and our re-analyses. Just as in Figure 2, the top panels present t-distributions while the bottom panels present p-curves, and the left panels present the original studies while the right panels now present statistical significance for the re-analyses.[35] 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

[34]This could be due to at least three reasons. First, the extent of p-hacking and publication bias might be larger in our sample. Second, replicators might focus on the most central claim(s) in original studies, while Brodeur et al. (2020) focus on all claims. Arguably, the central claim(s) could be more p-hacked or suffer from more publication bias. Third, replicators might choose to reproduce studies finding an effect or focus on replicating claims that reject the null hypothesis.

[35]See Appendix Figure 7 for the weighted distributions. For the re-analyses, we use the inverse of the number of test statistics presented in the replication report to weigh observations.

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 Appendix Figure 8 reproduce the top panel of Figure 3 for economics and political science while the bottom panels of Appendix Figure 8 reproduce the bottom panel of Figure 3. 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.

Appendix Figure 9 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 Appendix Figure 9).

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 Appendix Figure 10 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. From both we see that the statistical significance of an original estimate with a re-analysis with an oppositely-signed effect are often statistically significant, but are also not the only drivers of the reduction in statistical significance when moving from original to re-analysis either as the positive t-statistics still exhibit the mass peak's disappearance when moving from original to re-analysis.

Overall, our graphical analysis suggest 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.

Table 3 explicitly presents the change in statistical significance from the original to a re-analysis at the test-statistic level.

For example, the first row indicates that of those original test statistics that were not statistically significant, 13.61% reversed sign during re-analysis while the majority (75%) remained statistically insignificant. Very few became more statistically significant at conventional levels, with roughly 5, 4, and 3 percent becoming statistically significant at the 10%, 5%, and 1% significance thresholds,

respectively. The Total column indicates that the sum of the row values is normalized to 100%.[36]

The most striking result comes from the second row (the $(0.05 < p < 0.10]$ region) for which we find that almost half (45.45%) of re-analyses become statistically insignificant while an additional 6.91% flip sign and only 28.00% remain the same level of significance. This result suggests that estimates just marginally significant at the 10% level are the most likely to lose significance.

In the third row (the $(0.01 < p < 0.05]$ region) more than a quarter (27.89%) of re-analyses become statistically insignificant, 12.06% become just marginally significant at the 10% level, 41.08% remain significant at the 5% level, and a small share (16.21%) becomes statistically significant at the 1% level.

In the fourth row (the $[0 < p < 0.01]$ region), 12.89% of re-analyses become statistically insignificant, with another 4.43% of re-analyses remaining only marginally significant at the 10% level. 8.07% fell from this highest level of statistical significance to the two-star level, while almost 70% remained statistically significant at the original level.

5.2 Relative t-Statistics

As an alternative measure of robustness reproducibility, we rely on relative t-statistics. As there can be multiple re-analysis estimates per original estimate, we first take the average of the re-analysis estimates by original estimate and take the ratio.[37] Then, in order for all re-analyses to have the same effect, we average those ratios at the re-analysis level.[38]

In the movement from original to re-analysis statistical significance, we find that on average a re-analysis finds a statistical significance around 77% the size of the original (at the paper level, 95% CI [0.72,0.83], significantly different from 100%, $p < 0.000$). This average number no doubt conceals considerable heterogeneity, which we display in Appendix Figure 11. Displayed is the distribution of the relative t-statistic between the re-analysis and original estimates (but only if the originally published estimate was statistically significant at the 5% level).

5.3 Relative Effect Size

We now turn from a comparison of statistical significance to a comparison of effect sizes between the original studies and their re-analyses (only if originally published estimates were statistically significant at the 5% level). A direct comparison is possible for most types of re-analyses, for example when replicators change the control variables, or the weighting scheme applied by the original study. We have 5,511 tests (rows/observations) which are directly comparable *and* have statistics (coefficients and p-values). Due to the freedom afforded to replicators in their re-analyses, a direct

[36] Appendix Table 11 does not make this normalization, which shows that statistical insignificance represents 31.09% of all observations.

[37] This aggregation provides advantages from reporting multiple correlated observations from the same claim/article (without distinguishing them from independent observations) and allows for straightforward calculation of confidence intervals.

[38] For example, if a re-analysis reproduces two original estimates with statistical significance $t_1 = 1.5$ and $t_2 = 2.0$ to find $t_1^1 = 1.3$, $t_1^2 = 1.2$, $t_2^1 = 1.8$, and $t_2^2 = 1.7$, then we would calculate a relative t-statistics of 0.833 for t_1 and 0.875 for t_2 and 0.854 overall.

comparison is not possible for about 15.6% ($\approx 1072/6583$) of re-analyses.[39] The following analysis includes only those tests which are directly comparable and have coefficients and p-values.

For those re-analyses for which a direct comparison is possible (and we have statistics), we present the relative effect sizes in Figure 4 (see Appendix Figure 12's first panel for the weighted version). By construction, the relative effect sizes are normalized so that a value of one equals the original effect size. A positive value indicates that the re-analysis estimate is in the same direction as in the original study (93%), while a negative value indicates that the re-analysis estimate is not in the same direction (7%). In our sample, the median relative effect size of the re-analyses is 0.98 and the mean is 0.95 (when winsorizing about and below by 5%). There is a large mass around one,[40] with about 17% of re-analyses smaller or equal to 0.5 and a remarkable 48% of re-analyses reporting a ratio greater than one, suggesting that many original authors were potentially conservative when publishing their point estimates.[41]

Appendix Figure 12's second and third panel reproduces our Figure 4 for economics and political science to find similar patterns; the median relative effect size for economics is 0.98 against 1.00 for political science. Appendix Figure 11 presents relative effect sizes at the level of the paper to find similar results as the test-statistic level.

Overall, we find large heterogeneity in relative effect size. The evidence suggests that a large share of original authors are being conservative, while over 15% of re-analyses lead to coefficients less than half the size of the original estimate.

5.4 Robustness Reproducibility and Replicability Rate

We now turn to our analysis of the reproducibility and replicability rates. Here, we rely on four distinct definitions of reproducibility and replicability (reflecting the four dependent variables listed in Section 3.3). We begin with the definition of replicability as whether (or not) a re-analysis estimate is in the same direction as its associated original estimate and remains statistically significant at the 5% level. The second definition is similar but applies to the 10% statistical significance level. In both, we exclude original estimates that were not statistically significant at the 5% (or 10%) level.

Table 4 presents our reproducibility rates (see Appendix Table 12 for the number of re-analyses in addition to the rates and Appendix Table 13 for article weighted rates).

In the first row, we show that for the full sample 71% of original results that were significant at the 5% level had a re-analysis estimate that was of the same sign and retained statistical significance at the 5% level. In the second column, which includes re-analyses that *at least* changes the control variables, this rate increases to 76%. The type of robustness check which offers the lowest replication rate was changing the dependent variable (where only 45% of originally significant estimates

[39] 423 rows have coefficients and p-values *but* are not comparable. Examples include tests where the replicators might have standardized the dependent variable, leading to the original and re-analysis coefficients being incomparable. 398 rows do not have statistics *but* are comparable. Examples include figures or non-estimated target parameters which contribute importantly to the arguments in the original paper. In 251 rows, results are neither comparable *nor* are there statistics.

[40] A two-tailed t-test comparing the relative effect sizes to a hypothesized mean of one returns $p = 0.114$.

[41] At the article level, the average relative effect size is 97% (95% CI [0.89,1.07] not statistically different from 1, $p = 0.6172$)

survived) in comparison to the approximately 75% seen for most other types of robustness checks. The third row finds a similar result for estimates originally significant at the 10% level.

In the second row, we show that for the full sample 88% of original results that were not significant at the 5% level had a re-analysis estimate that was of the same sign and retained statistical *in*significance at the 5% level. We now see that for originally statistically insignificant results, the replicability rate seems to be around 90% (as compared to the mid 70's of statistically significant ones), regardless of the type of robustness check applied (even the dependent variable, which reduced the replication rate of statistically significant original results by almost half). This trend continues in the fourth row which examines re-analysis and originally not statistically significant results at the 10% level - again with replication rates around 90%. While this means that the remaining approximately 10% of re-analyses become statistically significant, we note with interest the very different replication rates between originally statistically significant results, and statistically insignificant results.

Our rates of robustness reproducibility and replicability are relatively high in comparison to previously published replications (e.g., economics laboratory experiments using new data Camerer et al. (2016)). We provide a more direct comparison to the literature in the next subsection by splitting our re-analyses by group, including re-analyses which incorporate new data. Nonetheless, we take as a positive sign for the re-analyzed literature that the re-analysis rate is as high as it is.

The total unique articles that have been re-analyzed is 104, and while 82 articles have at least one non-comparable estimate, we find that only a small proportion (10 re-analyses) were not directly comparable for all reported re-analysis estimates.[42]

For not directly comparable re-analyses, we report the proportion that replicators indicated were of the same statistical significance as the original and same sign. For our four definitions of reproducibility and replication rates these are: When the original estimate is statistically significant at the 5% level, 85% of those we considered not directly comparable indicated their re-analysis was of the same significance (93% for the 10% level). When the original estimate was not statistically significant at the 5% level, 88% of those we considered not directly comparable indicated their re-analysis was of the same (non)significance (92% for the 10% level).

5.5 Re-Analyses, P-Hacking, and Publication Bias

We now turn to formally documenting how re-analyses display a markedly different presence of p-hacking and publication bias. We first rely on caliper tests Gerber and Malhotra (2008b) 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.

[42] A simple t-test of mean p_{rep} by whether the re-analysis was not comparable (or included elsewhere in the analysis) reveals an average difference of 0.045, where those excluded p_{rep} were *more* statistically significant (had lower values on average) than the ones included. This means that excluding those leads to reproducibility rates that are underestimated.

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

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

where $Significant_{iajt}$ 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 $Reanalysis$, 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 Appendix Table 14 for the 5% significance threshold. In column 1, we restrict the sample to [0.05±0.04]. The other columns repeat the specification in column 1 but with narrower bandwidths. We find that re-analysis test statistics are about 10-20 percentage points less likely to be statistically significant than an originally published test statistic.[43]

We then rely on an application of Andrews and Kasy (2019). The results are presented in Appendix Table 16. 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% 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% 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% from 17%. This trend continues for the $(1.645, 1.96]$ interval, where we observe a 64% relative publication probability in a re-analysis versus 38%. For the relative publication probability of test statistics significant at the 5% level, the original analyses offer an almost equal probability of 107%, whereas the re-analysis is now slightly lower than the original at 89%.[44]

5.6 Types of Re-analyses

Replicators are afforded freedom in their re-analyses. From what was ultimately submitted, we categorize each re-analysis estimate into one (or more) of seven types (we discuss that categorization

[43] See Appendix Table 15 for the 10% threshold. The point estimates for the 10% level are all small and statistically insignificant.

[44] 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.

in detail in Appendix A.7).

In this section, we investigate the differences in robustness reproducibility by *type* of re-analysis. We begin with statistical significance, where we split Figure 9 into its components in Appendix Figure 13 (which also offers an additional analysis of re-analyses that introduced new data, which is not quite as directly comparable as the remainder of those we discuss at length). We then continue onto relative effect sizes, where we split Figure 4 into its components in Appendix Figure 14 to illustrate relative effect sizes by type of re-analysis (but only when effect sizes are directly comparable - e.g., not when changing the dependent variable since that would make comparison of effect sizes dubious at best).

We find striking patterns for statistical significance. For context, while the average difference in p-values depicted in the first panel of Appendix Figure 9 is 0.053, this average masks considerable heterogeneity apparent in the figure (for example, 22% of $p_{\text{rep}} - p_{\text{orig}}$ are greater than 0.10 which guarantees a loss of statistical significance regardless of original statistical significance level). In the fourth panel of Appendix Figure 13 we present the type of re-analysis that has the most striking distribution of the p-value difference. The mean difference is 0.15, representing an average shift of 15 percentage points *less* statistically significant (towards one) following re-analysis. Unsurprisingly, this large shift is composed of shifts as large as 0.25, 0.5, and close to 1, representing a statistically insignificant re-analysis regardless of the level of significance of the original result. A total of 32% of re-analyses that change the dependent variable result in a shift greater than 0.10, enough to ensure loss of statistical significance regardless of original statistical significance level. The remaining average increase in p-values range from 0.022 (changing estimation method) to 0.085 (changing sample).

We also find striking patterns for relative effect size. For context, the average relative effect size was approximately one (see Figure 4 for the test-level and Appendix Figure 11 for the paper level). There is significant heterogeneity in the relative effect size by type of re-analysis. The type of re-analysis with the lowest relative effect size is when the dependent variable is changed, with an average of only 29.8%. The type of re-analysis with the lowest relative effect size is when the main independent variable is changed, with an average of only 81.1% (though this too could be dubious as the reported coefficient's units may have changed). The type of re-analysis with the lowest relative effect size that we considered to be valid is when changing inference method (at 91% and depicted in the fifth panel of Appendix Figure 14). In contrast, some types of re-analysis provided an average relative effect size that was *greater* than the originally published estimates (for example when changing the sample (136%) and changing the estimation method (124%)).

Table 4 provides robustness reproducibility and replicability rates by type of re-analysis for the four definitions of reproducibility and replicability. We highlight here three key results. First, robustness reproducibility rates are almost always lower for originally statistically significant results when compared to their complement. Second, within a definition (for example in the first row) reproducibility rates vary widely from 45% to 87%. Third, robustness reproducibility rates are markedly lower when replicators change the (coding of the) dependent variable (45%) and the sample (64%). In contrast, replicability rate is the highest for re-analyses that introduced new data (87%).

These findings highlight the relative importance of different types of specification checks in confirming the robustness of originally published claims. Nonetheless, this by-type analysis suffers from numerous shortcomings, which we briefly highlight. First, the re-analyses are potentially categorized into types that are 'too broad.' Going forward, additional observations will allow for finer categorization and perhaps more nuanced discussion by type of re-analysis (for instance, differentiating between increasing or decreasing the sample or differentiating between changes in time or geographical units of analysis). Additional observations may even allow for productive discussion of reproducibility rates by research field and identification method. Second, replicators did not systematically implement these types of re-analyses (nor could they have been aware of these potential categorizations, since we conceived of them only after viewing their submissions), but rather had freedom to chose which (if any) to implement, and so selection along many (perhaps unobservable dimensions) is no doubt present. Third, many of the re-analyses are implemented simultaneously, making it hard to disentangle their relative importance.

In summary, we believe the patterns displayed here point to several optimistic results for the re-analyzed body of research. While remaining aware that replicators were free to choose which types of re-analysis to attempt, the most striking result of around one third of original p-values becoming non-significant also says that two-thirds remained statistically significant - a proportion higher than seen in many previous mass replication efforts.

6 Discussion

We aim for high-quality replication 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 replicators' work. In practice, original authors and replicators 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 reproducibility and replicability. Second, A.B. or a co-director at I4R checks the tone of both the original authors' response and replicators' report.[45] Third, while replicators may make mistakes, so do reviewers and editors. Our replicators 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.[46] 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 replicator, but not to an editor or reviewer. Relatedly, our teams of replicators 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.[47] Fourth, replicators learn throughout the process and

[45] A.B. and D.M. virtually meet with original authors and replicators upon request.

[46] Or are buried in a footnote.

[47] As an example, the *Canadian Journal of Economics* writes in its instructions to reviewers that the "amount of time taken with a paper can vary enormously - anything from a couple of hours to a couple of days of full-time effort. A typical report should probably take 3 or 4 hours." See https://www.economics.ca/cpages/cje-referees. Obviously, the journals

benefit from this experience. (See Section 7.) This, in itself, is a positive contribution.

6.1 Barriers to Reproducibility and Replicability

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)" Figure 15 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 replication report for I4R. That said, the lack of raw data restricted most what replicators 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 replication and 13% of teams believed it inhibited their ability to perform extensions.[48] The remaining reasons for potential hurdles replicators could have faced (like no intermediate data, no data dictionary, unclear documentation, and/or unclear replication package) only affected more than 5% of teams in one category. 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.

6.2 Communication with Original Authors

We asked replicators 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.

Interestingly, about two-thirds of replicators mentioned that interacting with the original authors improved the quality of their report. Reasons provided include providing missing information on variables and procedure and providing data or instructions on how to obtain the data. Some teams also reported that original authors rightfully helped them adjust the tone of their report.[49]

7 On the Benefits for Replicators

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

in our sample are higher ranked and we only focus on published manuscripts.

[48] Fišar et al. (2023) also provide evidence that non-reproducibility for the journal *Management Science* is due to non-availability/accessibility of data.

[49] One set of original authors also performed at the request of the replicators additional robustness checks in their anonymous (non-public) data files.

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. Appendix Table 17 provides a breakdown of the responses. We find that about 70% of replicators 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. Replicators 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%).

8 Many-Analysts Approach: Authors' Experience and Prestige, and Data and Code Availability

In this section, we tackle additional research questions using a "many-analysts" approach where six research teams use our Meta Database to answer the same research questions. A many-analysts approach may be less vulnerable to specification searching and may mitigate the influence of individual-researcher biases, such as confirmation bias by the proponent of a theory (Hoogeveen et al. (2023)).

Our approach and research questions, which we detail below, were pre-registered.[50] See Section A.8 for more information on the methodology and illustrative examples on how a few teams coded their analysis.

8.1 Research Questions

The six meta-analyst teams tackled the following eight questions:

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

[50] Our pre-analysis plan was pre-registered here: https://osf.io/8wsqx/. The pre-analysis plan was pre-registered prior to sharing the Meta Database with analysts.

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

8.2 Data for Meta-Analysts

Meta-analysts were not given access to raw data (Meta Database, team leader surveys, individual surveys). Rather, they were given access to intermediate/analytical data which was cleaned and merged in a manner which would be consistent for analysis and meta-analysis. Giving researchers a downstream dataset allowed A.B. and D.M. to make restrictions on what the meta-analysts could do. The clearest example of this would be defining dependent variables which were not allowed to be changed - providing a consistent definition between meta-analysts. Asking certain research questions also restricted the data given to the meta analysts. These restrictions were done in ways so that any analysis done would be more comparable.

The backbone of the data provided to meta 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.[51]

8.3 Method

As in Botvinik-Nezer et al. (2020), Breznau et al. (2022), Huntington-Klein et al. (2021), Menkveld et al. (Forthcoming) and Silberzahn et al. (2018), the goal is to have each team answer each research question independently. Each team received the same instructions and data. We allow full flexibility to all teams. Teams are 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.

[51] The data given to the meta analysts changed as replication reports, team leader and individual surveys were completed. In total, we provided 13 updated Meta Databases for meta analysts between November 6th, 2023 and February 12th, 2024. We did this to give meta 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.

We provided the four dependent variables and the Meta 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.

8.4 Results

Each row in Table 5 represents one of the eight research questions. The four columns represent four broad categories regarding research teams' coefficient estimate(s) to the research question: (1) negative and statistically significant, (2) negative and not-statistically significant, (3) positive and not statistically significant and (4) positive and statistically significant. The left-to-right order of the column categories corresponds to where the associated analyst t-statistic would fall on the real number line. While the dependent variable (which does not change in this table) is the same for each team, each team chooses their own primary independent variable. Each cell represents the proportion of analyst-estimated relationships by category. The cells are team-weighted so that if a many-analyst team presents three estimates and another team presents a single estimate, the first team's estimates enter the proportion as 1/3 each.

The cell in the first row and first column tells us that 42.8% of results from the many-analysts find a negative and statistically significant relationship between the coding experience of a replicator and the reproducibility rate for estimates that were originally statistically significant at the 5% level (i.e., lower reproducibility rate for more experienced replicators).[52] From the second column, it becomes clear that, if there is a relationship between replicators experience coding and the reproducibility rate, it seems to be almost definitively negative with a combined proportion of 86% of results returned as negative and statistically significant or negative and not statistically significant at the 5% level. Only 14% of estimates find a positive relationship between the replicators experience coding and the reproducibility rate - of which none of the estimated positive relationships estimated were statistically significant. (The associated row in Appendix Table 18, which looks at the replication for the 10% threshold finds the same pattern.) This result potentially suggests that replicators with more experience coding are better suited to detecting and correcting less-than-robust estimations - possibly because of having greater expertise with the methods used.

For the second research question - whether the replication rate depends on the replicators' academic experience, a somewhat similar albeit less starkly negative result is found with some proportion moving into the positive and statistically significant category. That said, the ratio of negative-and-significant results to positive-and-significant results remains above 4 to 1. The associated row in Appendix Table 18, which looks at the replication for the 10% threshold finds the same pattern, although with 75% of many-analysts results being negatively signed.

For the third research question - whether the replication rate depends on the author's experience seems to be centered on the null. Combined, the negative and not statistically significant and the

[52] Table 5 presents results where the dependent variable takes a value of one if an originally 5% statistically significant result was reproduced by a replicator also at the 5% level. Appendix Table 18 has the same structure, but uses the 10% threshold. Appendix Table 19 then examines whether an originally *not* 5% statistically significant result was reproduced, while Appendix Table 20 continues this with the 10% threshold.

positive and not statistically significant cells contain 97.2% of results. The null hypothesis dominates in Appendix Tables 18, 19, and 20 (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 replicator and original author experience) there seems to be a positive relationship when authors have more or the same level of experience as the replicator (research question 4a and 4b). This relationship, however, weakens to a likely null when authors have comparatively less experience than their replicators. Appendix Tables 18, 19, and 20 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 replicators, the same (albeit weaker) pattern is found. When authors have more prestige than their replicators, there is a very positive relationship with replication rate. When original authors and replicators 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 replicators, then the relationship seems to be negative: 22% finding a negative and statistically significant relationship. In Appendix Table 18, 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 replication rate does not seem to have a relationship for whether the authors provided raw data (research question 6), both raw and intermediate data (research question 7) or cleaning codes (research question 8).[53,54] This result 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.[55]

To sum up, we provide evidence suggesting a negative relationship between replicators' experience and robustness reproducibility, while provision of raw or intermediate data, or the necessary cleaning codes, does not seem to affect the reproducibility of research.

9 Conclusion

> *False facts are highly injurious to the progress of science, for they often long endure*
> The Descent of Man (1871), Vol. 2, 385. by Charles Darwin

[53] The null hypothesis clearly dominates for these final research questions in Appendix Tables 18, 19, and 20 as well.

[54] In Table 21, we reproduce the analyses in Table 5 and Appendix Tables 18, 19, and 20 while only including estimates if the analyst team indicated that, in their opinion, the estimated effect size was economically meaningful. Results are broadly consistent as those described above without the restriction.

[55] Our results are consistent with Brodeur et al. (2024b) 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.

Substantial information asymmetry exists between the authors of an article and the rest of the academic community (Brodeur et al. (2016)). This leads reviewers and editors to require several robustness checks prior to acceptance. Unfortunately, reviewers may not be aware of important coding decisions and do not have access to the codes and data for their review. A related concern is that some manuscripts' programs contain major coding errors or discrepancies between the codes and the articles.

We see mass reproducibility and replicability as a new hope for the social sciences, partly dealing with the concerns highlighted above. Our paper describes a new initiative and methods to reaching the goal of mass reproducibility and replicability. While our initiative is just starting, we document several important patterns using a sample of 110 replication reports.

In terms of impact, the scale of this ongoing project has the potential to change research norms and researchers' behavior through the adoption of more rigorous methodologies and deterring questionable research practices.

While our sample of journals is selective, our results are optimistic. They suggest a high level of reproducibility and a low prevalence of major coding errors. We argue that these results and this project may have a positive effect on trust in scientific results. We ask all replicators in the individual survey about the quality of the replication package they reproduced and their views of the discipline. We find that just over 40% report that the quality of the replication package gave them a more optimistic view of the discipline. About 45% report that the quality of the replication package did not affect their views of the discipline. These results suggest that mass reproduction may significantly increase trust in scientific results among scientists.

Equally important, our project has the potential to advance science and improve equity issues. The posting of data and code and its re-analysis are likely to advance science not only through course correction but also through learning and understanding new approaches more quickly. Reproducing the original authors' work in another (open source) software also has the potential to level the playing field by allowing researchers from lower-level universities, those in developing nations, and others who cannot afford expensive licenses to learn from elite scholars.

Our results suffer from several limitations. To this date and despite some recent progress on the matter, only a small number of economics and political science journals request data and codes (Askarov et al. (2023); Brodeur et al. (2024b)), and a very small fraction check whether the results are reproducible (Vilhuber et al. (2020)). This is even though this has been a long-standing issue; in fact, Ragnar Frisch wrote as early as 1933 that "In statistical and other numerical work presented in Econometrica the original raw data will, as a rule, be published, unless their volume is excessive. This is important to stimulate criticism, control and further studies." (Introductory editorial to Econometrica). Our results should thus be seen as describing patterns for leading journals in the field of open science and data sharing. Future research should aim to draw conclusions about reproducibility and replicability more broadly by reproducing and replicating a random sample of papers from journals that do and do not have a data availability policy.

References

Andrews, Isaiah and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review*, 2019, *109* (8), 2766–94.

Ankel-Peters, Jörg, Nathan Fiala, and Florian Neubauer, "Is Economics Self-Correcting? Replications in the American Economic Review," 2023. Ruhr Economic Papers, No. 1005.

Arel-Bundock, Vincent, Ryan C Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D Stanley, "Quantitative Political Science Research is Greatly Underpowered," 2022. I4R DP 6.

Askarov, Zohid, Anthony Doucouliagos, Hristos Doucouliagos, and TD Stanley, "The Significance of Data-sharing Policy," *Journal of the European Economic Association*, 2023, *21* (3), 1191–1226.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford et al., "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams," *Nature*, 2020, *582* (7810), 84–88.

Brandon, Alec and John A List, "Markets for Replication," *Proceedings of the National Academy of Sciences*, 2015, *112* (50), 15267–15268.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Hung HV Nguyen, Muna Adem, Jule Adriaans et al., "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty," *Proceedings of the National Academy of Sciences*, 2022, *119* (44), e2203150119.

Brodeur, Abel, Anna Dreber, Fernando Hoces de la Guardia, and Edward Miguel, "Replication Games: How to Make Reproducibility Research More Systematic," *Nature*, 2023, *621* (7980), 684–686.

_ **et al.**, "Mass Reproducibility and Replicability: A New Hope," 2024. I4R Discussion Paper 107.

_ **, Mathias Lé, Marc Sangnier, and Yanos Zylberberg**, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, January 2016, *8* (1), 1–32.

_ **, Nikolai Cook, and Anthony Heyes**, "Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 2020, *110* (11), 3634–3660.

_ **,** _ **, and Carina Neisser**, "P-Hacking, Data Type and Data-Sharing Policy," *Economic Journal*, 2024, *134* (659), 985–1018.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al., "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 2016, *351* (6280), 1433–1436.

_ **,** _ **, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek et al.**, "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015," *Nature Human Behaviour*, 2018, *2* (9), 637–644.

Chang, Andrew C and Phillip Li, "Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"," *Critical Finance Review*, 2022, *11* (1), 185–206.

Christensen, Garret and Edward Miguel, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–80.

Clark, Cory J and Philip E Tetlock, "Adversarial collaboration: The next science reform," in "Ideological and Political Bias in Psychology: Nature, Scope, and Solutions," Springer, 2023, pp. 905–927.

Clemens, Michael A, "The Meaning of Failed Replications: A Review and Proposal," *Journal of Economic Surveys*, 2017, *31* (1), 326–342.

Coffman, Lucas C, Muriel Niederle, and Alistair J Wilson, "A Proposal to Organize and Promote Repli-

cations," *American Economic Review: Papers & Proceedings*, 2017, *107* (5), 41–45.

Dafoe, Allan, "Science Deserves Better: the Imperative to Share Complete Replication Files," *PS: Political Science & Politics*, 2014, *47* (1), 60–66.

de la Guardia, Fernando Hoces, Yong Sung Seung, Abel Brodeur, Edward Miguel, and Lars Vilhuber, "Standardizing and Crowd-sourcing Analysis to Assess Reproducibility in Economics," 2024. Mimeo: UC Berkeley.

Doucouliagos, C. and T.D. Stanley, "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity," *Journal of Economic Surveys*, 2011, *27* (2), 316–339.

Dreber, Anna and Magnus Johannesson, "A Framework for Evaluating Reproducibility and Replicability in Economics," *Economic Inquiry*, 2023.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich, "Detecting p-Hacking," *Econometrica*, 2022, *90* (2), 887–906.

Ferguson, Joel, Rebecca Littman, Garret Christensen, Elizabeth Levy Paluck, Nicholas Swanson, Zenan Wang, Edward Miguel, David Birke, and John-Henry Pezzuto, "Survey of Open Science Practices and Attitudes in the Social Sciences," *Nature Communications*, 2023, *14* (1), 5401.

Fišar, Miloš, Ben Greiner, Christoph Huber, Elena Katok, Ali I Ozkes, and Management Science Reproducibility Collaboration, "Reproducibility in Management Science," *Management Science*, 2023.

Freese, Jeremy and David Peterson, "Replication in Social Science," *Annual Review of Sociology*, 2017, *43*, 147–165.

Gerber, A. and N. Malhotra, "Do Statistical Reporting Standards Affect what is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 2008, *3* (3), 313–326.

Gerber, A. S. and N. Malhotra, "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?," *Sociological Methods & Research*, 2008, *37* (1), 3–30.

Gertler, Paul, Sebastian Galiani, and Mauricio Romero, "How to Make Replication the Norm," *Nature*, 2018, *554* (7693), 417–9.

Hamermesh, Daniel S, "Replication in Economics," *Canadian Journal of Economics/Revue canadienne d'économique*, 2007, *40* (3), 715–733.

Havránek, Tomas and Anna Sokolova, "Do Consumers Really Follow a Rule of Thumb? Three Thousand Estimates from 144 Studies Say "Probably Not"," *Review of Economic Dynamics*, 2020, *35*, 97–122.

Hoogeveen, Suzanne, Alexandra Sarafoglou, Balazs Aczel, Yonathan Aditya, Alexandra J Alayan, Peter J Allen, Sacha Altay, Shilaan Alzahawi et al., "A Many-Analysts Approach to the Relation Between Religiosity and Well-Being," *Religion, Brain & Behavior*, 2023, *13* (3), 237–283.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch et al., "The Influence of Hidden Researcher Decisions in Applied Microeconomics," *Economic Inquiry*, 2021, *59* (3), 944–960.

Ioannidis, John PA, Tom D Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *Economic Journal*, 2017, *127* (605), F236–F265.

King, Gary, "Replication, Replication," *PS: Political Science & Politics*, 1995, *28* (3), 444–452.

Maniadis, Zacharias and Fabio Tufano, "The Research Reproducibility Crisis and Economics of Science," *Economic Journal*, 2017, *127* (605).

_ , _ , **and John A List**, "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *American Economic Review*, 2014, *104* (1), 277–290.

Menkveld, Albert J, Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss et al., "Non-Standard Errors," *Journal of Finance*, Forthcoming.

Moonesinghe, Ramal, Muin J Khoury, and A Cecile J W Janssens, "Most Published Research Findings Are False—but a Little Replication Goes a Long Way," *PLoS Medicine*, 2007, *4* (2), e28.

Mueller-Langer, Frank, Benedikt Fecher, Dietmar Harhoff, and Gert G Wagner, "Replication Studies in Economics—How Many and Which Papers Are Chosen for Replication, and Why?," *Research Policy*, 2019, *48* (1), 62–83.

Munafò, Marcus R, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John Ioannidis, "A Manifesto for Reproducible Science," *Nature Human Behaviour*, 2017, *1* (1), 1–9.

Nosek, Brian A, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen et al., "Promoting an Open Research Culture," *Science*, 2015, *348* (6242), 1422–1425.

_ , **Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B Nuijten et al.**, "Replicability, Robustness, and Reproducibility in Psychological Science," *Annual Review of Psychology*, 2022, *73*, 719–748.

Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, 2015, *349* (6251), aac4716.

Pérignon, Christophe, Kamel Gadouche, Christophe Hurlin, Roxane Silberman, and Eric Debonnel, "Certify Reproducibility with Confidential Data," *Science*, 2019, *365* (6449), 127–128.

_ , **Olivier Akmansoy, Christophe Hurlin, Anna Dreber, Felix Holzmeister, Juergen Huber et al.**, "Computational Reproducibility in Finance: Evidence from 1,000 Tests," 2023. HEC Paris Paper.

Peterson, David and Aaron Panofsky, "Self-Correction in Science: The Diagnostic and Integrative Motives for Replication," *Social Studies of Science*, 2021, *51* (4), 583–605.

Silberzahn, Raphael, Eric L Uhlmann, Daniel P Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník et al., "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (3), 337–356.

Simonsohn, Uri, Joseph P Simmons, and Leif D Nelson, "Specification Curve Analysis," *Nature Human Behaviour*, 2020, *4* (11), 1208–1214.

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

Vilhuber, Lars, "Reproducibility and Replicability in Economics," *Harvard Data Science Review*, 2020, *2* (4).

_ , **James Turrito, and Keesler Welch**, "Report by the AEA Data Editor," *AEA Papers and Proceedings*, May 2020, *110*, 764–75.

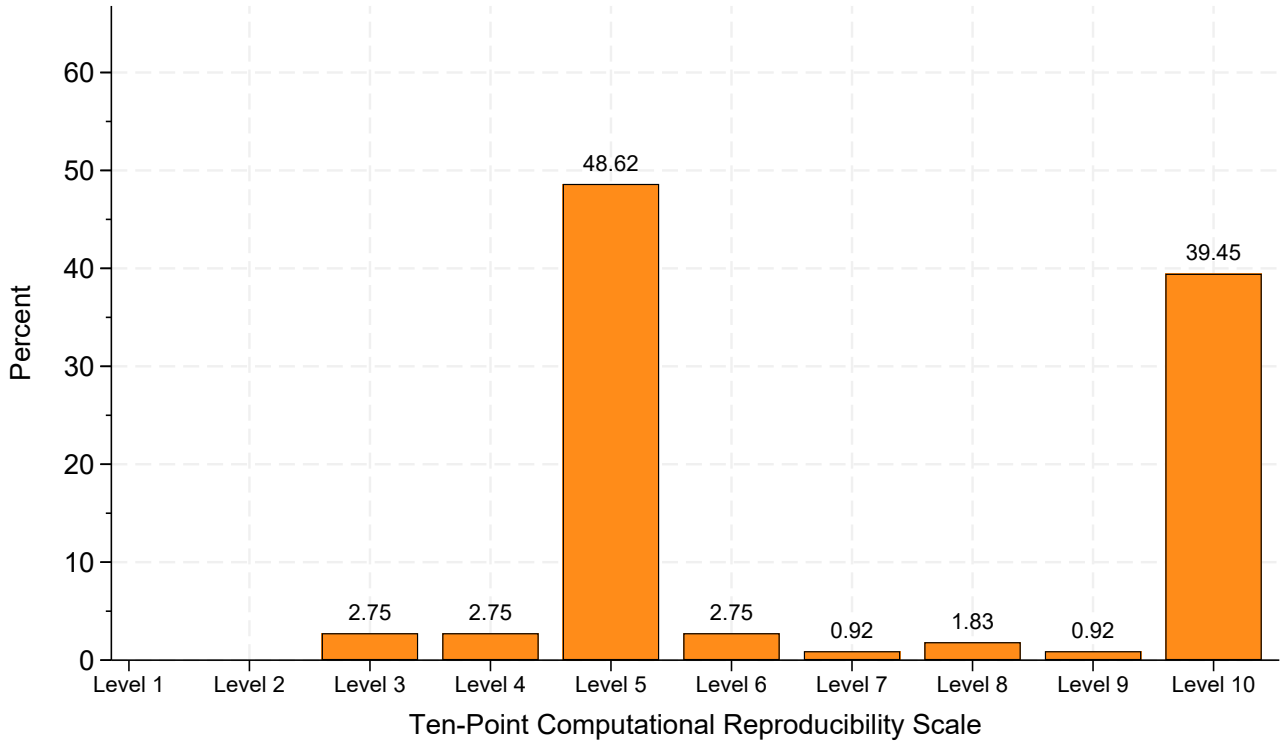
Vivalt, Eva, "Specification Searching and Significance Inflation Across Time, Methods and Disciplines," *Oxford Bulletin of Economics and Statistics*, 2019, *81* (4), 797–816.

Wood, Benjamin DK, Rui Müller, and Annette N Brown, "Push Button Replication: Is Impact Evaluation Evidence for International Development Verifiable?," *PloS one*, 2018, *13* (12), e0209416.

Young, Alwyn, "Consistency Without Inference: Instrumental Variables in Practical Application," *European Economic Review*, 2022, *147*, 104112.

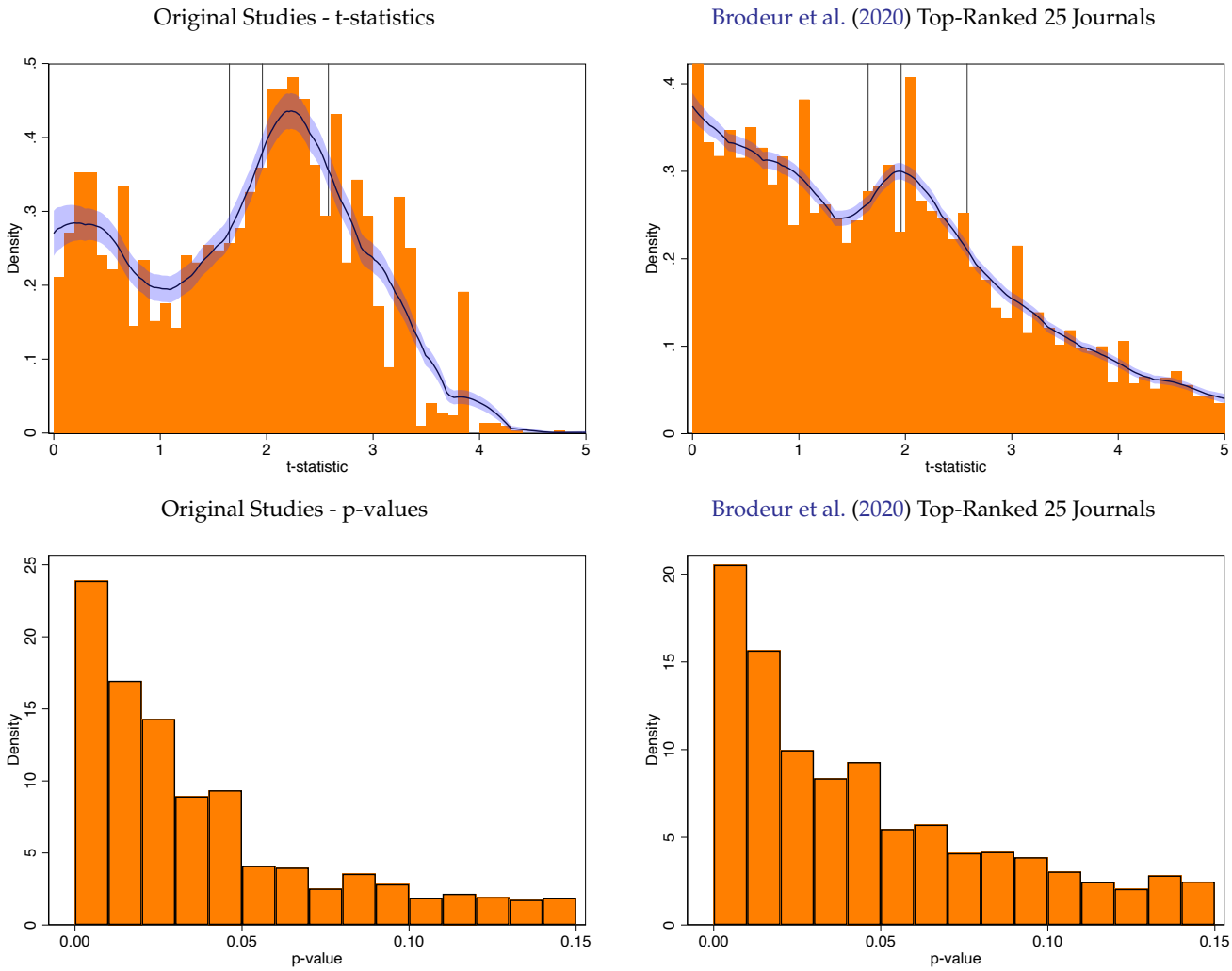
Figures

Figure 1: 10-Point Computationally Reproducibility Score



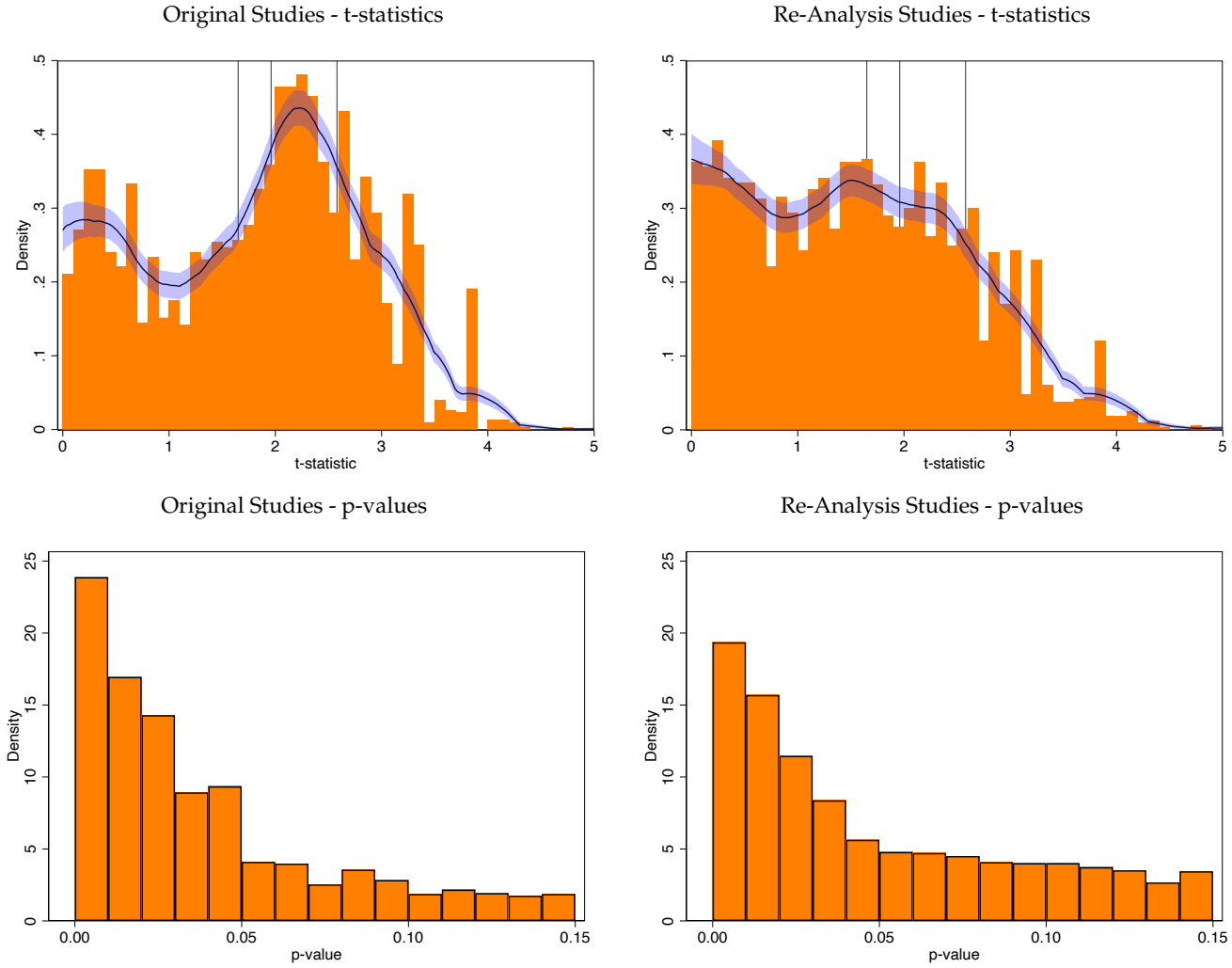
Notes: Each team assigned a reproducibility score on a scale of one to ten to the paper reproduced. See Online Appendix A.3 and Online Appendix Table 9 for a description of each score. Level 10 (L10) means that all necessary materials are available and produce consistent results with those presented in the paper, while level 5 (L5) means that analytic data sets and analysis code are available and they produce the same results as presented in the paper.

Figure 2: Distributions of t-Statistics and p-Values for Original Studies and Brodeur et al. (2020)



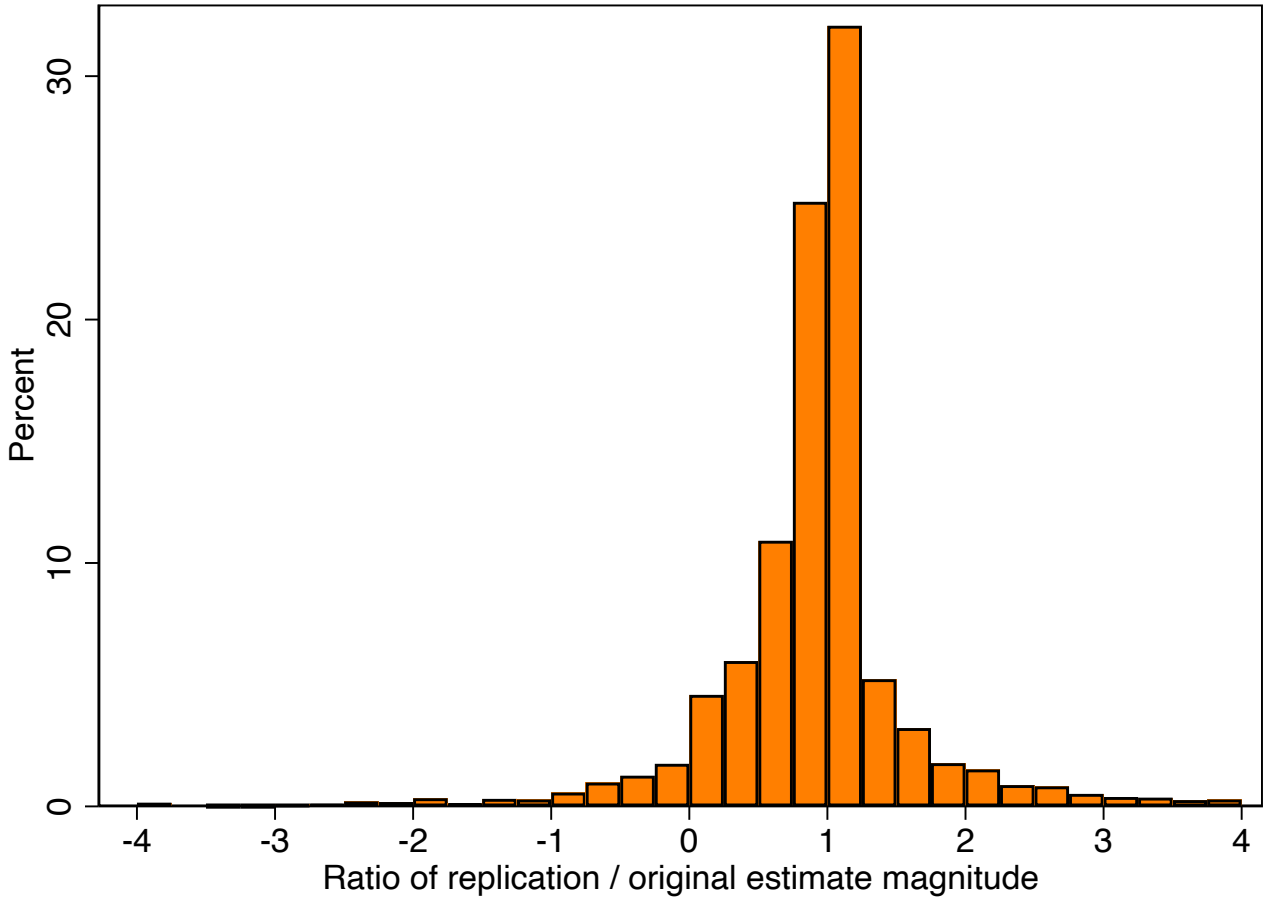
Notes: The top figures display a histogram of test statistics for $t \in [0, 5]$, with bins of width 0.1. The top left figure includes all original studies in our data set. As a comparison, the top right figure plots the corresponding histogram of z-statistics from the top-ranked 25 economics journals published in 2015 and 2018 (from Brodeur et al. (2020)). Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve. 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 Brodeur et al. (2020), respectively.

Figure 3: 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. 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.

Figure 4: Relative Effect Size



Notes: 48% of relative effect sizes are exactly equal to or greater than 1. This figure illustrates the ratio of robustness reproduction/replication 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.

Tables

Table 1: Summary Statistics: Original Authors and Replicators

	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 replicators, 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 replication report. Average GS citations in the bottom panels refers to the number of Google Scholar citations at the time the report is completed.

Table 2: 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 3: Shifts in Statistical Significance Regions

	Re-Analysis Significance Level					
Original Significance Level	Sign Change	Not Sig.	Sig. at 10%	Sig. at 5%	Sig. at 1%	Total
Not Significant	13.61	75.00	4.59	3.91	2.89	100.00
Significant at 10%	6.91	45.45	28.00	12.73	6.91	100.00
Significant at 5%	2.76	27.89	12.06	41.08	16.21	100.00
Significant at 1%	4.95	12.89	4.43	8.07	69.66	100.00
Total	7.32	37.72	7.80	14.06	33.10	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 4: Robustness Reproducibility and Replicability Rates

	(1) Full Sample	(2) Change Control	(3) Dep. Var.	(4) Change Estim.	(5) Infer. Method	(6) Ind. Var.	(7) Change Sample	(8) Change Weights	(9) New Data
Rep. if Orig. Sig. 5%									
Estimate	0.71	0.76	0.45	0.76	0.74	0.78	0.64	0.74	0.87
Confidence Interval	[0.70,0.73]	[0.73,0.79]	[0.35,0.55]	[0.72,0.81]	[0.67,0.82]	[0.72,0.85]	[0.61,0.67]	[0.64,0.85]	[0.84,0.91]
Rep. if Orig. Not Sig. 5%									
Estimate	0.88	0.92	0.80	0.85	0.88	0.77	0.86	0.97	0.83
Confidence Interval	[0.87,0.90]	[0.89,0.94]	[0.64,0.96]	[0.80,0.90]	[0.83,0.94]	[0.64,0.89]	[0.83,0.89]	[0.91,1.03]	[0.75,0.91]
Rep. if Orig. Sig. 10%									
Estimate	0.75	0.78	0.45	0.83	0.74	0.80	0.70	0.73	0.89
Confidence Interval	[0.74,0.77]	[0.75,0.81]	[0.36,0.55]	[0.79,0.86]	[0.67,0.82]	[0.74,0.86]	[0.67,0.73]	[0.63,0.83]	[0.86,0.92]
Rep. if Orig. Not Sig. 10%									
Estimate	0.85	0.88	0.93	0.82	0.84	0.54	0.82	0.92	0.75
Confidence Interval	[0.83,0.87]	[0.85,0.91]	[0.80,1.06]	[0.76,0.88]	[0.77,0.91]	[0.38,0.70]	[0.78,0.86]	[0.81,1.03]	[0.64,0.85]

Notes: Robustness reproducibility and replicability rates for four definitions by type of re-analyses. Columns present robustness reproducibility rates by type of re-analyses, which are not mutually exclusive. Columns 1-8 do not include re-analysis that use new data, while column 9 does. In (2), the re-analysis changed the control variables. In (3), the re-analysis changed the dependent variable. In (4), the re-analysis changed the estimation method. In (5), the re-analysis changed the inference method. In (6), the re-analysis changed the main independent variable. In (7), the re-analysis changed the sample. In (8), the re-analysis changed the weights applied, or applied weights for the first time. In (9), we present robustness replicability rates for re-analyses that introduced new data. 95% confidence intervals presented in square brackets.

Table 5: Many-Analysts' Replication Rate And Replicator Characteristics For Published Results Originally Statistically Significant at the 5% Level

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

Notes: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between replication rate and selected author/replicator 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 results find a positive but not statistically significant relationship. 42.78% find a negative and statistically significant relationship, and 43.33% of many-analyst results 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.

A ONLINE APPENDIX A

A.1 Authors' Contribution

Preparation of tables, figures, and manuscript. Abel Brodeur (University of Ottawa and Institute for Replication), Nikolai Cook (Wilfrid Laurier University), Derek Mikola (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 (Institute for Replication), Lars Vilhuber (Cornell University)

Analysis or interpretation of 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 (Aix Marseille Univ, CNRS, AMSE, Marseille), Lenka Fiala (University of Bergen), Jacopo Gabani (Centre for Health Economics, University of York; Department of Economics and Related Studies, University of York), Romain Gauriot (Deakin University), Goncalo Lima (European University Institute), Derek Mikola (Institute for Replication)

Author multiple replication reports. Douglas Campbell (New Economic School), Nikolai Cook (Wilfrid Laurier University), Joanne Haddad (ECARES, Université Libre de Bruxelles), Lamis Kattan (School of Foreign Service, Georgetown University Qatar), Diego Marino Fages (Durham University), Fabian Mierisch (Catholic University Eichstaett-Ingolstadt), Pu Sun (University of Ottawa), Taylor Wright (Brock University)

Author one replication report. Alejandro Abarca (Oregon State University), Mahesh Acharya (University of Calgary), Sossou Simplice Adjisse (University of Wisconsin-Madison and African School of Economics), Ahwaz Akhtar (George Washington University), Eduardo Alberto Ramirez Lizardi (University of Oslo), Sabina Albrecht (University of Queensland), Synøve Nygaard Andersen (University of Oslo), Zubaria Andlib (Lancaster University and Federal Urdu University of Arts, Science and Technology), Falak Arrora (University of Warwick), Thomas Ash (Anderson School of Management, UCLA), Etienne Bacher (Luxembourg Institute of Socio-Economic Research), Sebastian Bachler (University of Innsbruck), Félix Bacon (Laval University), Manuel Bagues (University of Warwick), Timea Balogh (UC Davis), Alisher Batmanov (UC San Diego), Mara Barschkett (Federal Institute for Population Research & DIW Berlin), B. Kaan Basdil (Mastercard), Jaromír Baxa (Institute of Economic Studies, Faculty of Social Sciences, Charles University, and Institute of Information Theory and Automation AS CR), Sascha Becker (Monash U and U Warwick), Monica Beeder (NHH Norwegian School of Economics), Louis-Philippe Beland (Carleton University), Abdel-Hamid Bello (Université Laval), Daniel Benenson

Markovits (Columbia University), Grant Benjamin (University of Toronto), Thomas Bergeron (University of Toronto), Moussa P. Blimpo (University of Toronto), Marco Binetti (University of the Bundeswehr Munich), Carl Bonander (University of Gothenburg), Joseph Bonneau (UC Davis), Endre Borbáth (Heidelberg University & WZB Berlin Social Science Center), Nicolai Topstad Borgen (Oslo Metropolitan University and University of Oslo), Solveig Topstad Borgen (University of Oslo), Jonathan Borowsky (University of Minnesota), Thomas Brailey (University of Oxford), Ryan Briggs (University of Guelph), Elisa Brini (University of Oslo and University of Florence), Myriam Brown (Laval University), Martin Brun (Universitat Autònoma de Barcelona), Stephan Bruns (Hasselt University), Nino Buliskeria (Institute of Economic Studies, Faculty of Social Sciences, Charles University), Andrea Calef (University College London), Alistair Cameron (Monash University), Pamela Campa (Stockholm Institute of Transition Economics), Santiago Campos-Rodríguez (University of California, Irvine), Giulio Giacomo Cantone (University of Sussex), Fenella Carpena (Oslo Business School, Oslo Metropolitan University), Perry Carter (Princeton University), Paul Castañeda Dower (University of Wisconsin-Madison), Ondrej Castek (Masaryk University), Jill Caviglia-Harris (Salisbury University), Gabriella Chauca Strand (University of Gothenburg), Shi Chen (Queen's University), Asya Chzhen (University of East Anglia), Jong Chung (Auburn University), Jason Collins (University of Technology Sydney), Alexander Coppock (Yale University), Hugo Cordeau (University of Toronto), Ben Couillard (University of Toronto), Jonathan Crechet (University of Ottawa), Lorenzo Crippa (University of Glasgow), Jeanne Cui (University of Ottawa), Christian Czymara (Tel Aviv University), Haley Daarstad (UC Davis), Danh Chi Dao (Queen's University), Dong Dao (University of Strathclyde and Coventry University), Marco David Schmandt (TU Berlin), Astrid de Linde (University of Oslo), Lucas De Melo (University of Nottingham, NICEP), Lachlan Deer (Tilburg University), Alexandra de Gendre (The University of Melbourne), Micole De Vera (CEMFI), Velichka Dimitrova (UCL SRI), Jan Fabian Dollbaum (European University Institute), Jan Matti Dollbaum (University of Fribourg and LMU Munich), Michael Donnelly (University of Toronto), Luu Duc Toan Huynh (Queen Mary University of London), Tsvetomira Dumbalska (University of Oxford), Jamie Duncan (University of Toronto), Kiet Tuan Duong (University of York), Yannick Dupraz (Aix Marseille Univ, CNRS, AMSE, Marseille, France), Thibaut Duprey (Bank of Canada), Christoph Dworschak (University of York), Sigmund Ellingsrud (BI Norwegian Business School), Ali Elminejad (Institute of Economic Studies, Faculty of Social Sciences, Charles University), Yasmine Eissa (American University in Cairo), Andrea Erhart (University of Innsbruck), Giulian Etingin-Frati (University of Zurich), Elaheh Fatemi-Pour (University of Warwick), Alexa Federice (UC Davis), Jan Feld (Victoria University of Wellington), Guidon Fenig (University of Ottawa), Lenka Fiala (University of Bergen), Mojtaba Firouzjaeiangalougah (Masaryk University), Erlend Fleisje (University of Oslo), Alexandre Fortier-Chouinard (University of Toronto), Julia Francesca Engel (Kiel University), Tilman Fries (LMU Munich), Reid Fortier (VisualAIM), Nadjim Fréchet (University of Montreal), Jacopo Gabani (Centre for Health Economics, University of York; Department of Economics and Related Studies, University of York), Thomas Galipeau (University of Toronto), Sebastián Gallegos (UAI Business School), Areez Gangji (Independent Researcher), Xiaoying Gao (University of York), Cloé Garnache (Oslo Metropolitan University), Attila Gáspár (HUN-REN Centre for Economic and Regional Studies), Romain Gauriot (Deakin University), Evelina Gavrilova (NHH Norwegian School of Economics), Arijit Ghosh (RWI - Leibniz Institute for Economic Research), Garreth Gibney (University of Galway), Grant Gibson (Canadian Research Data

Centre Network and McMaster University), Geir Godager (University of Oslo), Leonard Goff (University of Calgary), Da Gong (University of California, Riverside), Javier González (Department of Economics, Southern Methodist University), Jeremy D. Gretton (Public Health Agency of Canada), Cristina Griffa (University of Nottingham), Idaliya Grigoryeva (UC San Diego), Maja Grøtting (The Norwegian Institute of Public Health), Eric Guntermann (UC Berkeley), Jiaqi Guo (University of Birmingham), Alexi Gugushvili (University of Oslo), Hooman Habibnia (WU Vienna University of Economics and Business), Sonja Häffner (University of the Bundeswehr Munich), Jonathan D. Hall (University of Alabama), Olle Hammar (Linnaeus University and Institute for Futures Studies), Amund Hanson Kordt (University of Oslo), Barry Hashimoto (Independent), Jonathan S. Hartley (Stanford University), Carina I. Hausladen (ETH Zurich, work conducted while at California Institute of Technology), Tomáš Havránek (Institute of Economic Studies, Faculty of Social Sciences, Charles University), Harry He (University of California, San Diego), Matthew Hepplewhite (University of Oxford), Mario Herrera-Rodriguez (CREST-Ecole polytechnique, IP Paris), Felix Heuer (RWI – Leibniz Institute for Economic Research), Anthony Heyes (University of Birmingham), Anson T. Y. Ho (Toronto Metropolitan University), Jonathan Holmes (University of Ottawa), Armando Holzknecht (University of Innsbruck), Yu-Hsiang Dexter Hsu (National Taiwan University), Shiang-Hung Hu (California Institute of Technology), Yu-Shiuan Huang (UC Davis), Mathias Huebener (Federal Institute for Population Research (BiB) & IZA Bonn), Christoph Huber (WU Vienna University of Economics and Business), Kim P. Huynh (Bank of Canada), Zuzana Irsova (Institute of Economic Studies, Faculty of Social Sciences, Charles University, and Anglo-American University, Prague), Ozan Isler (The University of Queensland), Niklas Jakobsson (Karlstad University), Michael James Frith (University of Oslo), Raphaël Jananji (Université de Montréal), Tharaka A. Jayalath (University of Saskatchewan), Michael Jetter (University of Western Australia), Jenny John (University of Ottawa), Rachel Joy Forshaw (Heriot-Watt University), Felipe Juan (Howard University), Valon Kadriu (University of Kassel and INCHER), Sunny Karim (Carleton University), Edmund Kelly (University of Oxford), Duy Khanh Hoang Dang (King's College London), Tazia Khushboo (University of Calgary), Jin Kim (Northeastern Univeristy), Gustav Kjellsson (University of Gothenburg), Anders Kjelsrud (Oslo Metropolitan University), Jori Korpershoek (Erasmus University Rotterdam), Andreas Kotsadam (Ragnar Frisch Centre for Economic Research), Lewis Krashinsky (Princeton University), Suranjana Kundu (Indian Institute of Technology Delhi), Alexander Kustov (University of North Carolina at Charlotte), Nurlan Lalayev (Monash University), Audrée Langlois (Université Laval), Jill Laufer (UC Davis), Blake Lee-Whiting (University of Toronto), Andreas Leibing (DIW Berlin and Freie Universität Berlin), Gabriel Lenz (UC Berkeley), Joel Levin (UC San Diego), Peng Li (University of Bath), Tongzhe Li (University of Guelph), Yuchen Lin (University of Warwick), Goncalo Lima (European University Institute), Ariel Listo (University of Maryland), Dan Liu (Australian National University), Xuewen Lu (University of Calgary), Elvina Lukmanova (New Economic School), Alex Luscombe (University of Toronto), Lester R. Lusher (University of Pittsburgh), Ke Lyu (University of Nevada, Reno), Hai Ma (McGill University), Nicolas Mäder (Knauss School of Business, University of San Diego), Clifton Makate (Norwegian University of Life Sciences and Norwegian Geotechnical Institute), Alice Malmberg (UC Davis), Adit Maitra (The University of Melbourne), Marco Mandas (University of Cagliari), Jan Marcus (Freie Universität Berlin), Shushanik Margaryan (University of Potsdam), Lili Márk (Central European University), Diego Marino Fages (Durham University), Andres Martignano (University of Nottingham), Abi-

gail Marsh (Univerisy of Ottawa), Isabella Masetto (London School of Economics and Political Science), Anthony McCanny (University of Toronto), Emma McManus (Health Organisation, Policy and Economics, The University of Manchester), Ryan McWay (University of Minnesota), Lennard Metson (University of Oxford), Fabian Mierisch (Catholic University Eichstaett-Ingolstadt), Jonas Minet Kinge (University of Oslo), Sumit Mishra (Krea University), Myra Mohnen (University of Ottawa), Jakob Möller (WU Vienna University of Economics and Business), Rosalie Montambeault (Université Laval), Sébastien Montpetit (Toulouse School of Economics), Louis-Philippe Morin (University of Ottawa), Todd Morris (University of Queensland), Scott Moser (University of Nottingham, School of Politics and International Relations), Fabio Motoki (Norwich Business School at the University of East Anglia), Lucija Muehlenbachs (University of Calgary and Resources for the Future), Andreea Musulan (University of Toronto), Marco Musumeci (Erasmus University Rotterdam), Munirul Nabin (Deakin University), Karim Nchare (Vanderbilt University), Florian Neubauer (RWI - Leibniz Institute for Economic Research), Quan M. P. Nguyen (University of Sussex), Tuan Nguyen (Hasselt University), Viet Nguyen-Tien (London School of Economics), Ali Niazi (University of Calgary), Giorgi Nikolaishvili (University of Oregon), Ardyn Nordstrom (Carleton University), Patrick Nüß (Kiel University), Angela Odermatt (University of Oxford), Matt Olson (University of Pennsylvania Wharton), Henning Øien (Department of Health Management and Health Economics, University of Oslo), Tim Ölkers (University of Göttingen), Miquel Oliver i Vert (University of Nottingham), Emre Oral (University of Mannheim), Christian Oswald (University of the Bundeswehr Munich), Ali Ousman (McGill University), Ömer Özak (Department of Economics, Southern Methodist University, IZA and GLO), Shubham Pandey (Indian Institute of Technology Bombay), Alexandre Pavlov (Université de Montréal), Martino Pelli (Asian Development Bank, Université de Sherbrooke), Romeo Penheiro (University of Houston), RyuGyung Park (UC Davis), Eva Pérez Martel (Universitat Autonoma de Barcelona), Jörg Ankel-Peters (RWI - Leibniz Institute for Economic Research), Tereza Petrovičová (UCSD), Linh Phan (UC Davis), Alexa Prettyman (Towson University), Jakub Procházka (Masaryk University), Aqila Putri (University of Maryland), Julian Quandt (WU Vienna University of Economics and Business), Kangyu Qiu (University of Calgary), Loan Quynh Thi Nguyen (Queen Mary University of London), Andaleeb Rahman (Cornell University), Carson H. Rea (Emory University), Adam Reiremo (University of Oslo), Laëtitia Renée (Université de Montréal), Joseph Richardson (Lancaster University), Nicholas Rivers (University of Ottawa), Bruno Rodrigues (Ministry of Research and Higher Education, Luxembourg), William Roelofs (University of Toronto), Tobias Roemer (University of Oxford), Ole Rogeberg (Ragnar Frisch Centre for Economic Research), Julian Rose (RWI - Leibniz Institute for Economic Research), Andrew Roskos-Ewoldsen (UC Davis), Paul Rosmer (Ludwig Maximilian University of Munich), Barbara Sabada (Bank of Canada), Soodeh Saberian (University of Manitoba), Nicolas Salamanca (The University of Melbourne), Georg Sator (University of Nottingham), Daniel Scates (UC Davis), Elmar Schlüter (Justus Liebig University, Giessen), Cameron Sells (Indepenent Researcher), Sharmi Sen (Monash University), Ritika Sethi (Rice University), Anna Shcherbiak (WU Vienna University of Economics and Business), Moyosore Sogaolu (McMaster University), Matt Soosalu (Carleton University), Erik Ø. Sørensen (NHH Norwegian School of Economics), Manali Sovani (Tufts University), Noah Spencer (University of Toronto), Stefan Staubli (University of Calgary), Renske Stans (Erasmus University Rotterdam), Anya Stewart (UC Davis), Felix Stips (Luxembourg Institute of Socio-Economic Research), Kieran Stockley (University of Nottingham), Stephenson Strobel

(Cornell University), Ethan Struby (Carleton College, Boston College, and Minnesota Supercomputing Institute), John Tang (Utrecht University), Idil Tanrisever (University of California, Irvine), Thomas Tao Yang (Australian National University), Ipek Tastan (University of Calgary), Dejan Tatić (WU Vienna University of Economics and Business), Benjamin Tatlow (University of Nottingham), Féraud Tchuisseu Seuyong (Université de Montréal), Rémi Thériault (Université du Québec à Montréal), Vincent Thivierge (University of California, Berkeley), Wenjie Tian (University of Ottawa), Filip-Mihai Toma (California Institute of Technology), Maddalena Totarelli (University of Amsterdam), Van-Anh Tran (Monash University), Hung Truong (Simon Fraser University), Nikita Tsoy (INSAIT, Sofia University), Kerem Tuzcuoglu (Bank of Canada), Diego Ubfal (World Bank), Laura Villalobos (Salisbury University), Julian Walterskirchen (University of the Bundeswehr Munich), Joseph Tao-yi Wang (National Taiwan University), Vasudha Wattal (The University of Manchester), Matthew D. Webb (Carleton University), Bryan Weber (College of Staten Island - CUNY), Reinhard Weisser (University of the West of England), Wei-Chien Weng (National Taiwan University), 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 (Ludwig Maximilian University of Munich and ifo 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 (National Taiwan University), Myra Yazbeck (University of Ottawa), Luther Yap (Princeton University), Kareman Yassin (Alexandria University and Carleton 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 (ECARES, Université Libre de Bruxelles), Pu Sun (University of Ottawa)

Local organizer Replication Games. Marie Connolly (UQAM), Romain Gauriot (Deakin University), Leonard Goff (University of Calgary), Christoph Huber (WU Vienna University of Economics and Business), Andreas Kotsadam (Ragnar Frisch Centre for Economic Research), Diego Marino Fages (Durham University)

A.2 Guidelines for Choosing a Study

For the replication games, participants are assigned to a small team of about 3–5 researchers. Ideally, all researchers on a team are working in a similar field/subfield and have similarly preferred programming languages. Participants are then offered a short list of (about 5) studies in their field of interest about three weeks before the games. They are asked to choose a paper as a team. They are provided the following guidelines for choosing a study:

> Please read the Readme files to check for (i) (too) large data set/running time, (ii) software being used, (iii) completeness of the raw data. If none of these studies is interesting enough, please let us know ASAP so that we can suggest other studies with publicly available codes/data.
>
> The choice of which paper to replicate is very important. Avoid choosing a study using (i) methods you are not familiar with, (ii) use super computer or very long running time, (iii) only share final data set or (iv) data set in a language none of you can read.
>
> Last, avoid choosing a paper for which you have a conflict of interest (e.g., friend, coauthor).

A.3 Computational Reproducibility

Computational reproducibility is defined following the Guide for Accelerating Computational Reproducibility in the Social Sciences (https://bitss.github.io/ACRE/).

Note that the assessment is made at the journal article level using responses from the team survey. The assessment employs a 10-point scale, with 1 indicating that, given the existing conditions, replicators have no access to any reproduction package. On the other end of the scale at level 10, the replicators have full access to all essential materials, enabling faithful computational reproduction starting from the raw data.

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.

A.4 Formal Tests for P-Hacking and Publication Bias

We adopt diverse methodologies introduced by Brodeur et al. (2020) and Elliott et al. (2022) as our foundation. Our initial focus is on randomization tests, as designed by Brodeur et al. (2020) 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's 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 Table 22 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, 77.9% 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 Elliott et al. (2022), 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 Appendix Figure 16. This figure present 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 monotone (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.

A.5 Robustness Reproducibility for Figures

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

A.6 Non Comparable Re-Analysis

As mentioned earlier, a direct comparison is not possible between the original analysis and the replicators' analysis for about 15% of re-analyses. In applied microeconomics and politics papers, this may be due to a change in the estimator or a change in the scale of the dependent or main independent variable.

There are also scenarios where the original paper uses methods where coefficient estimates and p-values are not the objective of the analysis. This is apparent in a few empirical macroeconomics papers teams looked at. A common "robustness check" would be to adjust parameters which enter a model, possibly using accepted values in the field or estimated from an alternative dataset.

A.7 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. Replicators 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 replicators 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 and using a composite index of several indicators as the dependent variable.

Change (coding of) the main independent variable: The replicators may change the coding of the main independent variable. In our sample, there are 264 new re-analyses involving changing the main independent variable. An example is using a continuous variable instead of a dummy variable for treatment.

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

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

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

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

A.8 Many-Analysts: Methodology

A.8.1 Team Construction

We asked a subset of coauthors on this paper (replicators) if they would like to help analyse our Meta 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 *and* (2) had completed, or were near completion of, their replication report. We sent invitations (a simple sign-up form) in an email which also asked the replicators to respond to individual and team leader surveys which formed parts of our previous analysis. As a crude lower bound on the number of individuals who were invited between September 21st and October 8th, we had 87 individual surveys completed.[56] When we closed the period for volunteering on October 8th, we had 10 individuals sign-up as "meta-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 Meta-Analysts Research.

A.8.2 Meta Database

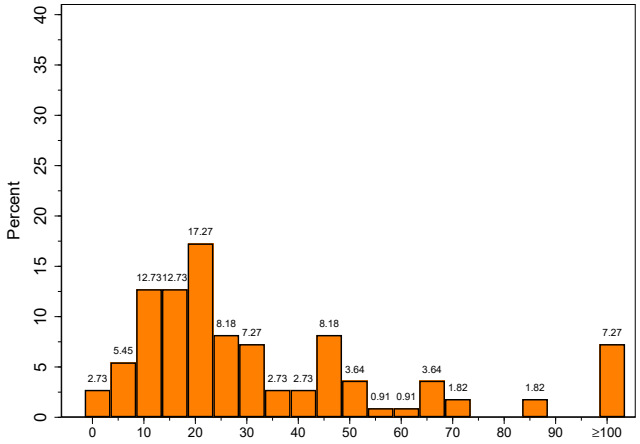
After pre-registering our procedures (https://osf.io/8wsqx/), we provided all of our analyst teams with the link to a folder which contains four documents: (1) Meta Database as a *.dta document; (2) Clarifying Questions and Comments document with a *.txt extension; (3) Reporting Guidelines excel file showing how we liked teams to report their results; and (4) an Analysts Document for Variables and Variable Labels as a *.docx.

While we had constructed the majority of the Meta Database when sharing it to all teams (October 2023), we still had replication reports and surveys being entered. That is, the dataset initially provided to all teams was not yet completely built.

[56] We also had 36 team leader surveys completed in that time. With an average of 3 people per team, another crude estimate would be about 108 individuals.

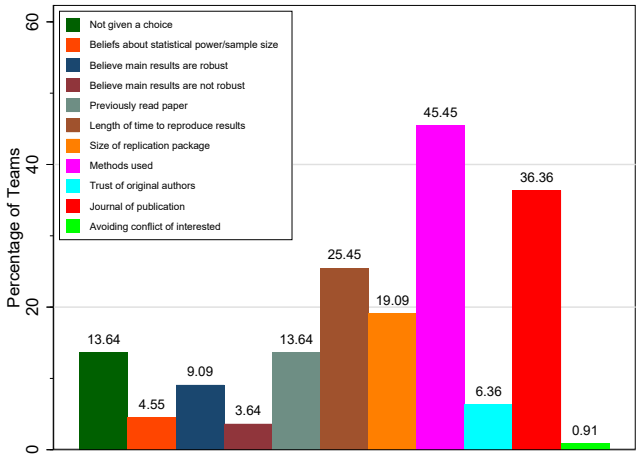
Appendix Figures

Figure 5: Histogram of Number of Active Work Days



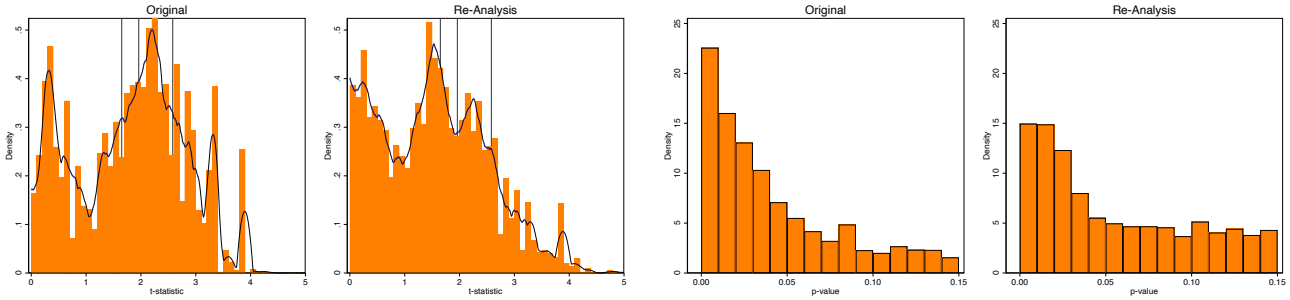
Notes: Data collected *via* survey of our replicators after completing their reports. This figure illustrates the number of active days each team worked on their report.

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



Notes: Data collected *via* survey of our replicators 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?"

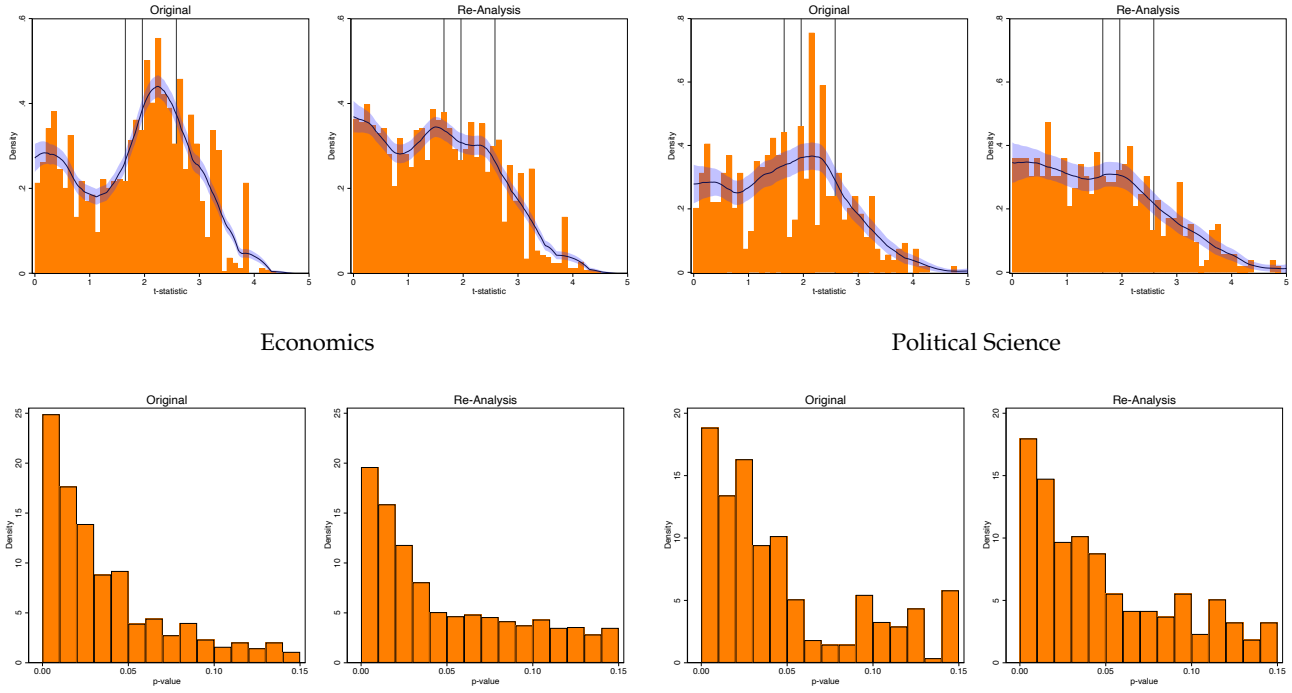
Figure 7: Weighted Distributions of Statistics for Original Studies and Re-Analyses



t-statistic p-value

Notes: Top panels 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. We use the inverse of the number of tests presented in the same article to weight observations. Bottom panels 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.

Figure 8: Distributions of t-Statistics and p-values for Original Studies and Re-Analyses

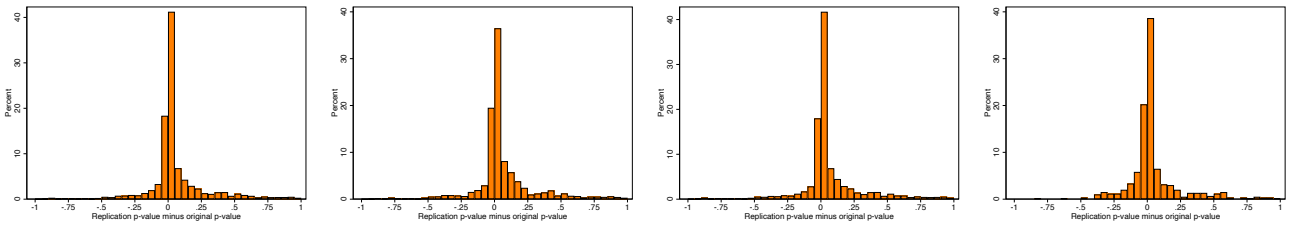


Economics Political Science

Economics Political Science

Notes: We restrict the sample to articles published in the indicated field. journals. Top panels display histograms of test statistics for $t \in [0, 5]$, with bins of width 0.1 respectively. Vertical reference lines are displayed at conventional two-tailed significance levels. We superimpose an Epanechnikov kernel density curve. Bottom panels display histograms of test statistics for p-values $\in [0.0025, 0.1500]$, with bins of width 0.0025.

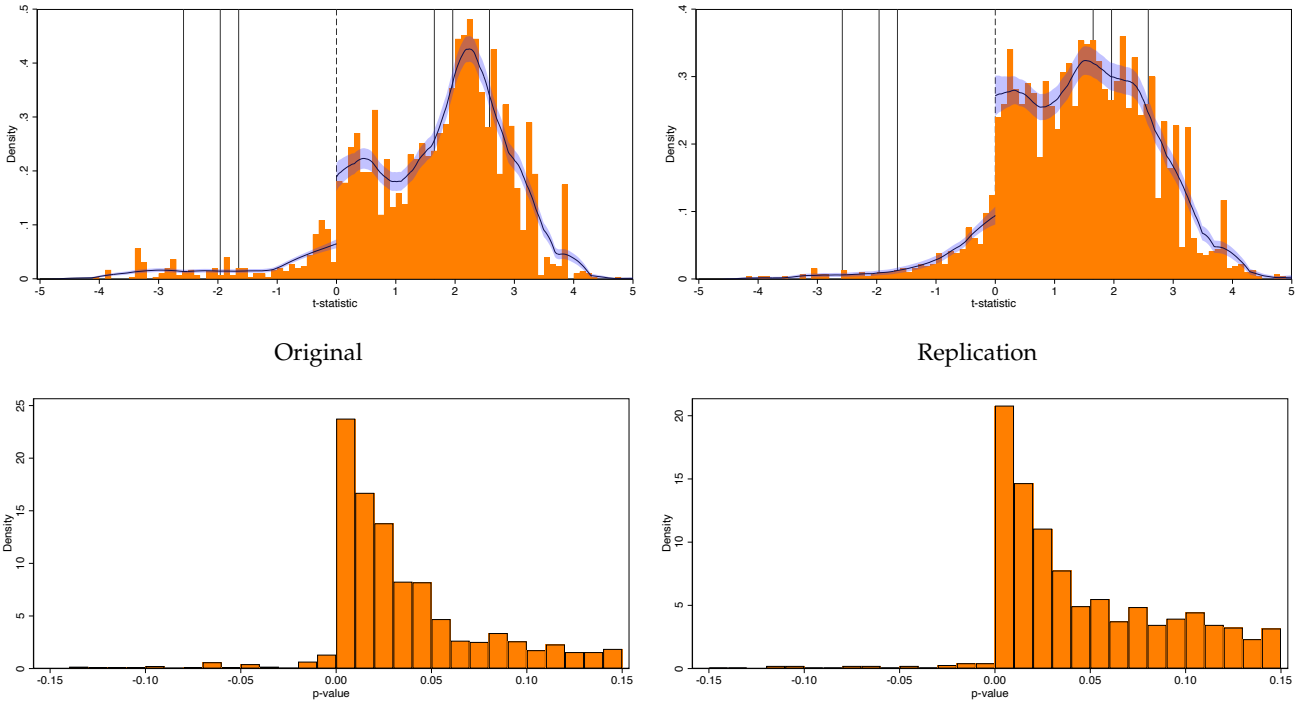
Figure 9: Distribution of $p_{\text{replication}} - p_{\text{original}}$ by Weights and Fields



Full Sample | Weights | Economics | Political Science

Second panel: We use the inverse of the number of test statistics in each replication 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 $(p_{\text{replication}} - p_{\text{original}})$

Figure 10: t and p-curves where negative represents a sign change from original to replicator

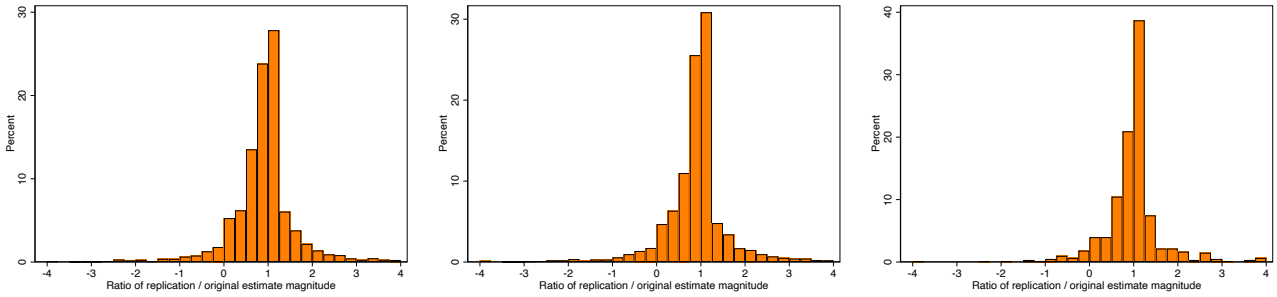


Original | Replication

Original | Replication

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 replicators' 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, 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 replicators' estimates. If the replicator's estimated effect was of the opposite sign than the originally published estimate, we set the sign of the associated statistic to be negative.

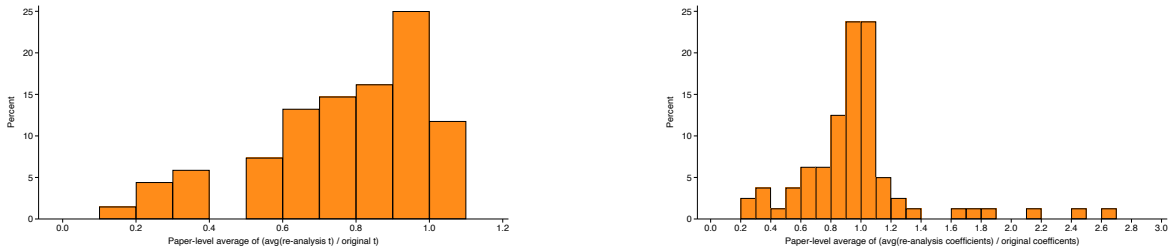
Figure 12: Relative Effect Size Weights and Fields



Weights | Economics | Political Science

First panel: We use the inverse of the number of test statistics in each replication report to weight observations. Second and third panel: The sample is restricted to original articles published in the indicated field. All panels: This figure illustrates the ratio of reproduction/replication 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.

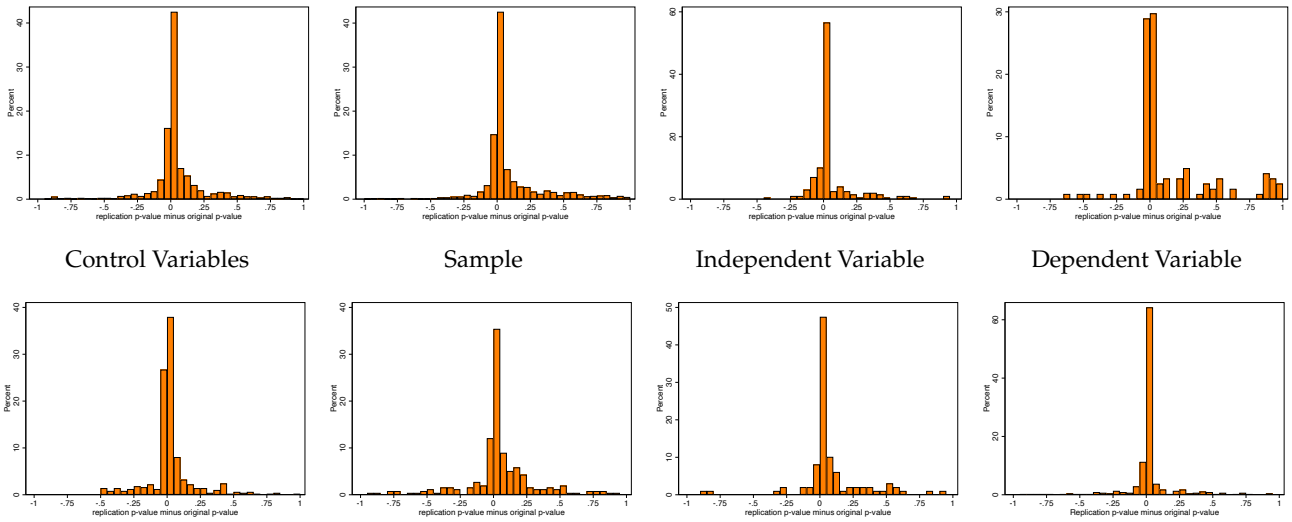
Figure 11: Relative t-statistics and effect sizes at the paper level



Relative t-statistics | Relative effect size

Notes: The data contains multiple re-analysis t-statistics (effect sizes) for every original result. We first take the average of the re-analysis t-statistics (effect sizes) by original result (if the re-analysis and original coefficients were of opposite sign, we assign the original to be positive and the re-analysis to be negative, otherwise everything is in absolute terms). We then take this average and divide by the original result . These values are then averaged at the paper level to get a paper's relative t-statistic (effect size) when replicated.

Figure 13: Difference in *p*-values

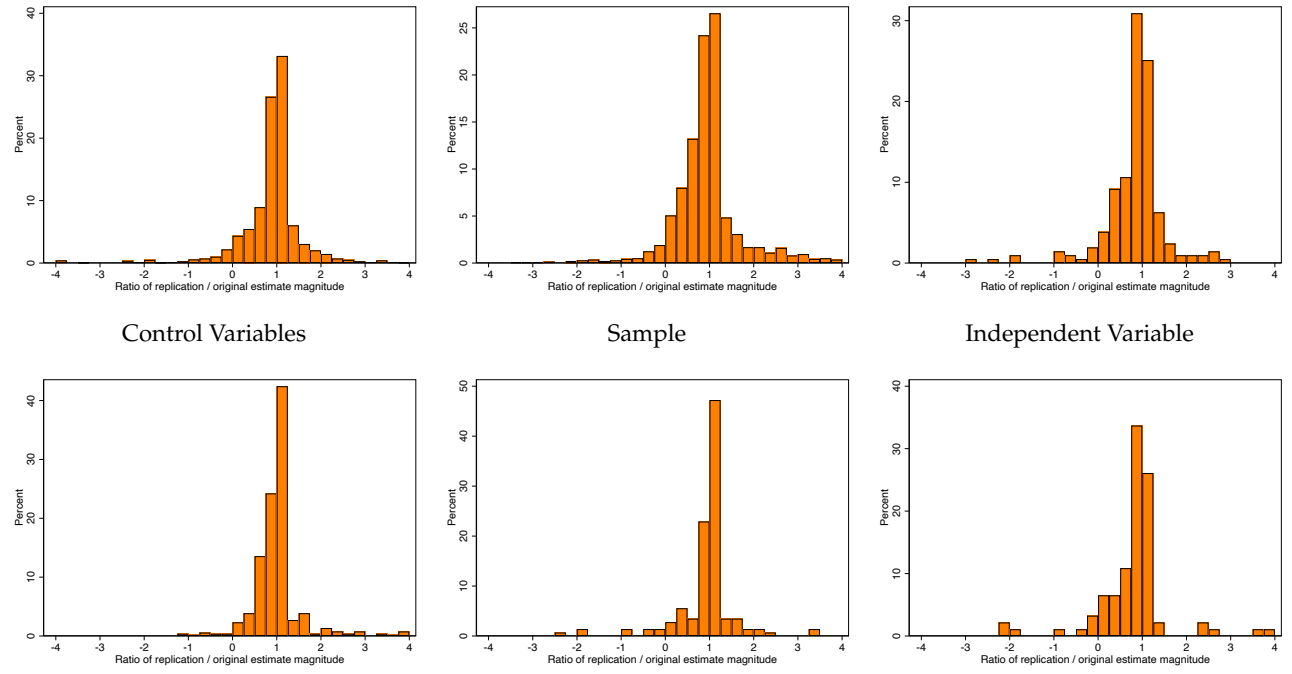


Control Variables | Sample | Independent Variable | Dependent Variable

Estimation Method | Inference Method | Weighting Scheme | New Data

Notes: In each panel, the sample is restricted to re-analyses for which the replicators changed the indicated research aspect. Depicted are the differences in p-values of the reproduction/replication and original estimates.

Figure 14: Relative Effect Size Components

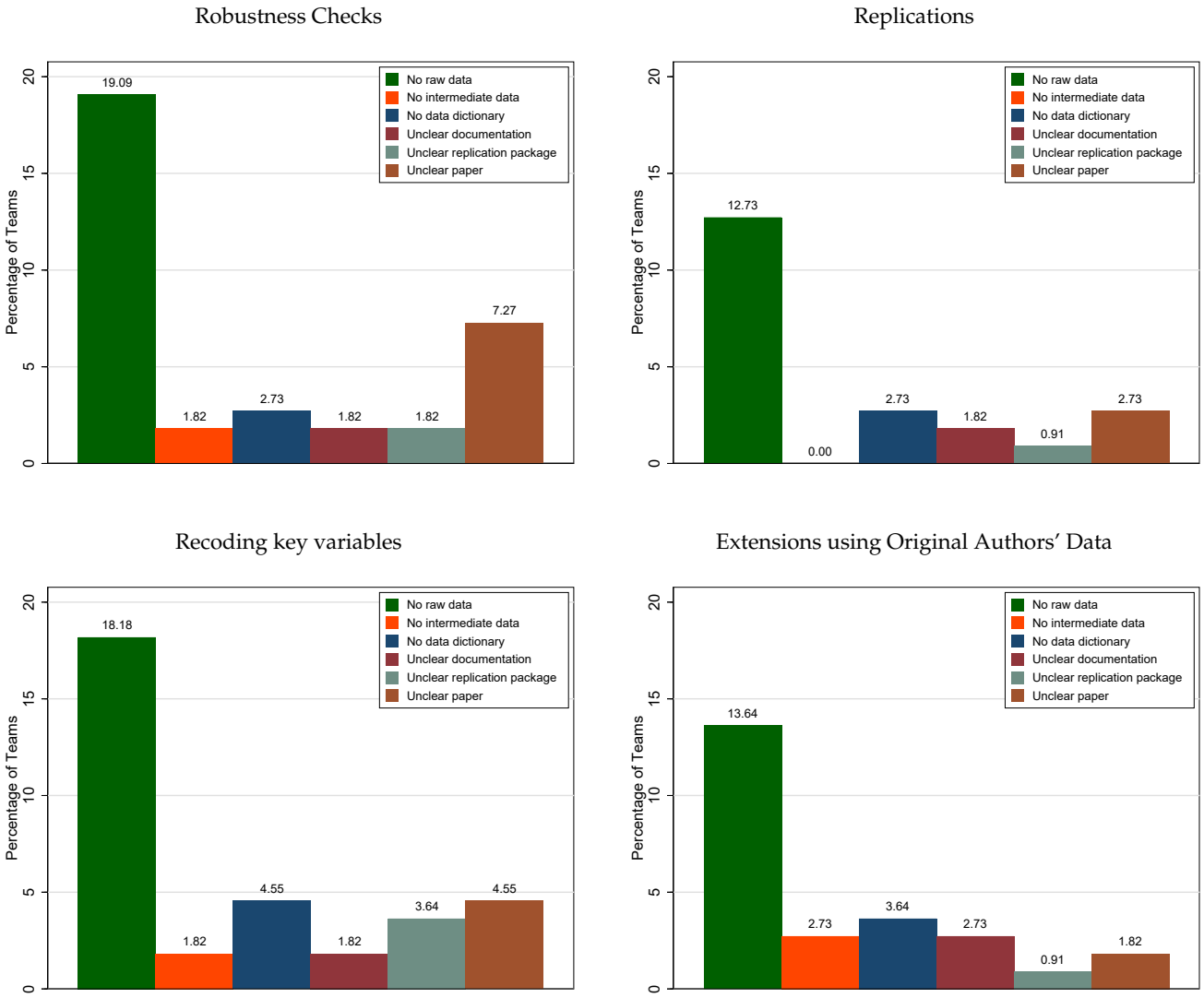


Control Variables | Sample | Independent Variable

Estimation Method | Inference Method | Weighting Scheme

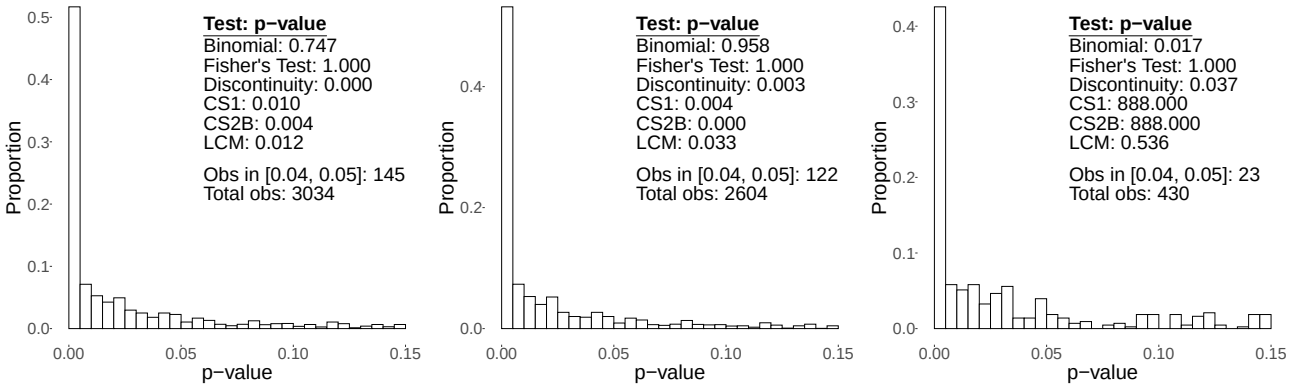
Notes: In each panel, the sample is restricted to re-analyses for which the replicators changed the indicated research aspect. Depicted are the ratio of reproduction/replication and original estimates. The standardized effect sizes are normalized so that 1 equals the original effect size. A positive 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.

Figure 15: 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.

Figure 16: Applying Elliott et al. (2022)'s Tests



Notes: This figure present p-curves and results for the battery of p-hacking tests proposed in Elliott et al. (2022) 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.

Appendix Tables

Table 6: 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 9: Levels of 10-point Computational Reproducibility Scale

	Availability of materials, and reproducibility									
	Analysis Code		**Analysis Data**		**CRA**	**Cleaning Code**		**Raw Data**		**CRR**
	P	C	P	C		P	C	P	C	
L1: No materials	–	–	–	–	–	–	–	–	–	–
L2: Only code	✓	✓	–	–	–	–	–	–	–	–
L3: Partial analysis data & code	✓	✓	✓	–	–	–	–	–	–	–
L4: All analysis data & code	✓	✓	✓	✓	–	–	–	–	–	–
L5: Reproducible from analysis	✓	✓	✓	✓	✓	–	–	–	–	–
L6: All cleaning code	✓	✓	✓	✓	–	✓	✓	–	–	–
L7: Some raw data	✓	✓	✓	✓	–	✓	✓	✓	–	–
L8: All raw data	✓	✓	✓	✓	–	✓	✓	✓	✓	–
L9: All raw data + CRA	✓	✓	✓	✓	✓	✓	✓	✓	✓	–
L10: Reproducible from raw data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Computationally Reproducible from Analytic data (CRA): The output can be reproduced with minimal effort starting from the analytic datasets. Computationally Reproducible from Raw data (CRR): The output can be reproduced with minimal effort from the raw datasets. **P** denotes "partial", **C** denotes "complete".

Table 7: 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 (Hoces de la Guardia et al. (2024)). The JEL Codes are only available for some of the economic journals.

Table 8: 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 10: 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 11: Shifts in Statistical Significance Regions

	Re-Analysis Significance Level					
Original Significance Level	Sign Change	Not Sig.	Sig. at 10%	Sig. at 5%	Sig. at 1%	Total
Not Significant	4.23	23.31	1.43	1.22	0.90	31.09
Significant at 10%	0.50	3.30	2.04	0.93	0.50	7.27
Significant at 5%	0.58	5.87	2.54	8.64	3.41	21.04
Significant at 1%	2.01	5.23	1.80	3.28	28.28	40.60
Total	7.32	37.72	7.80	14.06	33.10	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 12: Robustness Reproducibility and Replicability Rates (with counts)

	(1) Full Sample	(2) Change Control	(3) Dep. Var.	(4) Change Estim.	(5) Infer. Method	(6) Ind. Var.	(7) Change Sample	(8) Change Weights	(9) New Data
Rep. if Orig. Sig. 5%									
Estimates	0.71	0.76	0.45	0.76	0.74	0.78	0.64	0.74	0.87
Confidence Intervals	[0.70,0.73]	[0.73,0.79]	[0.35,0.55]	[0.72,0.81]	[0.67,0.82]	[0.72,0.85]	[0.61,0.67]	[0.64,0.85]	[0.84,0.91]
Observations	2552	833	96	348	121	160	945	66	370
Rep. if Orig. Not Sig. 5%									
Estimates	0.88	0.92	0.80	0.85	0.88	0.77	0.86	0.97	0.83
	[0.87,0.90]	[0.89,0.94]	[0.64,0.96]	[0.80,0.90]	[0.83,0.94]	[0.64,0.89]	[0.83,0.89]	[0.91,1.03]	[0.75,0.91]
Observations	1453	594	25	174	129	47	468	33	83
Rep. if Orig. Sig. 10%									
Estimates	0.75	0.78	0.45	0.83	0.74	0.80	0.70	0.73	0.89
Confidence Intervals	[0.74,0.77]	[0.75,0.81]	[0.36,0.55]	[0.79,0.86]	[0.67,0.82]	[0.74,0.86]	[0.67,0.73]	[0.63,0.83]	[0.86,0.92]
Observations	2826	932	106	373	137	168	1068	74	382
Rep. if Orig. Not Sig. 10%									
Estimates	0.85	0.88	0.93	0.82	0.84	0.54	0.82	0.92	0.75
Confidence Intervals	[0.83,0.87]	[0.85,0.91]	[0.80,1.06]	[0.76,0.88]	[0.77,0.91]	[0.38,0.70]	[0.78,0.86]	[0.81,1.03]	[0.64,0.85]
Observations	1179	495	15	149	113	39	345	25	71

Notes: Robustness reproducibility and replicability rates for four definitions by type of re-analyses. Columns present robustness reproducibility rates by type of re-analyses, which are not mutually exclusive. Columns 1-8 do not include re-analysis that use new data, while column 9 does. In (2), the re-analysis changed the control variables. In (3), the re-analysis changed the dependent variable. In (4), the re-analysis changed the estimation method. In (5), the re-analysis changed the inference method. In (6), the re-analysis changed the main independent variable. In (7), the re-analysis changed the sample. In (8), the re-analysis changed the weights applied, or applied weights for the first time. In (9), we present robustness replicability rates for re-analyses that introduced new data. 95% confidence intervals presented in square brackets.

Table 13: Robustness Reproducibility and Replicability Rates (with counts, weighted)

	(1) Full Sample	(2) Change Control	(3) Dep. Var.	(4) Change Estim.	(5) Infer. Method	(6) Ind. Var.	(7) Change Sample	(8) Change Weights	(9) New Data
Rep. if Orig. Sig. 5%									
Estimate	0.67	0.68	0.42	0.71	0.69	0.83	0.62	0.81	0.91
Confidence Intervel	[0.66,0.69]	[0.65,0.71]	[0.32,0.52]	[0.66,0.76]	[0.60,0.77]	[0.77,0.89]	[0.59,0.65]	[0.71,0.91]	[0.88,0.94]
Observations	2552	833	96	348	121	160	945	66	370
Rep. if Orig. Not Sig. 5%									
Estimate	0.88	0.92	0.74	0.87	0.88	0.72	0.85	0.95	0.83
Confidence Interval	[0.86,0.90]	[0.89,0.94]	[0.57,0.92]	[0.82,0.92]	[0.82,0.94]	[0.59,0.85]	[0.82,0.88]	[0.88,1.03]	[0.75,0.91]
Observations	1453	594	25	174	129	47	468	33	83
Rep. if Orig. Sig. 10%									
Estimate	0.72	0.71	0.42	0.81	0.69	0.84	0.69	0.81	0.91
Confidence Interval	[0.70,0.73]	[0.68,0.74]	[0.33,0.52]	[0.77,0.85]	[0.62,0.77]	[0.79,0.90]	[0.66,0.71]	[0.72,0.90]	[0.89,0.94]
Observations	2826	932	106	373	137	168	1068	74	382
Rep. if Orig. Not Sig. 10%									
Estimate	0.84	0.88	0.93	0.84	0.82	0.53	0.80	0.91	0.74
Confidence Interval	[0.82,0.86]	[0.85,0.90]	[0.80,1.06]	[0.78,0.90]	[0.75,0.89]	[0.37,0.69]	[0.76,0.84]	[0.80,1.02]	[0.64,0.84]
Observations	1179	495	15	149	113	39	345	25	71

Notes: This is the same as Table 12 except applying article weights. Robustness reproducibility and replicability rates for four definitions by type of re-analyses. Columns present robustness reproducibility rates by type of re-analyses, which are not mutually exclusive. Columns 1-8 do not include re-analysis that use new data, while column 9 does. In (2), the re-analysis changed the control variables. In (3), the re-analysis changed the dependent variable. In (4), the re-analysis changed the estimation method. In (5), the re-analysis changed the inference method. In (6), the re-analysis changed the main independent variable. In (7), the re-analysis changed the sample. In (8), the re-analysis changed the weights applied, or applied weights for the first time. In (9), we present robustness replicability rates for re-analyses that introduced new data. 95% confidence intervals presented in square brackets.

Table 14: Caliper Tests, Significance at 5% Level

	Significant at 5% Level			
	(1)	(2)	(3)	(4)
Re-Analysis=1	-0.092***	-0.112**	-0.084*	-0.126*
	(0.030)	(0.044)	(0.050)	(0.070)
Observations	1,973	1,306	786	412
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 15: Caliper Tests, Significance at 10% Level

	Significant at 10% Level			
	(1)	(2)	(3)	(4)
Re-Analysis=1	-0.055	-0.052	-0.069	-0.122
	(0.050)	(0.060)	(0.063)	(0.093)
Observations	814	628	436	201
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 16: Applying Andrews and Kasy (2019)

	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 Andrews and Kasy (2019). The columns μ, τ, and df represent the model's estimated parameters (using an underlying t-distribution and symmetric sign probabilities). The fourth column $[0, 1.645]$ presents the relative publication probability for a t-statistic in the $[0, 1.645]$ interval compared to one in the reference interval of $(2.576, \infty)$.

Table 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 replicators' 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: Many-Analysts' Replication Rate And Replicator Characteristics For Published Results Originally Statistically Significant at the 10% Level

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

Notes: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between replication rate and selected author/replicator 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 19: Many-Analysts' Replication Rate And Replicator Characteristics For Published Results Originally **Not** Statistically Significant at the 5% Level

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

Notes: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between replication rate and selected author/replicator 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 20: Many-Analysts' Replication Rate And Replicator Characteristics For Published Results Originally **Not** Statistically Significant at the 10% Level

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

Notes: Six many-analyst teams independently answered eight pre-registered research questions concerning the possible relationship between replication rate and selected author/replicator 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 21: Many-Analysts' Replication Rate And Replicator Characteristics - Only if Analyst Indicated the Effect Size was Meaningful

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

Dependent Variable: Original Result Statistically Significant at 10% Level					
	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	50.00	50.00	0.00	0.00	100.00
2	55.00	25.00	10.00	10.00	100.00
3	0.00	41.67	25.00	33.33	100.00
4a	0.00	0.00	12.50	87.50	100.00
4b	25.00	0.00	25.00	50.00	100.00
4c	8.33	58.33	0.00	33.33	100.00
5a	6.67	13.33	20.00	60.00	100.00
5b	25.00	25.00	25.00	25.00	100.00
5c	16.67	63.33	0.00	20.00	100.00
6	0.00	20.00	60.00	20.00	100.00
7	20.00	0.00	26.67	53.33	100.00
8	0.00	37.50	50.00	12.50	100.00

Dependent Variable: Original Result *Not* Statistically Significant at 5% Level					
	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	0.00	83.33	16.67	100.00
2	0.00	100.00	0.00	0.00	100.00
3	0.00	41.67	58.33	0.00	100.00
4a	0.00	33.33	33.33	33.33	100.00
4b	0.00	33.33	33.33	33.33	100.00
4c	0.00	33.33	33.33	33.33	100.00
5a	0.00	0.00	72.22	27.78	100.00
5b	11.11	72.22	0.00	16.67	100.00
5c	0.00	37.50	16.67	45.83	100.00
6	12.50	75.00	12.50	0.00	100.00
7	0.00	50.00	33.33	16.67	100.00
8	50.00	33.33	5.56	11.11	100.00

Dependent Variable: Original Result *Not* Statistically Significant at 10% Level					
	Category				
RQ	Neg. & Sig.	Neg. & Not Sig.	Pos. & Not Sig.	Pos. & Sig.	Total
1	0.00	0.00	83.33	16.67	100.00
2	0.00	100.00	0.00	0.00	100.00
3	0.00	75.00	25.00	0.00	100.00
4a	0.00	0.00	83.33	16.67	100.00
4b	0.00	50.00	33.33	16.67	100.00
4c	0.00	12.50	75.00	12.50	100.00
5a	0.00	12.50	50.00	37.50	100.00
5b	0.00	83.33	0.00	16.67	100.00
5c	0.00	37.50	25.00	37.50	100.00
6	0.00	87.50	12.50	0.00	100.00
7	0.00	38.89	44.44	16.67	100.00
8	33.33	50.00	0.00	16.67	100.00

Notes: This table presents the same analysis as in Tables 5, 18, 19, and 20 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 5. The second panel corresponds to Table 18. The third panel corresponds to Table 19. The fourth panel corresponds to Table 20. 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 22: Randomization Tests, Significance at 5% Level

	Original Analysis
Proportion Significant in .05 $\pm.04$	0.779
One-Sided p-value against 0.50	0.000
Number of Tests in .05 $\pm.04$	1973.000
Proportion Significant in .05 $\pm.03$	0.747
One-Sided p-value against 0.50	0.000
Number of Tests in .05 $\pm.03$	1306.000
Proportion Significant in .05 $\pm.02$	0.680
One-Sided p-value against 0.50	0.000
Number of Tests in .05 $\pm.02$	786.000
Proportion Significant in .05 $\pm.01$	0.671
One-Sided p-value against 0.50	0.000
Number of Tests in .05 $\pm.01$	412.000

Notes: Following Brodeur et al. (2020), 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.