Article:
Miske, O., Abatayo, A.L., Daley, M. et al. (2026) Investigating the reproducibility of the social and behavioural sciences. Nature, 652. pp. 126-134. ISSN: 0028-0836

https://doi.org/10.1038/s41586-026-10203-5

Investigating the reproducibility of the social and behavioral sciences

Olivia Miske, Anna Lou Abatayo, Mason Daley, Mirka Dirzo, Nicholas Fox, Noah Haber, Krystal M. Hahn, Melissa Kline Struhl, Brinna Mawhinney, Priya Silverstein, Theresa Stankov, Andrew H. Tyner, Matúš Adamkovič, Shilaan Alzahawi, Saule Anafinova, Eli Awtrey, Erick Axxe, James Bailey, Bert N. Bakker, Akshaya Balaji, Gabriel Banik, František Bartoš, Henk Berkman, Zachariah Berry, Felix S. Bethke, Timothy F. Brady, Nate Breznau, Laura Caquelin, Sara Capitan, Tabaré Capitán, Kent Jason Cheng, William J. Chopik, Gwen-Jiro Clochard, Tom Coupé, Jamie Cummins, Elif Gizem Demirag Burak, Jianhua Duan, Kevin M. Esterling, Thomas R. Evans, Nathan Fiala, James Field, Victor Gay, Jing Geng, Johanna Gereke, Ilka Helene Gleibs, Amélie Gourdon-Kanhukamwe, Dmitry Grigoryev, Nicholas Gunby, Paul H. P. Hanel, Sanghyun Hong, Sean Dae Houlihan, Nick Huntington-Klein, Kamil Izydorczak, Kristin Jankowsky, Kai Jonas, Pavol Kačmár, Hansika Kapoor, Sebastian Karcher, Marta Kołczyńska, David Kretschmer, Ljiljana Lazarevic, Katelin E. Leahy, Jessica C. Lee, Christopher Limnios, An-Chiao Liu, John Wills Lloyd, Ruben Lopez-Nicolas, Nigel Mantou Lou, Richard E. Lucas, Maximilian Maier, Daniel J. Mallinson, Marcel Martončik, Michael C. McCall, Nikita Mehta, Esteban Méndez, Johannes Michalak, Daniel C. Molden, Faisal Mushtaq, Claudia Neuendorf, Austin Lee Nichols, Gustav Nilsonne, Ernest O'Boyle, Jeewon Oh, Thomas Ostermann, Abiola Oyebanjo, Radoslaw Panczak, Yuri G. Pavlov, Zoran Pavlović, Noemi Peter, Kim Peters, Nathaniel D. Porter, Mariah Purol, Arathy Puthillam, Marco Ramljak, Arran T. Reader, W. Robert Reed, Jan Philipp Röer, Ivan Ropovik, Alexander O. Savi, Kathleen Schmidt, Landon Schnabel, Eric L. Sevigny, Samuel Shaki, Shishir Shakya, Andrew Soh, Angela Somo, Fatih Sonmez, Eirik Strømland, Jordan W. Suchow, Anna Szabelska, Anirudh Tagat, Melba Verra Tutor, Karolina Urbanska, Pieter Van Dessel, Elisabeth Julie Vargo, Diem Thi Hong Vo, Victor Volkman, Ke Wang, Aaron L. Wichman, Jamal R. Williams, Fabian Winter, Ferdinand Wintermantel, Nan Zhang, Ignazio Ziano, Cristina Zogmaister, Zorana Zupan, Brian A. Nosek, and Timothy M. Errington

Abstract

Published claims should be *reproducible*, yielding the same result when applying the same analysis to the same data. We assessed reproducibility in a stratified random sample of 600 papers published from 2009 to 2018 in 62 journals spanning the social and behavioral sciences. Authors of 146 (24.3% [95% CI 21.1 - 27.9%]) papers made data available to assess reproducibility, and for some others, we obtained source data to reconstruct the dataset. From 145 papers assessed, 76.3 (52.6% [95% CI 44.7 - 59.9%]) papers were rated as precisely reproducible and 104.5 (72.1% [95% CI 64.5 - 79.0%]) papers as at least approximately reproducible (within 15% of the original effects or within .05 of original p-values) after inverse weighting each of the 553 claims by the number of claims per paper. We observed higher reproducibility for papers from Political Science and Economics than other disciplines, for more recent than older papers, and for papers from journals that required data sharing.

Keywords: credibility, reproducibility, reliability, validity, economics, political science, psychology, marketing, sociology, finance, management, public administration, organizational behavior, education, criminology, health research

Readers of quantitative research are skeptical: Does a research design justify the authors' conclusions? Are the chosen measures valid assessments of the constructs of interest? Would the findings be the same with a different analytic specification? Will the findings generalize to other circumstances? Skepticism identifies weaknesses, roots out errors, and suggests alternative explanations to investigate. However, even skeptical readers will ordinarily assume that the quantitative analyses and outcomes are reported precisely.

Productive scholarly dialogue is difficult if readers of papers are left wondering whether the reported sample size is the same as the sample size in the dataset, whether the reported means reflect the actual means from the data, or whether the reported model is the model that the authors used in their analysis. Ideally, readers should be able to assume that the described analysis, applied to the original data, consistently produces the reported outcomes. This paper investigates how close we are to this ideal.

Investigations in Economics, Finance, Political Science, Cognitive Science, Psychology, Social Sciences, Health, Ecology, and elsewhere suggest that *reproducibility,*[1,2] defined as observing the same results when applying the same analysis to the same data, cannot be taken for granted.[3–18] Irreproducible outcomes can occur because of coding mistakes, transcription errors, or faulty record keeping; many of which are unintentional, all of which are unwelcome.

Investigations of reproducibility hinge on the accessibility of the data, and often the analytic code, because descriptions of analytic methods may be incomplete or difficult to translate back to code.[8,19,20] We assessed availability of author-provided data to enable an independent test of reproducibility. Author-provided data refers to data that authors made available via a website, repository, or direct correspondence by email. This could have been raw or source data, prepared or derived data, or a combination. Data availability rates well below 50%, and sometimes in single digits, have been reported in the fields of Biomedicine, Cancer Biology, Ecology, Business, Economics, and across the social sciences more generally.[7,11,21–25]

As part of the DARPA-funded Systemizing Confidence in Open Research and Evidence (SCORE) program,[26] we conducted a systematic investigation of reproducibility in the social and behavioral sciences.

Table 1. 62 journals included in the sample for selecting papers and claims.

Business	Education	Psychology
Academy of Management Journal	American Educational Research Journal	Child Development
Journal of Business Research	Computers and Education	Clinical Psychological Science
Journal of Management	Contemporary Educational Psychology	Cognition
Leadership Quarterly	Educational Researcher	European Journal of Personality
Management Science	Exceptional Children	Evolution and Human Behavior
Organization Science	Journal of Educational Psychology	Journal of Applied Psychology
Journal of the Academy of Marketing Science	Learning and Instruction	Journal of Consulting and Clinical Psych.
Journal of Consumer Research		Journal of Environmental Psychology
Journal of Marketing		Journal of Experimental Psychology: General
Journal of Marketing Research		Journal of Experimental Social Psychology
Journal of Organizational Behavior		Journal of Personality and Social Psychology
Org. Behavior and Human Decision Processes		Psychological Science
		Health Psychology
		Psychological Medicine
		Social Science and Medicine

Economics	Political Science	Sociology
American Economic Journal: Applied Economics	American Journal of Political Science	American Journal of Sociology
American Economic Review	American Political Science Review	American Sociological Review
Econometrica	British Journal of Political Science	Demography
Experimental Economics	Comparative Political Studies	European Sociological Review
Journal of Finance	Journal of Conflict Resolution	Journal of Marriage and Family
Journal of Financial Economics	Journal of Experimental Political Science	Social Forces
Journal of Labor Economics	World Politics	Criminology
Journal of Political Economy	Journal of Public Admin. Research and Theory	Law and Human Behavior
Quarterly Journal of Economics	Public Administration Review	
Review of Financial Studies		
World Development		

Caption: For primary reporting, Economics and Finance were combined as "Economics," Sociology and Criminology were combined as "Sociology," Management, Marketing, and Organizational Behavior were combined as "Business," Psychology and Health were combined as "Psychology," and Political Science and Public Administration were combined as "Political Science." Outcomes are reported separately by subdiscipline in the supporting information.

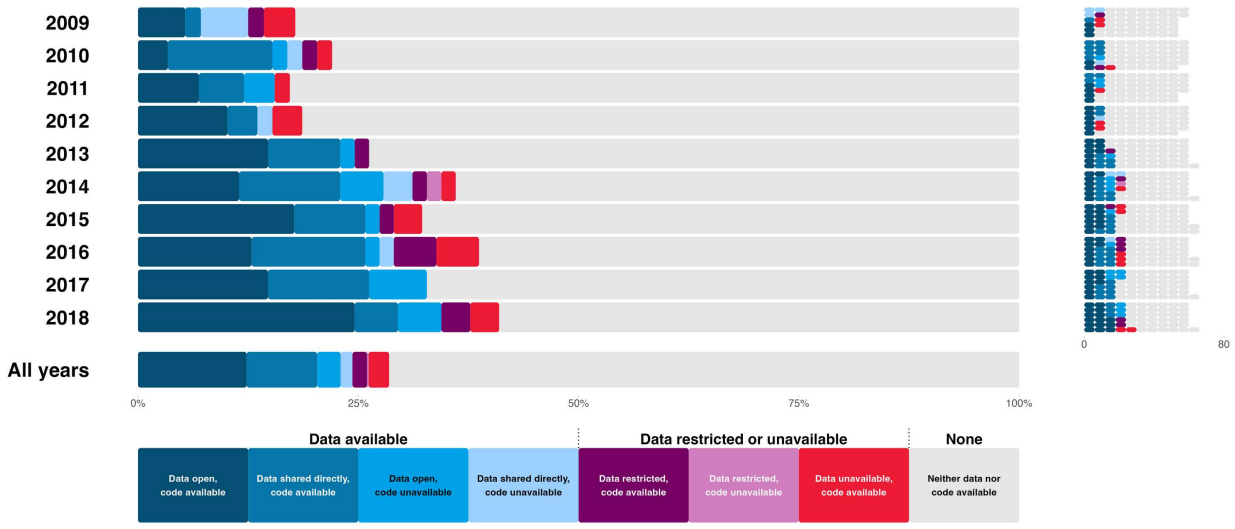
Results

Data and code availability

We assessed data availability of 600 papers from a stratified random sample published from 2009 to 2018 in 62 journals across the social and behavioral sciences (Table 1; see also Table S4). For each paper, we searched for data collected or prepared by the original author(s) for the analyses reported in the paper. This *author-provided* data was considered available if it was publicly available or made accessible by the authors upon request. We examined the paper's main text and supplementary materials for the data and links to repositories, and we recorded whether authors explicitly stated that some or all of the original data sources were restricted or unable to be shared due to ethical or legal reasons. Restricted data was counted as not available (11 cases observed, all from Economics). We did not systematically document whether *source* data were available if the authors did not make it available themselves. We conducted a similar search for analytic code that implemented the analyses reported in the paper.

We obtained both data and code for 122 (20.3% [95% CI 17.3 - 23.7%]), just data for 24 (4.0% [95% CI 2.7 - 5.9%]), just code for 24 (4.0% [95% CI 2.7 - 5.9%]), and neither for 430 (71.7% [95% CI 67.9 - 75.1%]). Thus, data was available for 146 papers (24.3% [95% CI 21.1 - 27.9%]) and unavailable for 454 papers (75.7% [95% CI 72.1 - 78.9%]).

Figure 1. Data and code availability by year of publication. The left panel shows data and code availability as a percentage of papers; the right panel shows raw counts of papers with author-provided data and code available and not available. Note that purple reflects restricted data, which did not count as available data, but might be accessible in principle.



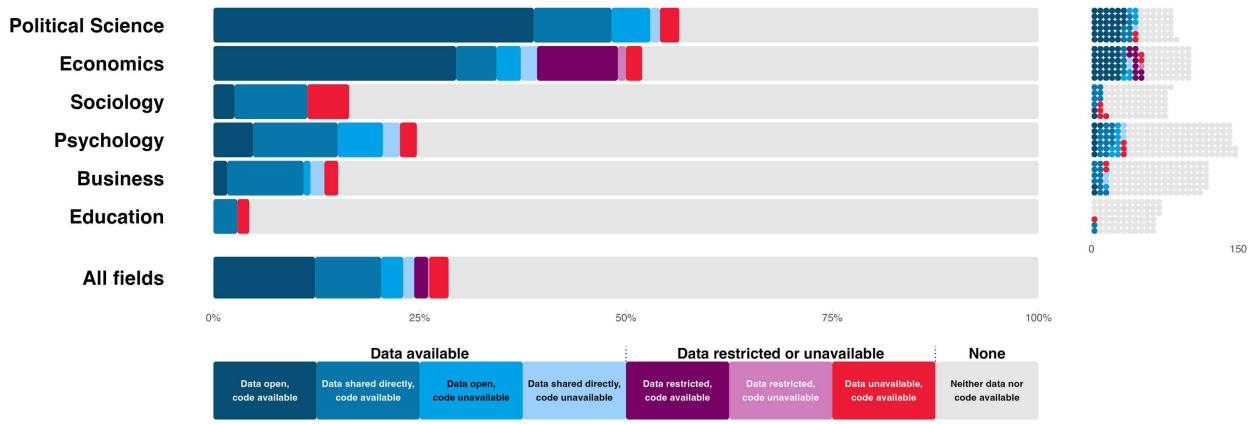
Data and code availability by year of original publication

Prior investigations suggest that data and code availability is higher for more recent papers – perhaps due to poor archival practices that lead to data loss over time or to

standards in data sharing improving over time.[11,27] Our sample spans a 10-year time period of widespread discussion of reproducibility and changes to journals reproducibility policies, offering a window for observing variation due to both of these factors. We replicated the association between year of publication and availability of data and code, with more recent papers having higher rates of sharing data, code, or both (Figure 1). This is also reflected in modest positive correlations between year and percent of papers with data available (ρ = 0.16 [95% CI 0.08 - 0.24]), code available (ρ = 0.17 [95% CI 0.09 - 0.24]), or both available (ρ = 0.16 [95% CI 0.08 - 0.24]).

Conditional on data *or* code being available, we did not observe clear evidence of greater data *and* code availability for more recent papers. Overall, 122 of 170 papers (71.8% [95% CI 64.6 - 78.0%]) with either data or code available had both data and code available and the correlation with year of publication was ρ = 0.07 [95% CI -0.08 - 0.22]. In sum, considering papers for which some sharing occurred, the comprehensiveness of sharing was not significantly higher for more recent papers.

Figure 2. Data and code availability by field. The left panel shows data and code availability as a percentage of papers; the right panel shows raw counts of papers with data and code available and not available. Note that purple reflects restricted data, which did not count as available data, but might be accessible in principle.



Data and code availability by discipline

Papers from the 62 included journals were aggregated into 6 discipline categories for expository purposes: Business (including Marketing, Management, Organizational Behavior), Economics (including Finance), Education, Political Science (including Public Administration), Psychology (including Health), and Sociology (including Criminology). Figure 2 illustrates that Political Science (46 of 85 papers, 54.1% [95% CI 43.6 - 64.3%]) and Economics (40 of 102 papers, 39.2% [95% CI 30.3 - 48.9%]) had higher data availability rates than the other disciplines (combined: 14.5% [95% CI 11.5 - 18.3%]) and Education had the lowest (2.9% [95% CI 0.8 - 10.0%]). The Figure also illustrates that Political Science and Economics have similar data availability rates, if the 11 cases of restricted data in Economics are considered available, in principle. In the supporting information, Figure S6 illustrates even higher data availability in Political Science 69.2% [95% CI 57.2 - 79.1%] and Economics 51.4% [95% CI 40.1 - 62.6%]

after separating them from Public Administration 5.0% [95% CI 0.3 - 23.6%] and Finance 10.0% [95% CI 3.5 - 25.6%], respectively (see also Figures S7, S8).

Assessing reproducibility

Whereas we evaluated data availability only at the paper level, reproducibility was also assessed for individual claims within papers. For most papers, we extracted and evaluated a single key claim, but for a subset of papers, multiple claims per paper were extracted.[28] 59 of 145 (40.7%) papers had >1 claim assessed for reproducibility (mean claims per paper = 3.8, SD = 5.8, range = 1-37). In total, there were 553 claims from 145 papers assessed for reproducibility.

The reproducibility of multiple claims within a paper are potentially statistically dependent because they arise from the same project and authors. We assessed reproducibility at the (1) claim level and (2) paper level by weighting claims-level data to the paper level (e.g., if there were 4 claims in a paper, each was weighted to be equivalent to 0.25 observations) and clustering to account for interclass correlation among claims. As such, reproducibility for papers can be fractional based on the outcomes of claims within the same paper. We report paper-level outcomes in the main text and claim-level outcomes in the supporting information (Figures S1-S3).

Reproducibility assessments in comparison with the sample

Because assessing reproducibility requires access to data, which varies by discipline and time, the subset of papers ultimately tested for reproducibility may not be fully representative of the broader sample. Table 2 shows, by discipline, how representativeness shifted at successive stages of the research process.[29] The initial stages of selecting papers and identifying claims maintained representativeness by discipline (see also Tables S3 and S5).

The primary determinant of inclusion in the reproducibility tests was successful acquisition of author-generated data (n = 146). We also included 37 papers for which authors data were unavailable but source data were obtainable to reconstruct the datasets. Together, these constitute the “papers with source or author data available” in Table 2.

Relative to the broader sample, Political Science and Economics became a larger share of the tested sample; Sociology was largely unchanged; and the remaining fields decreased in share. At the claim level, Economics accounts for a higher proportion because its papers contained more reanalyzed claims per paper. An analogous assessment of representativeness by publication year appears in the Supporting Information (Table S3).

Observed reproducibility

Analysts were matched with data to reanalyze and followed a structured protocol. Reproducibility was investigated with three possible outcomes: precise reproducibility, approximate reproducibility, and not reproduced. Precise reproducibility was achieved if all of the statistical outcomes of the reproduction were the same as originally reported.

This could include, for example, the sample size, focal regression coefficient, test statistic, effect size, and p-value for a single claim. Approximate reproducibility was defined *a priori* as achieved if at least one statistical outcome was not precisely reproduced and all outcomes for a claim were reproduced within ± 15% of what was originally reported and, for p-values, a difference of no more than .05. If any of the statistical outcomes were neither precisely nor approximately reproduced, then the claim was coded as not reproduced.

Table 2. Number of papers at each stage of the selection process and number and percentage of papers and claims reproduced by discipline.

	Business	Economics	Education	Political Science	Psychology	Sociology	Total
				n (%)			
Papers with claims	591 (19.7%)	520 (17.3%)	342 (11.4%)	424 (14.1%)	727 (24.2%)	396 (13.2%)	3000 (100%)
Papers eligible for reproduction	119 (19.8%)	102 (17.0%)	69 (11.5%)	85 (14.2%)	146 (24.3%)	79 (13.2%)	600 (100%)
Papers with multiple claims	38 (19.0%)	33 (16.5%)	23 (11.5%)	32 (16.0%)	49 (24.5%)	25 (12.5%)	200 (100%)
Papers with single claim	81 (20.2%)	69 (17.2%)	46 (11.5%)	53 (13.2%)	97 (24.2%)	54 (13.5%)	400 (100%)
Papers with source or author data available	17 (9.3%)	42 (23.0%)	10 (5.5%)	50 (27.3%)	41 (22.4%)	23 (12.6%)	183 (100%)
Papers with at least one claim reproduction started	15 (9.1%)	38 (23.0%)	11 (6.7%)	46 (27.9%)	31 (18.8%)	24 (14.5%)	165 (100%)
Papers with at least one claim reproduction completed	14 (9.5%)	33 (22.3%)	9 (6.1%)	43 (29.1%)	28 (18.9%)	21 (14.2%)	148 (100%)
Total reproductions of claims	46 (7.4%)	174 (27.9%)	23 (3.7%)	199 (31.9%)	121 (19.4%)	60 (9.6%)	623 (100%)
Reproductions of unique claims	40 (7.2%)	162 (29.1%)	23 (4.1%)	177 (31.8%)	102 (18.3%)	53 (9.5%)	557 (100%)

While 148 papers and 557 claims had at least one outcome reproduction attempt, 4 claims had none of our eligible statistical outcomes and were not counted for the quantitative assessment of reproducibility. Claims were re-weighted to the paper level after excluding these claims. As such, 145 papers consisting of 553 claims were assessed for reproducibility. For 7 papers, an outcome reproduction attempt began, but the analysts' determined that the material they had was not sufficient to assess reproducibility. This could occur if the provided data was incomplete or otherwise compromised for conducting a reproduction, or the provided code was not usable or adaptable. These were counted as reproducibility failures.

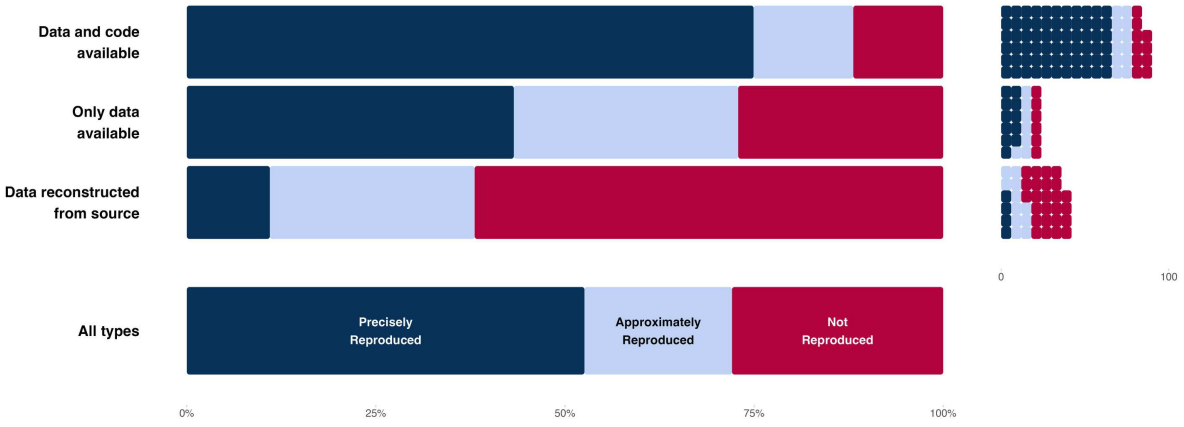
Of the 145 papers that were assessed, we observed approximate or precise reproducibility for 104.5 papers (72.1% [95% CI 64.5 - 79.0%]) and precise reproducibility for 76.3 papers (52.6% [95% CI 44.7 - 59.9%]).

Figure 3 shows reproducibility results separately for different circumstances of conducting the reproduction. When code and data were available, we attempted to execute the original code or adapt it if necessary. We observed approximate or precise reproducibility for 73.3 of the 83.2 papers (88.1% [95% CI 81.4 - 94.6%]) and precise reproducibility for 62.4 of the 83.2 papers (75.0% [95% CI 66.4 - 82.6%]). For 52.1 (62.6% [95% CI 53.6 - 72.0%]) of these papers, we were able to reproduce the findings with minimal effort other than executing the code on the data, a standard known as *push button reproducibility*.[30]

When only data were available, we attempted to reproduce the findings by generating new code following the analyses described in the paper. Of these, we observed approximate or precise reproducibility for 16.1 of the 22.1 papers (72.9% [95% CI 54.8 - 89.4%]) and precise reproducibility for 9.5 of the 22.1 papers (43.2% [95% CI 25.5 - 63.0%]).

When author-provided data were unavailable, but source data were available, we attempted to reproduce the findings by preparing the data and generating new code. Of these, we observed approximate or precise reproducibility for 15.1 of the 39.7 papers (38.1% [95% CI 24.2 - 52.4%]) and precise reproducibility for 4.4 of the 39.7 papers (11.0% [95% CI 3.2 - 20.6%]). In summary, reproducibility rates were comparatively high when data and code were both available, and comparatively low when needing to reconstruct the data and code.[8]

Figure 3. Reproducibility by data and code availability. Reproducibility as a percentage of attempts (left), and reproducibility as counts (right).

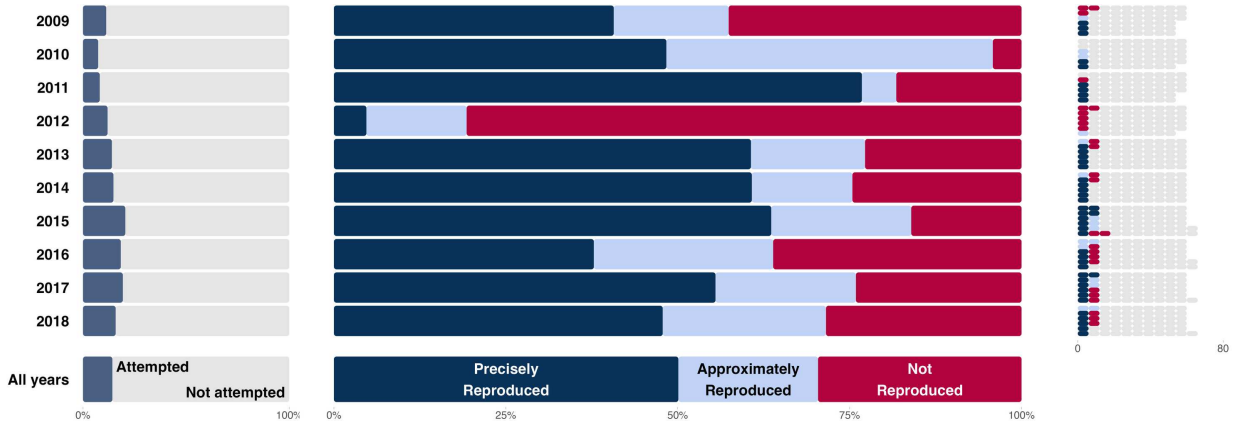


In addition to the empirically defined reproducibility criteria, we asked analysts to provide their subjective assessment of whether they successfully reproduced each claim. This included papers and claims that did not have eligible statistical outcomes for our quantitative evaluation. Excluding missing or undetermined cases, analysts reported successful reproductions of 83 of 134 papers (61.9% [95% CI 53.5 - 69.7%]) and 433 of 537 claims (80.6% [95% CI 77.1 - 83.8%]).

Reproducibility by year of original publication

Figure 4 presents reproducibility by year. The number of reproduction attempts per year is quite small. Considering only papers with an attempt, the prevalence of precise reproducibility (Spearman's ρ = -0.040 [95% CI -0.211 - 0.158]) and the prevalence of approximate and precise reproducibility (Spearman's ρ = 0.007 [95% CI -0.173 - 0.205]) were not significantly associated with time.

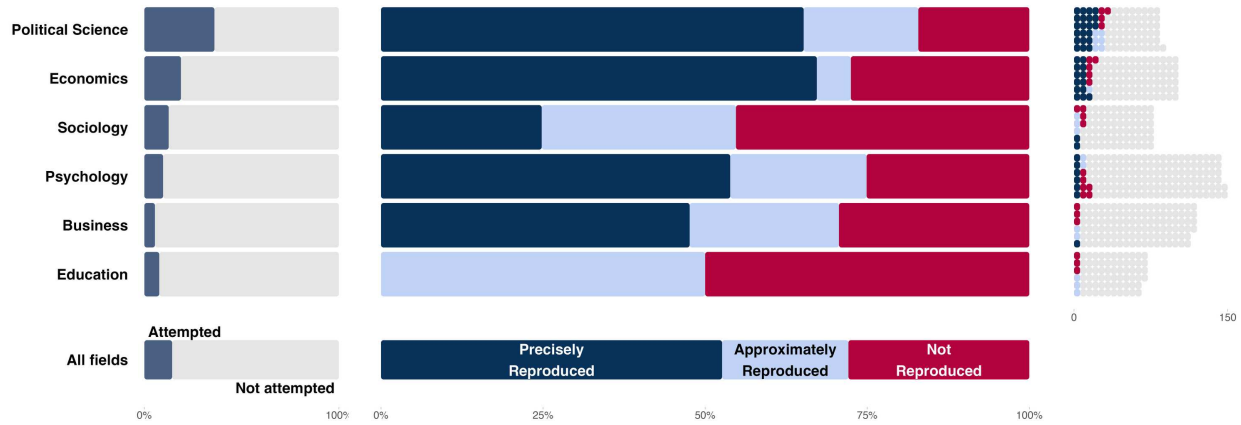
Figure 4. Reproducibility by year of publication. Left: Proportion of reproduction attempts from the sample of papers. Middle: Reproducibility as a percentage of the attempts. Right: Reproducibility as counts compared with the sample of papers. Note that papers with multiple claims could be partly reproducible, but color coding of the dots showing paper counts in the right panel is rounded to the nearest paper.



Reproducibility by discipline

Figure 5 presents reproducibility by discipline (see also Figures S4, S5, S9, S10). Political Science and Economics had much higher rates of reproduction attempts than other fields due to greater data availability. Considering only papers with a reproduction attempt, we observed approximate or precise reproducibility for 34.7 of 41.9 (82.8% [95% CI 71.8 - 92.8%]) Political Science papers and 23.9 of 33.0 (72.5% [95% CI 57.7 - 86.2%]) Economics papers. We observed precise reproducibility for 27.3 of 41.9 (65.1% [95% CI 51.9 - 78.5%]) Political Science papers and 22.2 of 33.0 (67.2% [95% CI 52.6 - 82.1%]) Economics papers. Combining the data across the other four disciplines, we observed approximate or precise reproducibility for 45.8 of 70.0 (65.4% [95% CI 54.9 - 76.6%]) papers and precise reproducibility for 26.7 of 70.0 (38.2% [95% CI 28.6 - 48.6%]) papers.

Figure 5. Reproducibility by discipline. Left: Proportion of reproduction attempts from the sample of papers. Middle: Reproducibility as a percentage of the attempts. Right: Reproducibility as counts compared with the sample of papers. Note that 1) papers with multiple claims could be partly reproducible, but color coding of the dots showing paper counts in the right panel is rounded to the nearest paper, and 2) the weighting scheme in the right column includes claims without reproduction attempts, so the proportions will not exactly match the middle bars.

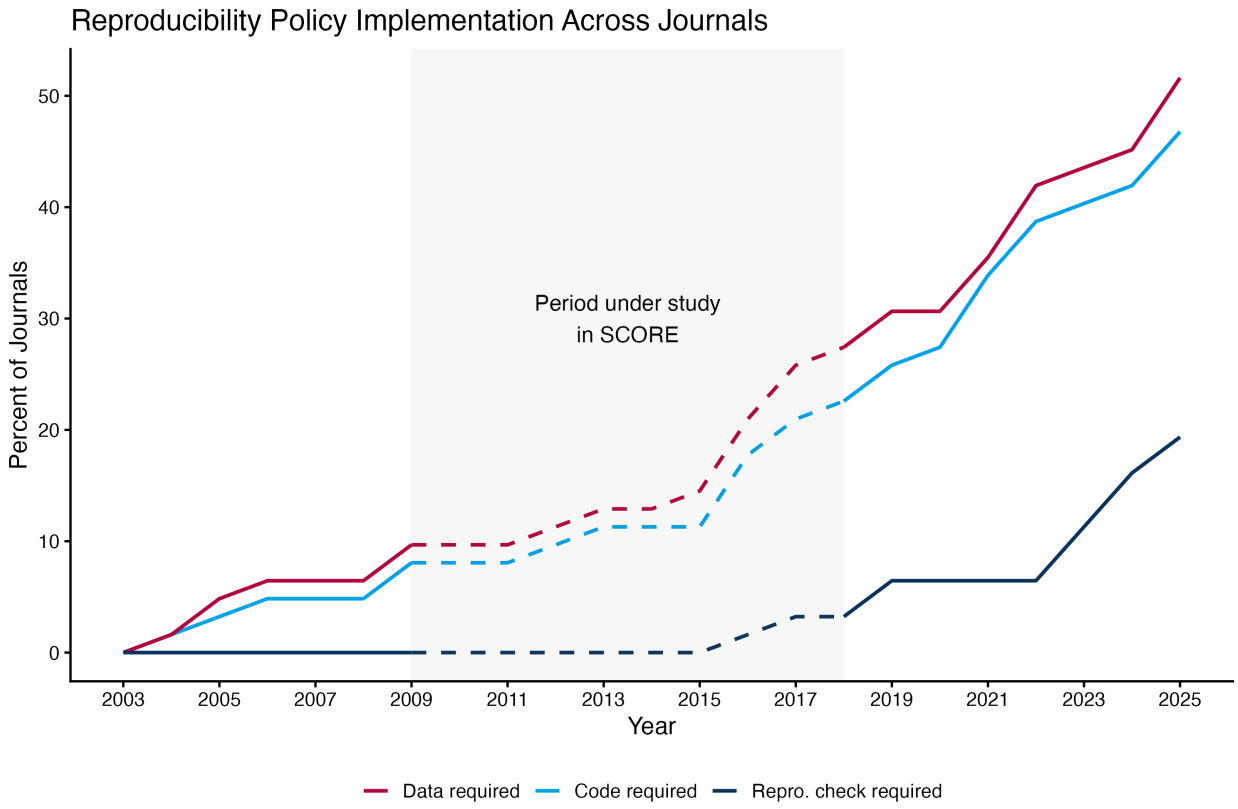


Data and code availability and reproducibility by journal policies

We conducted a follow-up exploratory investigation of the relationship between journal policies and reproducibility (Figure S11; Tables S7-S13). Data availability was observed more consistently for papers published in journals requiring data sharing (87.5%), data and code sharing (66.2%), or data and code sharing and reproducibility checks (100.0%) than for papers published in journals with none of those policies (16.0%; $\chi^2(3) = 99.8$, $p < .001$). Conditional on attempting a reproduction, precise reproducibility was observed more frequently for papers published in journals requiring data sharing (70.5%), data and code sharing (71.9%), or data and code sharing and reproducibility checks (65.0%) than for papers published in journals with none of those policies (40.3%) though the evidence was only suggestive ($\chi^2(6) = 14.6$, $p = 0.023$). Similar analyses at the claim level were consistent and statistically significant ($\chi^2(6) = 46.4$, $p < .001$). Analyses using an ordered logistic regression model at both the paper and claim-levels suggested that the association between journal policy and reproducibility was primarily observed in distinguishing papers and claims that were not reproduced from papers and claims that were approximately and precisely reproduced.

The prevalence of journal policies requiring data sharing, code sharing, and reproducibility checks has increased over time (Figure 6). Reproducibility checks refer to the journal employing an internal process to assess reproducibility prior to publication. In 2018, the last year of our sample of papers, 27.4% of journals had a data sharing requirement, 22.6% had a code sharing requirement, and 3.2% conducted reproducibility checks. 77.8% of journals from Economics and Political Science in our sample had at least one of the policies, and 6.8% of journals from other fields did so. As of mid-2025, rates had increased to 51.6% of journals having a data sharing requirement, 46.8% having a code sharing requirement, and 19.4% conducting reproducibility checks. 94.4% of journals from Economics and Political Science had at least one of the policies, and 43.2% of journals from other fields did so.

Figure 6. Percentage of 62 journals with data sharing, code sharing, and reproducibility check requirements from 2003 to 2025. Three of the 62 journals in the sampling frame had no reproduction attempts (Journal of Finance, Journal of Public Administration Research and Theory, and Law and Human Behaviour) but, even so, their policies are included here.



Discussion

A basic assumption of quantitative research is that repeating the same analysis on the same data will produce the same result as the original report; that is to say, that the reported result is *reproducible*. The most substantial barrier to observing reproducibility in our sample of social and behavioral science papers was the unavailability of author data preventing reproduction attempts. When reproductions could be attempted, availability of data and code was associated with greater but imperfect reproducibility compared with only data availability. Attempting to reproduce findings from source data had a lower success rate. Political Science and Economics papers were more likely to have data available and reproduce successfully than other fields, and exploratory evidence introduces the hypothesis that this could be due partly to the presence of journal policies requiring sharing data, code, or checking reproducibility prior to publication. Looking forward, the proportion of journals requiring these policies has increased since 2018, the year that the most recent papers in our sample were published. These findings provide several insights about reproducibility in the social and behavioral sciences.

Does lack of data availability mean that the outcomes cannot be trusted?

No. It is possible that the results would be perfectly reproducible if the data were available. The primary consequence of this is uncertainty: readers do not know whether the results are reported precisely.

Our criterion was the availability of author-provided data for conducting the analyses reported in the paper. It is likely that more data and code could have been accessed if we had adopted more assertive methods to obtain it. This includes pursuing restricted data that is ostensibly available by meeting the requirements to gain access. For example, data-use agreements for confidential data have been employed at AEA journals for verifying reproducibility.[31]

We could have relaxed the definition of what counted as data availability beyond author-provided datasets, such as including occasions for which source data could be found and obtained, though we observed much less reproducibility in such cases. There are often several data management steps between source data (or raw data) and a dataset that is prepared for inferential analyses (or processed data) that will be reported in the paper. Sometimes these steps are represented in available analytic code, sometimes they are not. There may be a tradeoff between data availability standards and reproducibility. Greater leniency on what counts as sufficient data sharing may be associated with greater failures in reproducing the outcomes. More complete and precise documentation of the data preparation and analysis pipeline improves transparency and possibly reproducibility.[19]

The estimated reproducibility differs dramatically depending on whether papers with no data available are included or ignored, 17.5% versus 72% for approximate or precise reproducibility, for example. Which is a more appropriate reproducibility estimate depends on one's perspective. If the question of interest is whether the outcomes can be verified, then the low estimate reflects the percentage of outcomes we were able to reproduce independently given the amount of effort we invested in gaining access to data and conducting reanalysis. If the question of interest is whether outcomes are reported precisely in papers, then the higher reproducibility estimate might be closer to reality. Presumably, some of the papers with unavailable data would have reproduced successfully if the data could have been obtained.

Does a reproducibility failure mean that the finding is wrong?

No. A reproducibility test can fail because the data used are not identical to the original data, the code or computational environment used to execute the code used is not aligned with the original analysis, the reproduction analyst makes an error or did not spend enough time troubleshooting, or the original description of the data preparation and analysis was incomplete or inaccurate. If those are the sole reasons that the reproduction test failed, then the original outcome may have been reported precisely despite the independent failure to reproduce them. Prior evidence suggests imperfect consistency across reproducibility analysts, and perhaps more so when working from source data rather than author-provided datasets, suggesting that this plays a

meaningful role.[7,32] Even so, a failure in this context creates undesirable uncertainty regarding the credibility of the original outcomes.

Also, we observed a sizable number of approximately correct reproductions. We defined approximately correct statistically, within 15% of the original effects or within .05 of the p-value. Prior investigations in which several analysts investigate the same question with the same data observe substantial variation in data preparation and analysis decisions that may not be reported clearly.[33–35] Without code provided, reproduction attempts may involve making reasonable inferences about how the data were prepared and analyzed based on what is written in the paper. Further, even when code is available, other research suggests that reproducibility is higher when researchers have better coding skills, investigate simpler research questions, and have simpler code.[7] With code provided, reproduction attempts may vary in complexity, increasing the odds of approximate rather than precise success. As such, even though all findings should be precisely reproducible in principle, there are practical challenges in conducting reproductions that may reduce precision without indicting the original finding.[36]

Does reproducibility success mean that the finding is correct?

No. Reproducibility means that the results are reported precisely. Precisely reported results can be wrong because the analysis strategy is invalid, there are coding errors in the data, the research design is confounded, the result is not robust to reasonable alternative analytic decisions, or the researcher selectively reported positive results from many analyses, inflating the likelihood of exaggerated findings.[37–41] Reproducibility is a baseline assessment of credibility for quantitative findings.[42]

What accounts for the differences in data availability and reproducibility across fields?

Papers from Political Science and Economics were more likely to have data available and achieve reproducibility than papers from other fields. A follow-up exploratory investigation suggested an association between reproducibility success and journal policies requiring data sharing, code sharing, and reproducibility checks. Economics and Political Science were more likely to have such policies. Also, transparency policies have become more popular over time across our sample of social and behavioral science journals. If transparency policies have a causal impact on reproducibility, then reproducibility success may be higher in a replication with the same journals using a more recent sample of papers. Future investigations into the causes of reproducibility could also assess the role of social norms, training, tools used in data preparation and analysis, and potential variation across research methodologies. There could be interactions between causes. For example, a transparency policy could ironically cause an increase in data availability and a decrease in reproducibility if implementing the policy is highly burdensome: Available data and code might be unusable by independent researchers thereby harming reproducibility attempts.

Constraints on generalizability

We conducted reproducibility tests on a stratified sample of papers published from 2009 to 2018 from 62 journals in the social and behavioral sciences. Included papers had to have a quantitative outcome associated with a primary claim in the abstract of the paper. Selection of the 62 journals followed a principled approach that was applied consistently across disciplinary boundaries. Nevertheless, the overall and discipline-specific rates may differ with a different sample of journals. Likewise, the exploratory findings that reproducibility rates vary by time and transparency policies imply that the observed outcomes may be different during other time periods. The papers subjected to data and code availability assessment remained representative of the sample, but we did not attempt to access a small number of datasets that were reported as restricted but could, in principle, be obtained. The papers subjected to reproducibility assessment were skewed because a test could be conducted only if data were available. The extent to which this affects the generalizability is unknown. In every field for which a reproducibility study has been conducted, both data availability and reproducibility have fallen short of perfection.[5,9,10,21] Given that not all papers could be assessed for reproducibility, our reproducibility estimate may not generalize to our full sample, or the social and behavioral sciences generally. Even so, this evidence does suggest that reproducibility practices can improve in all disciplines investigated.

Limitations to universal data accessibility and implications for reproducibility

Even if all findings are reported precisely, there are occasions in which reproducibility will not be easily verified because of barriers to data access (see e.g., Weissgerber et al. 2024 on the role of method provenance in responsible FAIR reuse).[43] Principal challenges recognized in open science policies and principles are privacy and proprietary data concerns.[44] In this project, we considered only data made available directly. Some data cannot be publicly shared because they contain personally identifiable or other sensitive information, or are proprietary data belonging to a firm or other private entity. There are a variety of solutions available to advance confidence in reproducibility even under these circumstances, though sometimes with substantial cost.[45,46] For example, some datasets can be anonymized to be publicly shareable for the purposes of demonstrating reproducibility of key findings.[47] Other datasets may not be anonymized, but can be archived and re-analyzed under protected conditions through a variety of data centers with appropriate security and ethical oversight.[48] For proprietary data, authors can spell out the process by which they obtained permission to use it so that an independent researcher could follow the same steps for verification purposes. In some cases, the raw data may not be shareable, but the code and derived data could at least enable verification of the analysis and reporting workflow. Synthetic datasets can be created that reproduce the statistical outcomes without violating confidentiality concerns.[49–51] These solutions demonstrate that the aims of open science and security are not inevitably in opposition. Innovative approaches to data access can maintain proper controls for privacy and security and still provide pathways to accessibility and confidence in reproducibility.

Conclusion

Even the most experienced researchers will make errors in data management, analysis, recordkeeping, and transcription. Implementing measures to verify that research is reproducible is not a statement that researchers are untrustworthy, but a recognition that high standards for quality control are needed because even the most diligent researchers will sometimes be unable to detect and correct mistakes.

A credo of the open scholarship movement is “as open as possible, as closed as necessary.”[52,53] Transparency and sharing enable independent observers to interrogate and verify the basis of research claims. Limitations in transparency and sharing may be inevitable in some cases, and deliberate efforts to maximize verifiability in those circumstances will benefit the trustworthiness of the research. Reproducibility failures add unnecessary uncertainty to the complex enterprise of knowledge production.

Method

We examined whether author data was available so that a reproduction could be attempted, and *reproducibility,* whether the same outcomes as reported originally were observed after conducting the same analysis on the same data.

This reproduction project was part of the DARPA SCORE program to generate and evaluate automated measures of confidence in research claims.[26] Evidence for reproducibility (same analysis, same data) was gathered as a secondary criterion of credibility for the program. Human and machine methods were evaluated on their assessments of replicability (same question, new data).[54,55] Data, materials, code, and other outputs from the program that can be shared without violating privacy or intellectual property restrictions are organized and publicly accessible for evaluation and re-use. This methods section summarizes sampling, conducting the reproducibility assessments, aggregating the data across reproducibility assessments, and evaluation of reproduction outcomes. Further details of these methods are available in the SI (Tables S1-S2).

Sampling frame and selection of claims for reproduction

Research claims were identified with a systematic selection process to reduce selection effects and to enhance generalizability to quantitative social and behavioral research. The project started with a sample of 3000 papers selected by a stratified random sampling of a larger set of papers to ensure representativeness across the 62 journals and publication dates from 2009 to 2018. The time period was defined as the 10 years prior to project onset and the journals were selected via an informal review and nomination process among authors of this paper and other researchers. We selected journals that were well-regarded, published quantitative research, published a sufficient volume of papers during the time period, and collectively represented the diversity of disciplines and quantitative approaches in the social and behavioral sciences.

Within each selected journal, we aimed to extract a single claim from each of five papers per year across the 10-year sampling frame, producing approximately 50 claims per journal depending on the availability of eligible papers. Each paper was reviewed by a trained coder who assessed whether the paper was eligible for SCORE. Eligible papers reported at least one inferential test using human or social data and reported a statistically significant test result that supported a claim made in the paper's abstract.[28] Journals with papers that did not produce an eligible claim in the given year were re-sampled from the same journal and year until 5 claims were extracted, or there were no more eligible papers. This process yielded 3000 claims from 3000 papers across the sample.

From the pool of 3000 papers, 600 were randomly selected as the papers eligible for conducting reproduction attempts with a similar stratified random sampling process to maintain representativeness. Within this pool of 600 papers, 200 were non-randomly sampled for additional coding. In this subset, we extracted all of the main claims regardless of evidence type (i.e., including non-inferential and non-significant evidence). These papers were selected because it appeared likely that we could attempt replications and reproductions of their findings based on feasibility and likely availability of relevant materials, with some adjustments made for representativeness (see Abatayo and colleagues [2026] for details on the sampling and selection process).[29] Other papers and data were gathered during the SCORE program, but they did not include reproduction attempts and are not discussed in this paper.

Data and code availability

We assessed data and code availability of all 600 papers in our stratified random sample. We coded contextual information about the search for data and code sharing such as where it was found, whether it was linked to or referenced from within the article, and whether the paper stated that the data were restricted. Coders first did a brief review of the paper looking for links or references to supplemental materials that included data or code. If either data or code were not located from the paper, coders searched for publicly available materials online, checking online sources such as the website of the publisher or journal where the paper was published, common online repositories, and personal or lab websites of authors. If either data or code were not found, then we emailed the corresponding author and requested the missing content. Retrieved or shared data and code were added to a private OSF project for that paper in preparation for reproducibility assessment.

Reproducibility assessment

192 papers were eligible for reproduction attempts because we had both author data and code, only author data, or source data that could be reconstructed to recreate author data. Of these, 145 papers were assessed for reproducibility. Here, random sampling is lost because selection for reproducibility assessment depends on data availability.

Papers were made available for analyst collaborators to conduct a reproduction attempt. Analysts agreed to attempt reproductions based on factors such as familiarity with the

methods, analytic software, and topical area. Reproduction teams preregistered the inference criteria for judging success. Reproductions conducted during the first half of the program, without the author code, also preregistered their analysis plans. These plans were put through a peer review process managed by an independent editor; otherwise, the preregistration documents were reviewed internally by the project coordinators. Approved preregistrations were registered on the OSF prior to conducting the reproduction attempts. For reproductions conducted during the second half of the program, we eliminated the preregistration and review of analysis plans and added a transparency report of their reproduction process.

Completed reproduction reports went through an internal quality control review. Data, materials, and code were archived on the OSF and made openly available to the maximum extent allowed without violating privacy of participants or intellectual property licenses for any original content.

Data aggregation

Occasionally (n = 62 claims from 49 papers), more than one analyst team conducted a reproduction of the same claim. For reporting purposes, we filtered multiple reproductions through a sequence of decision rules to arrive at a singular outcome for reproducibility. The decision rules were maximally generous to achieving reproducibility. First, we selected whichever reproduction attempt produced outcomes closest to the original, using the reproducibility thresholds of precisely, approximately, and not reproduced (n = 21 claims). Second, if multiple attempts produced equally close results, then we selected the attempt that relied most heavily on the authors' materials (n = 12 claims). Third, if multiple attempts produced equally close results with the same materials, then we selected the attempt that was part of a reproduction of multiple claims in the same paper (n = 23 claims). Finally, if there were multiple reproductions meeting the prior criterion, then we selected randomly among them (n = 6 claims).

Data analysis and inference

Presented statistics are mostly descriptive statistics and precision estimates, using two-tailed 95% confidence intervals. Code used to generate each statistic reported in this paper is provided in the data and code repository. Unless otherwise specified, confidence intervals for proportions claims-level analyses are two-tailed confidence interval estimates using the Wilson interval method.

For paper-level analyses, multiple reproduced claims in a single paper are weighted to the paper level to adjust the target population, and clustered at the paper level to account for interclass correlation. If there are 3 reproduction attempts for a given paper, each is weighted to ⅓. Unless otherwise indicated, confidence intervals are estimated through a simple clustered bootstrap clustered at the paper level, where 95% confidence intervals are estimated for the 2.5 and 97.5 percentile intervals of the bootstrapped sample distribution, using 1,000 bootstrap iterations for every statistic.

For data availability, only one claim per paper was assessed, so no weighting or clustering is necessary. For reproducibility, weighting is pegged to the full population of

claims and papers that were assessed for reproducibility including for subset analyses by discipline and year. Unless otherwise indicated, other analysis population weights are pegged to the full set of claims collected from the 600 papers in our randomized sample.

Inclusion and Ethics

Researchers from more than 24 nations participated in conducting reproductions. Joining the collaboration was an open process, promoted via social media primarily by the Center for Open Science and the corresponding author. A variety of roles were defined to maximize opportunity for researchers with varying skills, areas of interest, and access to resources to participate. Criteria for earning co-authorship was defined in advance so that researchers could make informed decisions about joining the collaboration. All reproduction studies reported in this manuscript involved secondary analysis of data of organizations, firms, or human participants. None involved primary data collection from human participants and all reproduction studies were considered not human subjects research by ethics review boards (BRANY SBER IRB Protocol # 20-030-749, Protocol # 20-019-749, and Protocol # 21-056-749; concurrence from MRDC HRPO and NIWC-PAC HRPO).

Summary Paragraph

Reproducibility refers to observing the same result as an original investigation when the same analysis is applied to the same data. Reproducibility is related to other forms of repeatability including robustness, which involves assessing variation in the results when alternative analyses are applied to the same data, and replicability, which involves testing the same question with independent data. Here we show that published results from the social and behavioral science papers were precisely reproduced 52.6% of the time, and that only 24.3% of published papers could be assessed for reproducibility because of data unavailability given our methodology. We also found that Economics and Political Science had higher reproducibility rates than other disciplines. Reproducibility failures add unnecessary uncertainty to the already complex enterprise of knowledge production. Exploratory evidence suggests that stronger norms and journal policies for sharing data and code could improve reproducibility rates.

Competing Interest Statement

A.H.T., M.D., N.H., K.H., O.M., T.Stankov, B.A.N., and T.M.E. are employees of the non-profit organization Center for Open Science that has a mission to increase openness, integrity, and reproducibility of research.

Data, Materials, and Code Availability Statement

Data, materials, and code associated with this research that can be shared without restriction is publicly available on our OSF repository (https://osf.io/ed8pj/). Also

included is all available documentation for reproduction attempts that were not completed. The repository includes a push button package with all code and data used to both produce all statistics, figures, and tables and code that populates them directly into the manuscript from a template. This includes most of the data and code from the individual reproduction attempts, save for any data that is proprietary or protected that will not be made available, or for which analyst teams were uncertain or unable to confirm that they were allowed to share secondary data. It is possible that some data, materials, or code that could be shared openly is not available at the time of publication. Readers are encouraged to contact the corresponding author or the authors of the relevant subproject (Table S2) to see if more research content can be shared.

Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreements No: N660011924015 (PI: Brian A. Nosek) and HR00112020015 (PI: Timothy M. Errington). The views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We thank Beatrix Arendt, Alexandria Denis, Samuel Field, Zachary Loomas, Bri Luis, Lesley Markham, E. Simon Parsons, Courtney Soderberg, and Adam Russell for their contributions to this project.

Additional Information

Contact Brian Nosek, nosek@cos.io, for correspondence concerning this paper.

Author ORCiDs and Institutions

Given name	Family Name	ORCID	Institution 1	Institution 2
Olivia	Miske	0000-0003-4787-3995	Center for Open Science	
Anna Lou	Abatayo	0000-0002-2686-5075	Wageningen University and Research	
Mason	Daley	0000-0002-3460-3673	Center for Open Science	
Mirka	Dirzo		Center for Open Science	
Nicholas	Fox	0000-0002-3772-8666	Center for Open Science	
Noah	Haber	0000-0002-5672-1769	Center for Open Science	
Krystal M.	Hahn	0009-0006-2551-4528	Center for Open Science	
Melissa	Kline Struhl	0000-0003-2217-9331	Massachusetts Institute of Technology	
Brinna	Mawhinney	0000-0002-4926-3026	Center for Open Science	
Priya	Silverstein	0000-0003-0095-339X	University of Coimbra, Portugal	Institute for Globally Distributed Open Research and Education
Theresa	Stankov		Center for Open Science	
Andrew H.	Tyner	0000-0001-9180-4490	Center for Open Science	
Matúš	Adamkovič	0000-0002-9648-9108	Slovak Academy of Sciences	University of Jyväskylä and Charles University
Shilaan	Alzahawi	0000-0002-6892-4643	Stanford University	
Saule	Anafinova	0000-0002-4466-3426	Budapest University of Technology and Economics (BME)	
Eli	Awtrey	0000-0002-6712-0256	University of Cincinnati	
Erick	Axxe	0000-0002-0426-5722	Hendrix College	
James	Bailey	0000-0002-6132-6026	Providence College	
Bert N.	Bakker	0000-0002-6491-5045	University of Amsterdam	
Akshaya	Balaji		Monk Prayogshala	
Gabriel	Banik	0000-0002-6601-3619	Pavol Jozef Safarik University, Slovakia	
František	Bartoš	0000-0002-0018-5573	University of Amsterdam	
Henk	Berkman		University of Auckland	
Zachariah	Berry	0000-0002-0827-6437	University of Southern California	
Felix S.	Bethke	0000-0002-4259-6071	Peace Research Institute Frankfurt (PRIF)	
Timothy F.	Brady	0000-0001-5924-5211	University of California, San Diego	
Nate	Breznau	0000-0003-4983-3137	German Institute for Adult Education - Leibniz Institute for Lifelong Learning	
Sara	Capitan	0000-0001-6519-6073	Swedish University of Agricultural Sciences	
Tabaré	Capitán	0000-0002-5055-3995	Swedish University of Agricultural Sciences	
Laura	Caquelin	0000-0003-4557-3315	Karolinska Institutet	
Kent Jason	Cheng	0000-0002-8931-4086	The Pennsylvania State University	
William J.	Chopik	0000-0003-1748-8738	Michigan State University	
Gwen-Jiro	Clochard	0009-0004-5513-4193	Osaka University	Joint Initiative for Latin American Experimental Economics
Tom	Coupé	0000-0002-9520-5556	University of Canterbury	UCMeta
Jamie	Cummins	0000-0002-4681-0725	University of Bern	University of Bern
Elif Gizem	Demirag Burak	0000-0001-9974-8956	University of Oklahoma	
Jianhua	Duan	0000-0002-4750-0243	Stats NZ	University of Canterbury
Kevin M.	Esterling	0000-0002-5529-6422	University of California Riverside	
Thomas R.	Evans	0000-0002-6670-0718	University of Greenwich	
Nathan	Fiala		University of Connecticut	
James	Field	0000-0001-8487-6648	West Virginia University	
Victor	Gay	0000-0001-9912-3841	Toulouse School of Economics	
Jing	Geng	0000-0002-7059-7725	Virginia Tech	
Johanna	Gereke	0000-0002-1058-9651	University of Mannheim	

Ilka Helene	Gleibs	0000-0002-9913-250X	London School of Economics	
Amélie	Gourdon-Kanhukamwe	0000-0002-3060-1320	Kingston University London	King's College London
Dmitry	Grigoryev	0000-0003-4511-7942	HSE University	
Nicholas	Gunby	0009-0001-6003-9068	Contact Energy	UCMeta
Paul H. P.	Hanel	0000-0002-3225-1395	University of Essex	
Sanghyun	Hong	0000-0003-0135-2617	University of Canterbury	
Sean Dae	Houlihan	0000-0001-5003-9278	Dartmouth College	
Nick	Huntington-Klein	0000-0002-7352-3991	Seattle University	
Kamil	Izydorczak	0000-0002-9870-3825	SWPS University	
Kristin	Jankowsky	0000-0002-4847-0760	University of Kassel	
Kai	Jonas	0000-0001-6607-1993	Maastricht University	
Pavol	Kačmár	0000-0003-0076-1945	Faculty of Arts, Pavol Jozef Šafárik University in Košice	
Hansika	Kapoor	0000-0002-0805-7752	Monk Prayogshala	University of Connecticut
Sebastian	Karcher	0000-0001-8249-7388	Syracuse University	
Marta	Kołczyńska	0000-0003-4981-0437	Institute of Political Studies of the Polish Academy of Sciences	
David	Kretschmer	0000-0002-8702-3007	University of Oxford	
Ljiljana	Lazarevic	0000-0003-1629-3699	Faculty of Philosophy, University of Belgrade, Serbia	
Katelin E.	Leahy	0000-0002-3638-3694	Michigan State University	
Jessica C.	Lee	0000-0003-4253-2008	University of Sydney	University of New South Wales
Christopher	Limnios	0000-0001-5387-1334	Providence College	
An-Chiao	Liu	0000-0003-4064-0515	Utrecht University	
John Wills	Lloyd	0000-0002-2597-6216	University of Virginia	
Ruben	Lopez-Nicolas	0000-0002-6963-7443	University of Murcia	
Nigel Mantou	Lou	0000-0003-1363-833X	University of Victoria	
Richard E.	Lucas	0000-0002-7995-3319	Michigan State University	
Maximilian	Maier	0000-0002-9873-6096	University College London	
Daniel J.	Mallinson	0000-0002-8094-6685	Penn State Harrisburg	
Marcel	Martončik	0000-0003-4869-6900	Institute of Social Sciences CSPS SAS	University of Jyväskylä
Michael C.	McCall	0000-0002-4668-4212	Syracuse University	
Nikita	Mehta	0000-0001-6208-747X	Monk Prayogshala	
Esteban	Méndez	0000-0002-7248-6092	Central Bank of Costa Rica	
Johannes	Michalak	0000-0003-4701-5464	Witten/Herdecke University	
Daniel C.	Molden	0000-0002-2182-5621	Northwestern University	
Faisal	Mushtaq	0000-0001-7881-1127	University of Leeds	
Claudia	Neuendorf	0000-0002-3024-0000	University of Potsdam	
Austin Lee	Nichols	0000-0003-4580-3301	Central European University	
Gustav	Nilsonne	0000-0001-5273-0150	Karolinska Institutet	Stockholm University
Ernest	O'Boyle	0000-0002-9365-1069	Indiana University	
Jeewon	Oh	0000-0001-8103-906X	Syracuse University	
Thomas	Ostermann	0000-0003-2695-0701	Witten/Herdecke University	
Abiola	Oyebanjo		Policy Innovation Center	
Radoslaw	Panczak	0000-0001-5141-683X	University of Bern	
Yuri G.	Pavlov	0000-0002-3896-5145	University of Tuebingen	
Zoran	Pavlović	0000-0002-9231-5100	Faculty of Philosophy, University of Belgrade	
Noemi	Peter	0000-0002-2743-4883	University of Groningen	
Kim	Peters		University of Exeter	
Nathaniel D.	Porter	0000-0002-0479-6777	Virginia Tech	
Mariah	Purol	0000-0003-2921-3600	Union College	
Arathy	Puthillam	0000-0003-2426-8362	Monk Prayogshala	UC San Diego
Marco	Ramljak	0009-0008-1502-6453	Utrecht University	Zeppelin Universität

Arran T.	Reader	0000-0002-0273-6367	University of Stirling	
W. Robert	Reed	0000-0002-6459-8174	University of Canterbury	UCMeta
Jan Philipp	Röer	0000-0001-7774-3433	Witten/Herdecke University	
Ivan	Ropovik	0000-0001-5222-1233	Charles University	Czech Academy of Sciences
Alexander O.	Savi	0000-0002-9271-7476	University of Amsterdam	
Kathleen	Schmidt	0000-0002-9946-5953	Southern Illinois University	Ashland University
Landon	Schnabel	0000-0002-2674-3019	Cornell University	
Eric L.	Sevigny	0000-0002-1596-0042	Georgia State University	
Samuel	Shaki	0000-0002-2340-5401	Ariel University	
Shishir	Shakya	0000-0002-6272-6654	Appalachian State University	
Andrew	Soh		Ateneo de Manila University	
Angela	Somo	0000-0002-9069-9462	San Diego State University	
Fatih	Sonmez	0000-0002-4054-0269	Muş Alparslan University	
Eirik	Strømland		Western Norway University of Applied Sciences	
Jordan W.	Suchow	0000-0001-9848-4872	Stevens Institute of Technology	
Anna	Szabelska	0000-0001-5362-3787	Psychological Science Accelerator	
Anirudh	Tagat	0000-0002-7707-453X	Monk Prayogshala	
Melba Verra	Tutor	0000-0001-7951-3690	Independent researcher	
Karolina	Urbanska	0000-0001-5063-4747	Independent Researcher	
Pieter	Van Dessel	0000-0002-3401-780X	Ghent University	
Elisabeth Julie	Vargo	0000-0002-5123-1170	Institute for Globally Distributed Open Research and Education (IGDORE)	
Diem Thi Hong	Vo	0000-0002-5289-2325	RMIT University Vietnam	UCMeta
Victor	Volkman	0000-0003-2781-535X	University of Connecticut	
Ke	Wang	0000-0002-5776-0815	University of Virginia	
Aaron L.	Wichman	0000-0002-2641-440X	Western Kentucky University	
Jamal R.	Williams	0000-0002-3034-511X	University of California, San Diego	
Fabian	Winter	0000-0002-4838-4504	University of Zurich	
Ferdinand	Wintermantel	0009-0002-6816-0185	Humboldt-Universität zu Berlin	Zeppelin Universität
Nan	Zhang	0009-0001-6883-1359	University of Mannheim	
Ignazio	Ziano	0000-0002-4957-3614	University of Geneva	
Cristina	Zogmaister	0000-0002-1540-7503	Università di Milano-Bicocca	
Zorana	Zupan	0000-0002-0763-8192	University of Belgrade	
Brian A.	Nosek	0000-0001-6797-5476	Center for Open Science	University of Virginia
Timothy M.	Errington	0000-0002-4959-5143	Center for Open Science	

Author Contributions: CReDiT Taxonomy

Full name	Conceptualization	Data curation	Formal Analysis	Funding acquisition	Investigation	Methodology	Project admin	Resources	Software	Supervision	Validation	Visualization	Original draft	Review & editing	Contribution Explanation
Olivia Miske	0	0	0	0	1	1	1	0	0	1	1	0	1	1	I was a Project Coordinator (and later, Assistant Research Scientist) for SCORE at COS. I was involved in a range of activities which included: Contributing to the conceptualization/design of the data availability assessment effort (e.g., determining coding scheme/workflow); providing oversight over the PR coding effort; communicating with reproduction analysts and assisting with coordination/tracking of projects; data entry/validation of reproduction outcomes for a subset of projects; assisting with the OSF audit process; contributing to the methods write-up.
Anna Lou Abatayo	0	1	0	0	1	1	0	0	1	0	1	0	0	0	Extracted claims, mostly from economics journals but helped out on non-economics journals / covid-preprints when needed. Reviewed others' extracted claims. Part of the team (along with Andrew Tyner) that created the process to replicate extracted claims. Collected data for replication (also collected data for reproduction when replication and reproduction data overlapped). Analyzed data for replication. Managed excel file where external individuals picked a claim / journal article to replicate. Read through and checked pre-analysis plans for replication by others before they were submitted. Read through and checked the analysis (for replication). Recruited external individuals that helped with replication (and they might have helped for reproduction too). Helped review related literature on journal policies regarding data sharing, code sharing, and conducting reproducibility checks.
Mason Daley	0	0	0	0	1	0	0	0	0	0	0	0	0	1	PR coding
Mirka Dirzo	0	0	0	0	0	0	0	0	0	0	0	0	0	1	As a Project Coordinator with Center for Open Science, I have helped with coding, created charts/graphs for presentations about SCORE, and will be possibly contributing to writing and review of publications.
Nicholas Fox	0	0	0	0	1	1	0	0	0	0	0	0	0	0	Research scientist in TA1, coding scientific papers to generate datasets for TA2 and TA3 usage, as well as managing research laboratories conducting replication attempts for validation
Noah Haber	0	0	0	0	0	0	0	0	0	0	1	1	0	0	I contributed to data curation, analysis code, figure design and generation, and producing the reproducible manuscript workflow and infrastructure.
Krystal M. Hahn	0	0	0	0	1	0	0	0	0	1	0	0	0	0	I've contributed to data availability efforts (e.g., IPR/DPR scoring), non-HSR coordination and sourcing, and the OSF audit.
Melissa Kline Struhl	1	0	1	0	1	1	0	0	1	1	0	0	0	0	I was one of four research scientists on the COS SCORE team, involved in designing and executing processes for SCORE claim selection, replication/reproduction design, data management & analysis of primary outcomes of finished replications/reproductions
Brinna Mawhinney	0	0	0	0	1	1	0	0	0	0	0	1	1	1	As a Project Coordinator with Center for Open Science, I have helped with the development of coding schemes relevant to SCORE's methodology, conducted data collection, created charts/graphs for presentations about SCORE, and will be contributing to writing and review of publications.
Priya Silverstein	0	0	0	0	0	1	0	0	0	0	1	0	0	1	I was involved in lots of different bits and bobs! CE (single-trace and bushel), had input on some P2 process stuff (especially bushel CE), variable coding (original, replication, reproduction), and some validation stuff (OSF audits, checking CE and variable coding, etc.)
Theresa Stankov	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Set up raw data to processed data pipeline for replication outcomes
Andrew H. Tyner	1	1	0	0	1	1	1	0	0	1	1	1	1	1	I was a Research Scientist and then Principal Research Scientist for the duration of SCORE. In these roles I was involved in most aspects of implementing the project, including claim extraction, facilitating replications and reproductions, quality assurance, data curation, analysis, visualizations, and oversight responsibilities.
Matúš Adamkovič	0	0	1	0	1	1	0	0	0	0	0	0	0	0	Conducted two ADRs (Swanson_JournEduPsych_2016_e2 and Seong_JournManage_2015_3B4j) with my colleague Ivan Ropovik
Shilaan Alzahawi	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted four reproduction studies and one secondary data replication (i.e. data analytic replication). 1. I completed a Source Data Reproduction of BERSANI_Criminology_2013_zmYY_2w9mo 2. I completed a Source Data Reproduction of Denson_AmEduResJourn_2009_zb3Y 3. I completed a Source Data Reproduction of Li_LeadQuart_2011_GQvr 4. I completed a Source Data Reproduction of Vadillo_JournExPsychGen_2016_BrGp 5. I completed a Data-Analytic Replication of Vadillo_JournExPsychGen_2016_BrGp
Saule Anafinova	0	1	1	0	0	0	0	0	0	0	1	0	1	1	I conducted one reproduction study as part of the SCORE project. I also acted as a peer-reviewer, and validated the correctness of entered data. I participated in the revision of the paper.
Eli Awtrey	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Reviewer for reproduction of a claim from Mosimann & Pontusson (2017)
Erick Axxe	0	0	1	0	1	1	0	0	1	0	1	0	0	0	I conducted one replication study. (O'Brien_AmSocioRev_2015_7X54 - Ramljak/Axxe - Data Analytic Replication - 93k7)
James Bailey	0	0	0	0	1	1	0	0	1	0	1	0	0	0	Prepared data for a replication study that included a reproduction. Reviewed a replication study that included a reproduction.
Bert N. Bakker	0	0	0	0	0	0	0	0	0	1	1	0	0	1	I have been a reviewer or editor on the following papers: McLaren (2012); Mosimann & Pontussen (2017); Zakharova et al. 2014; Ahlerup et al. (2016); Chung & Chuwonganant (2014); Tertytchnaya et al. (2018); Fitzgerald et al. (2018); Nagengast et al. (2014); Teney (2016); Bezu et al. (2012); Goeree and Yariv (2011);

Akshaya Balaji	0	0	1	0	0	0	0	0	1	0	0	0	0	0	Angrist, J., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. American Economic Review, 99(4), 1384-1414. [OSF Project], joint with Anirudh Tagat (Monk Prayogshala) LaFave, D., & Thomas, D. (2016). Farms, families, and markets: New evidence on completeness of markets in agricultural settings. Econometrica, 84(5), 1917-1960. [OSF Project], joint with Anirudh Tagat (Monk Prayogshala) McDevitt, R. C. (2014). "A" business by any other name: firm name choice as a signal of firm quality. Journal of Political Economy, 122(4), 909-944. [OSF Project], joint with Anirudh Tagat (Monk Prayogshala)
Gabriel Banik	0	0	0	0	1	0	0	0	0	0	0	0	0	1	I worked on reproductions of Bersani_Criminology_2013_zmYY, and Liao_JournOrgBehavior_2016_PkXJ with my colleagues Matus Adamkovic and Ivan Ropovik.
František Bartoš	0	0	1	0	1	1	0	0	1	0	0	0	0	0	My colleague and I conducted reproduction of one study (Hansen_JournExpSocPsych_2014).
Henk Berkman	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I verified a paper by Stambaugh et al. in the Journal of Financial Economics
Zachariah Berry	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I believe I reproduced several findings in one article
Felix S. Bethke	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I conducted a SCORE reproduction project in 2022. (i.e. Balcells_JournConflictRes_2014_0P4r_28884)
Timothy F. Brady	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted a reproduction of the research claim from 'Asymmetrical Body Perception: A Possible Role for Neural Body Representations', by Linkenauger et al. (2009).
Nate Breznau	0	0	0	0	1	1	0	0	1	0	1	1	0	1	Computational reproduction. Writing and editing of 1st and 2nd drafts. Data collection, lead analysis and visualization for the integration of journal policies for the R&R.
Sara Capitan	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I worked on reproductions.
Tabaré Capitán	0	0	1	0	1	1	0	0	0	1	1	0	0	0	I was an editor, reviewer, and conducted replication studies.
Laura Caquelin	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Provided review of analysis code for ensuring clarity, accuracy, and reproducibility.
Kent Jason Cheng	0	0	1	0	1	0	0	0	0	0	0	0	0	0	admittedly I do not have a clear recollection of the difference between the replication and reproduction study, but I was involved in the collection of data for the replication of 4 studies (Montez et al, Fielding-Miller et al, Carillo Vega, and Fitzgerald et al).
William J. Chopik	1	0	1	0	1	1	0	0	0	1	1	0	1	1	For our replication project (Nelson), I (with help from students) programmed the survey and edited/spliced/hosted the videos for the study. I coordinated data collection and training of RAs, and supervised data analysis/reporting (which was mostly done by the students). I reviewed several replication/pre-registrations. I encouraged those students (Mariah Purol, Jeewon Oh, Katelin Leahy) to complete this form.
Gwen-Jiro Clochard	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I conducted the replication analysis for one paper (Platt Boustan_AmEcoJourn_2012_PVQK_69y19_PBR). I also aimed at conducting a second analysis (Carrell_AmEcoJourn_2010_LmA2_2kgk8), but could not obtain data from the Alachua County Public Schools administration.
Tom Coupé	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Jamie Cummins	0	0	0	0	1	0	0	0	0	0	1	0	0	1	I reviewed protocols for source data reproductions, and led an eventually not-completed push-button reproduction (not completed due to lack of willingness from original authors to share relevant materials).
Elif Gizem Demirag Burak	1	0	0	0	1	0	0	0	0	1	1	0	0	1	I had 1 study. I ran the PBR on the claims for which it's possible and conducted an ADR for the remaining claims. The study is titled as Gender Differences in Political Knowledge: Bringing Situation Back In; Ihme_JournExpPoliSci_2018_xYbO
Jianhua Duan	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I co-led the team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Kevin M. Esterling	0	1	0	0	0	1	0	0	0	1	1	0	0	0	I served as editor for several papers, but I did not keep track of how many, and I was unable to recover the number from a search of my emails. I also served on teams to collect new data on journal policies at the revision stage, and assisted in the statistical analysis of the journal policies impacts.
Thomas R. Evans	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted a computational reproduction analysis for 4 studies: King_JournOrgBehaviour_2017_Q1dl Bertin_covid_zk94 Hou_ChildDev_2017_YOXI Plohl_covid_W3vr
Nathan Fiala	0	0	1	0	1	1	0	0	0	1	0	0	1	1	
James Field	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I completed three reproduction studies.

Victor Gay	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted two computational reproductions (dataset construction + reproduction analysis): Park_Demography_2010_ZdgL - Gay - Computational Reproduction - 2637 (https://osf.io/uyzc4/) and Mosimann_WorldPolitics_2017_z4dO - Gay - Computational Reproduction - 6m17 (https://osf.io/kzpf8/)
Jing Geng	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Performed data analytic replications of research claims from Anderson (2011) and Desmond (2015) in American Economic Journal (Systematizing Confidence in Open Research and Evidence program). Duties included IRB application, data cleaning and processing with R, and coordinating with Virginia Tech and external faculty.
Johanna Gereke	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I conducted 1 reproduction study.
Ilka Helene Gleibs	0	0	1	0	1	1	0	0	1	1	1	0	0	0	I conducted one replication study, and also reviewed a pre-registration. I have supervised my former PhD student Nihan Albayrak-Aydemir with whom I collected the data and published some of the results (https://doi.org/10.25384/SAGE.6263366.v2). I contacted editors and coded data about journals' policies regarding data requirements.
Amélie Gourdon-Kanhukamwe	0	0	0	0	0	0	0	0	0	1	1	0	0	0	According to my working hours records, I served as reviewer for 11 submissions and as editor for 22 submissions, although screening back payment agreements and Gdocs I have once worked on, I can confirm only 27 names (7 reviews and 20 editing jobs). Of these, one was a reproduction study (Rinaldi_Cognition_2016_Kj9d_5196): the full list of identified studies is at https://ameliegourdonkanhukamwe.notion.site/2fd4b161b8994bb39d75cb097ece5f22?v=1faf926789d74544be1bd377c2330d0e&pvs=4
Dmitry Grigoryev	0	1	1	0	0	0	0	0	1	0	1	0	0	1	I conducted two reproduction studies: 1) Push Button Reproduction Attempt to Evaluate a Claim from Hertel_ClinPsychSci_2018_YabW; 2) Source Data Reproduction Attempt to Evaluate a Claim from Stice_JournConsClinPsy_2009_q4X2. For the journal policy subproject, I contributed to: Data curation: collected and structured journal responses, verified completeness; Project administration: handled journal outreach and follow-up communication, monitored incoming responses and maintained records; Review & editing: checked the integrity of entered policy information, aligned language across entries, ensured consistency with project goals.
Nicholas Gunby	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I wrote code to reproduce the Baxter et. all Social Forces study - collaborated closely with Bob Reed and Jane Duan
Paul H. P. Hanel	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I have identified the relevant statistical analyses, conducted the analyses, and wrote up a report.
Sanghyun Hong	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Sean Dae Houlihan	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted one reproduction study on: Jared B Fitzgerald, Juliet B Schor, and Andrew K Jorgenson. (2018). Working Hours and Carbon Dioxide Emissions in the United States, 2007–2013. Social Forces. (Paper ID: Fitzgerald_SocialForces_2018_4q0L)
Nick Huntington-Klein	0	0	0	0	1	0	0	0	0	0	1	0	0	0	I conducted two reproduction studies.
Kamil Izydorczak	0	0	1	0	0	0	0	0	0	0	0	0	0	0	I performed push-button replications for following studies: Alves_PsychologSci_2018_AvOr - RRTeam_unassigned - Computational Reproduction - mzk9, Adida_CompPolitStu_2016_G0Kb - RRTeam_unassigned - Computational Reproduction - g2z. I also participated in Multi100 performing push-button replication and independent analysis for one study: Brough_JournConsRes_2016_9ey
Kristin Jankowsky	0	0	1	0	1	1	0	0	1	0	0	0	0	1	
Kai Jonas	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I served as both and editor and reviewer for many reproductions studies
Pavol Kačmár	0	0	1	0	0	0	0	0	1	0	0	0	0	1	I have conducted SDR (SCORE study id: Robinson_6owm3).
Hansika Kapoor	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Reviewer for 31: Stanley_covid_b3G4_98g Pennycook_covid_7NEL_9k1y Malik_covid_Y3jx_348 Pfattheicher_covid_yZD4_y006 Du_covid_2NAG_41z0 Bohnke_EurSocioRev_2017_xGGO_y11 O'Brien_AmSocioRev_2015_7X54_93k7 Denson_AmEduResJourn_2009_zb3Y_41k2 Du_covid_2NAG_41z0 Niehaus_AmEduResJourn_2014_BlRQ_546 Montez_Demography_2014_3aPw_05g8 Kim_CompEdu_2014_YWep_75g6 Past√∂tter_Cognition_2013_EQxa_3z3k Seaton_AmEduResJourn_2010_Blxd_6778 Berg_covid_qKPb_k127 Baxter_SocialForces_2015_z0v1_y410 Kausel_OrgBehavior_2015_5XEE Petit_JournBusRes_2017_9R9X Griffiths_JournExPsychGen_2011_J7ek Bhattacharjee_JournPerSocPsy_2017_Br0x King_JournOrgBehavior_2017_Q1dl Raley_JournMarFam_2012_D2LY Weidmann_2g7ky Montez_2y4gm Karraker_2g79y Anderson_329k Li_6m34m van Gastel_21487 Liang_23g12 BATESON_2k5g2 Andrews_95my
Sebastian Karcher	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I conducted two reproduction studies
Marta Kołczyńska	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I was analyst in 4 reproduction studies. In one of these reproduction studies I collaborated with Karolina Urbanska (kurbanska015@gmail.com).
David Kretschmer	0	0	1	0	0	0	0	0	1	0	0	0	0	0	Wrote the code and conducted all statistical analysis for reproduction of Smith et al. 2016: Ethnic composition and friendship segregation: differential effects for adolescent natives and immigrants; jointly with Johanna Gereke, Nan Zhang, Fabian Winter

Ljiljana Lazarevic	0	0	0	0	0	0	0	0	0	0	0	0	0	1	I served as peer reviewer of 15 reproduction designs (Bhattacharjee_JournPerSocPsy_2017_Br0x_7g66, Lee_EvoHumanBehavior_2018_AvWY_8g91, Pfattheicher_JournPerSocPsy_2018_521q_286, Vollmann_EurJournPersonality_2011_x3KP_k5z, Ståhl_JournPerSocPsy_2012_gbl9_393, seuntjens_journpersocpsy_2015_PNPz_5zg9, alves_psychologsci_2018_AvOr_92g, fritz_journorgbehavior_2010_zekm_k17, Sternisko_covid_GG7d_m489, Du_covid_2NAG_41z0 Torelli_JournPerSocPsy_2010_gbAY_g7g1 Pennycook_covid_7NEL_9k1y Pfattheicher_covid_yZD4_y006 Imhoff_covid_dPzV_z1k9 Axt_JournExpSocPsych_2018_zK2_m5g9)
Katelin E. Leahy	1	0	1	0	0	1	0	0	1	0	1	0	1	1	For our replication project (Nelson), I (with help from fellow graduate students) programmed the survey and edited/spliced/hosted the videos for the study. I coordinated data collection and training of RAs, and supervised data analysis/reporting (with help from fellow graduate students). I worked with my fellow graduate students, Jeewon Oh and Mariah Purol, to draft the pre-registration, conduct analysis, and write up results and the final report. I also helped compile the necessary files and syntax to submit for publication.
Jessica C. Lee	0	0	1	0	0	1	0	0	0	0	0	0	0	0	I conducted an author data reproduction of Sandra & Otto (2018). I sourced the data, and attempted to reproduce the statistical models and results.
Christopher Limnios	0	0	1	0	1	1	0	0	1	0	1	0	0	0	Wrote scripting code to replicate statistical claims made by the authors of the original paper.
An-Chiao Liu	0	0	1	0	1	1	0	0	1	0	0	0	0	0	
John Wills Lloyd	0	1	0	0	0	0	0	0	0	1	1	0	0	0	I was pleased to collaborate with teams of researchers and reviewers as they developed and refined their plans for replications. I coded data about journals' policies regarding availability-accesss to replication resources.
Ruben Lopez-Nicolas	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted one reproduction study. Specifically, I was in charge of the project: Hoffman_JournAppPsych_2015_DEmL.
Nigel Mantou Lou	0	0	1	0	0	0	0	0	0	0	0	0	0	1	I conducted a reproduction analysis (Wilfahrt_WorldPolitics_2018_k7wj)
Richard E. Lucas	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I edited reproduction studies. I cannot remember the exact number of studies I edited, but there were multiple.
Maximilian Maier	0	0	1	0	0	1	0	0	0	0	1	0	0	1	My colleague and I conducted reproduction of one study (Hansen_JournExpSocPsych_2014).
Daniel J. Mallinson	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted reproduction analysis as part of one replication study. I am unsure if the studies that I served as a peer reviewer for had a reproduction component or not.
Marcel Martončik	0	0	1	0	1	0	0	0	0	0	0	0	0	1	I conducted two reproduction studies (Ohtsubo_675wo, Nelson_2w9oz).
Michael C. McCall	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I conducted two reproduction studies, PIETRYKA_AmPoliSciRev_2017 and Li_JournExpPoliSci_2017. I worked with Sebastian Karcher on both.
Nikita Mehta	0	0	1	0	0	0	0	0	1	0	0	0	0	0	I worked on 4 reproduction studies: 1) Push-button/Author-data reproduction (PBR/ADR) for Bruner (2017) 2) Push-button/Author-data reproduction (PBR/ADR) for Lindqvist & Ostling (2010) 3) Source data reproduction (SDR) for Liu et al. (2016) 4) Source data reproduction (SDR) for Lindqvist & Ostling (2010) All of these reproduction studies were done in close collaboration with Arathy Puthillam.
Esteban Méndez	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
Johannes Michalak	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
Daniel C. Molden	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted the computational reproduction Sandra_Cognition_2018_0qar - Molden - Computational Reproduction - 756g. I also conducted the Zakharova_CompPolitStu_2014_qYr7 - Molden - Secondary Data Replication - 5z36.
Faisal Mushtaq	0	0	1	0	0	0	0	0	1	0	0	0	0	0	I performed a PBR of Rich_JournExPsychGen_2018_LbEB
Claudia Neuendorf	0	0	1	0	0	1	0	0	0	0	1	0	0	1	I conducted a reproduction of Roberts et al., 2010. Further, I validated the replication of Kim & Radoias 2015. Finally, I contributed to the review and editing of the manuscript.
Austin Lee Nichols	0	0	0	0	0	1	0	0	0	0	1	0	0	0	I served as a reviewer for Chittoor 2009 project
Gustav Nilsonne	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I served as editor and reviewer. I also did some data finding and preparation. But I am not sure which activities were for the reproduction and which for the replication part of the SCORE project. I have now gone over the reimbursement forms. I have found 1 study where I was listed as editor and 11 studies where I was listed as reviewer. There are a further 12 studies for which I have received reimbursement but where the form did not specify my role.
Ernest O'Boyle	0	0	0	0	1	0	0	0	0	0	0	0	0	0	I served as an AE for a number of replications. My duties were to independently review the protocols and submissions, then assist authors in addressing the reviewers' feedback as well as any of my own. Once all reviewer concerns were addressed, I accepted the paper for publication. I also played an incredibly minor role early in the process in terms of some procedures related to the review process.
Jeewon Oh	0	0	1	0	1	1	0	0	0	1	1	0	1	0	I worked with William J Chopik, Katelin Leahy and Mariah Purol to replicate a study. Our study team collected and analyzed data together.

Thomas Ostermann	0	0	1	0	1	0	0	0	1	0	1	0	0	1	
Abiola Oyebanjo	0	0	1	0	0	0	0	0	1	0	0	0	0	0	
Radoslaw Panczak	0	0	1	0	0	0	0	0	0	0	0	0	0	1	I worked on data preparation and analysis of Siedner_covid_P3NJ_1y2 study. I reviewed and edited final manuscript.
Yuri G. Pavlov	0	0	1	0	1	1	0	0	1	0	0	0	0	1	Reproduction of Linkenauger_PsychologSci_2009_7WjP - Pavlov_SDR - 2g9z2, https://osf.io/vh5u6/
Zoran Pavlović	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted five reproduction studies. Those were Rovny_WorldPolitics_2014_Aqgj, Robertson_BritJournPoliSci_2017_qggQ, Cohen_AmEcoRev_2015_2lb5, Gerber_BritJournPoliSci_2018_3WmY, and Bigoni_Econometrica_2015_VBx1.
Noemi Peter	0	0	1	0	0	0	0	0	0	0	0	0	0	1	I conducted the reproduction Anderson_AmEcoJourn_2011_bLe8, and I contributed to reviewing and editing the manuscript.
Kim Peters	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Served as a reviewer for a few reproduction studies.
Nathaniel D. Porter	0	0	1	0	1	1	0	0	1	1	1	0	0	0	I performed one reproduction study (Travers & Krezmein 2018), served as preregistration review editor for one reproduction study (Horvat 2011) in my role as lead preregistration review editor for sociology, and served as reviewer for one reproduction study (Teney 2016).
Mariah Purol	0	0	1	0	1	1	0	0	1	0	0	0	1	1	For our replication project (Nelson), I assisted in coordinated data collection and training of RAs, and completed data analysis/reporting.
Arathy Puthillam	0	0	0	0	1	0	0	0	1	0	0	0	0	0	
Marco Ramljak	0	0	1	0	0	1	0	0	1	0	0	1	0	0	I collaborated closely in all projects with Carolin Nast and Ferdinand Wintermantel. We conducted multiple replication projects and were involved in phase 1 and 2 of the overall projects.
Arran T. Reader	0	0	1	0	1	1	0	0	0	0	1	0	0	1	I conducted a reproduction study (Hurst_EvoHumanBehavior_2017_yypJ - Reader - 21952): https://osf.io/mr6fs/. I also provided feedback on a draft of the manuscript.
W. Robert Reed	0	0	1	0	1	0	0	0	0	1	0	0	0	0	I co-led the team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Jan Philipp Röer	0	0	1	1	1	1	0	0	1	1	1	0	0	1	I have planned and conducted a reproduction study together with Thomas Ostermann and Johannes Michalak (https://osf.io/anfk6/) and served as a reviewer for a couple of reproduction submissions. I also edited 20-30 submissions, but I haven't kept track of the exact number.
Ivan Ropovik	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I worked on several reproductions, namely Bersani_Criminology_2013_zmYY, Swanson_JournEduPsych_2016_e2, Seong_JournManage_2015_3B4j, Hofer_LearnInst_2012_rWbG, Liao_JournOrgBehavior_2016_PkXJ, and Ihme_JournExpPoliSci_2018_xYbO.
Alexander O. Savi	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted one reproduction study of Hansen_JournExpSocPsych_2014_EAa_675g9 with František Bartoš and Maximilian Maier.
Kathleen Schmidt	0	0	1	0	1	1	0	0	0	0	0	0	0	0	I completed a bushel reproduction: Savani_PsychologSci_2010_88xa - Schmidt - 6zzyw. I also prepared a second reproduction (Woltin_JournExpSocPsych_2011_Wre - Schmidt - 2y4om) but was unable to complete it because the dataset wasn't available.
Landon Schnabel	0	0	1	0	1	1	0	0	1	0	1	0	0	1	I conducted a reproduction study and reviewed reproduction studies.
Eric L. Sevigny	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I was the Co-PI on a Replication Project that also performed a Reproduction of the original study (BERSANI_Criminology_2013_zmYY_g5m-Shakya/Sevigny).
Samuel Shaki	0	0	1	0	0	1	0	0	0	1	1	1	1	1	
Shishir Shakya	0	0	0	0	0	1	0	0	1	0	0	0	0	1	I conducted Replication of a Research Claim from Bersani and Doherty (2013), from Criminology
Andrew Soh	0	0	0	0	1	0	0	0	0	0	0	0	0	0	Gathered data for replication/reproduction and cleaned up data gathered.
Angela Somo	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I conducted 20+ push-button reproduction studies. I cannot remember the exact number as I no longer have access to the email address that I had used during that time. I also completed one independent reproduction/robustness analysis for one study (Liu_JournMarket_2015_9DZl) but I believe this was for a Multi100 project (not sure how interconnected the SCORE and Multi100 projects are).
Fatih Sonmez	0	0	1	0	0	1	0	0	1	0	0	0	0	0	I managed the "Ku_JournEnvPsych_2014_YpZZ - Sönmez - Computational Reproduction - 1012" project. The data had been obtained from the original authors by the OSF fellows. I prepared the script, performed the reproduction, and reported the results.
Eirik Strømland	0	0	0	0	0	0	0	0	0	0	1	0	0	1	I was a peer-reviewer on at least one reproduction study (Li, 2011) and possibly others (but this was the one I easily found in my google mail). I also audited final reports checking for errors and reviewed and edited the final manuscript.
Jordan W. Suchow	0	0	0	0	1	1	0	0	1	0	0	0	0	1	Suchow performed reproduction studies and helped to refine the reproducibility auditing process.
Anna Szabelska	0	0	1	0	1	1	0	0	0	0	1	0	0	0	I conducted several reproduction studies (can't easily check how many because I'm moving house and have no access to my computer but will be able to check that later).

Anirudh Tagat	0	0	1	0	1	1	0	0	1	0	1	0	0	0	Abouk, R., & Heydari, B. (2021). The immediate effect of COVID-19 policies on social-distancing behavior in the United States. Public Health Reports, 136(2), 245-252. [OSF Project], joint with Varsha Ashok (Royal Holloway). Anderson, S. (2011). Caste as an Impediment to Trade. American Economic Journal: Applied Economics, 3(1), 239-63. [OSF Project], joint with Nathaniel Porter and Jing Geng (Virginia Tech); Angrist, J., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. American Economic Review, 99(4), 1384-1414. [OSF Project], joint with Akshaya Balaji (Monk Prayogshala); Gerhold, L. (2020, March 25). COVID-19: Risk perception and Coping strategies. https://doi.org/10.31234/osf.io/xmpk4 [OSF Project], joint with Hansika Kapoor (Monk Prayogshala); LaFave, D., & Thomas, D. (2016). Farms, families, and markets: New evidence on completeness of markets in agricultural settings. Econometrica, 84(5), 1917-1960. [OSF Project], joint with Akshaya Balaji (Monk Prayogshala); McDevitt, R. C. (2014). "A" business by any other name: firm name choice as a signal of firm quality. Journal of Political Economy, 122(4), 909-944. [OSF Project], joint with Akshaya Balaji (Monk Prayogshala); Thames, F. C., & Williams, M. S. (2010). Incentives for personal votes and women's representation in legislatures. Comparative Political Studies, 43(12), 1575-1600. [OSF Project], joint with Arathy Puthillam (Monk Prayogshala)
Melba Verra Tutor	0	0	0	0	1	0	0	0	1	0	0	0	0	0	I conducted several PBRs and PBR extensions for SCORE.
Karolina Urbanska	0	0	1	0	1	0	0	0	0	0	1	0	0	0	Led multiple projects - reviewing, finding datasets, preparing prereg, analysing data, reporting. Also involved in identifying claims in the earlier stage before replication kicked-off. Helped with auditing the results at the end as well.
Pieter Van Dessel	0	0	1	0	1	0	0	0	0	0	0	0	0	0	I conducted a reproduction study
Elisabeth Julie Vargo	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I reviewed several reproduction protocols. I have not kept record of how many or which. Please let me know if you would like me to retrieve this information.
Diem Thi Hong Vo	0	0	1	0	1	1	0	0	1	0	0	0	0	0	I was part of a team at the University of Canterbury who were involved in replications/reproductions of approximately 40 studies. We were involved in multiple aspects of the project, including data preparation, data analysis, and writing up of final reports for each study.
Victor Volkman	0	0	1	0	1	0	0	0	1	0	1	0	1	1	I was part of the replication projects for Liang, Lazear, and Wang(2016) and Benjamin, Choi, and Strickland(2010). In the former case, I did data analysis on entrepreneurship figures gathered from the countries used in the original experiment in the years following the finished paper. I took a much more active role in the latter case, not only adapting the original questions used in the original experiment to fit a sample of the general population instead of students, but programming an online version of this experiment using Qualtrics, facilitating meetings with the survey firm in charge of its implementation, and conducting the data analysis on the results returned.
Ke Wang	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I conducted one reproduction study on "Liu_JournMarket_2015_9DZI" (SCORE RR ID: 21474 OSF Project: https://osf.io/3mr7g).
Aaron L. Wichman	0	0	1	0	1	1	0	0	1	0	0	0	0	1	I think it was Steinmetz et al.
Jamal R. Williams	0	0	1	0	1	1	0	0	0	0	0	0	0	0	We conducted a reproduction of the research claim(s) in "Asymmetrical Body Perception: A Possible Role for Neural Body Representations", by Linkenauger, et al. (2009)
Fabian Winter	0	0	0	0	1	0	0	0	1	0	0	0	0	0	Replication of one specific result using the original data.
Ferdinand Wintermantel	0	0	1	0	0	0	0	0	1	1	1	0	0	0	I conducted one single ADR, a bushel ADR, a bushel DAR, and a single SDR together with Marco Ramljak.
Nan Zhang	0	0	1	0	1	0	0	0	0	0	0	0	0	0	Conducted 1 replication study
Ignazio Ziano	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I reviewed 5 reproduction and replication projects before they were conducted.
Cristina Zogmaister	0	0	0	0	0	0	0	0	0	1	1	0	0	0	I served as Reviewer for 1 reproduction study, and 1 study that contained both a replication and a reproduction, as well as Editor for 1 study that contained both a replication and a reproduction.
Zorana Zupan	0	0	0	0	0	0	0	0	0	0	1	0	0	0	I have served as a reviewer for 5 replication submissions, and 3 reproduction submissions. (Zhou et al., 2014, Morewedge et al., 2009, Smith et al., 2016, Muis et al., 2009, Seaton et al., 2010, Roberts et al, 2010, Al Tammemi et al, 2020, Travers&Kreizman, 2018)
Brian A. Nosek	1	0	0	1	0	1	1	0	0	1	0	1	1	1	PI of the TA1 team from the SCORE program (Center for Open Science). Contributed high-level design, visioning, and leadership for the project. Collaborated closely with the COS project leader (Tim Errington) on COS's contribution to the program. Coordinated across teams on project planning, executing, and reporting.
Timothy M. Errington	1	0	0	1	1	1	1	0	0	1	1	0	1	1	Project lead of the TA1 team from the SCORE program (Center for Open Science). Contributed to high-level design, visioning, leadership, and operationalization for the project. Coordinated across teams on project planning, executing, and reporting.

References

1. Dreber, A. & Johannesson, M. A framework for evaluating reproducibility and replicability in economics. *Econ. Inq.* **n/a**, (2024).

2. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. (The National Academies Press, Washington, DC, 2019). doi:10.17226/25303.

3. Chang, A. C. & Li, P. Is Economics Research Replicable? Sixty Published Papers FromThirteen Journals Say “Often Not”. *Crit. Finance Rev.* **10**, (2021).

4. McCullough, B. D., McGeary, K. A. & Harrison, T. D. Do economics journal archives promote replicable research?: Economics journal archives. *Can. J. Econ. Can. Déconomique* **41**, 1406–1420 (2008).

5. Vilhuber, L. Reproducibility and replicability in economics. *Harv. Data Sci. Rev.* **2**, 1–39 (2020).

6. Brodeur, A. *et al. Mass Reproducibility and Replicability: A New Hope*. https://econpapers.repec.org/paper/zbwi4rdps/107.htm (2024).

7. Pérignon, C. *et al.* Computational Reproducibility in Finance: Evidence from 1,000 Tests. *Rev. Financ. Stud.* **37**, 3558–3593 (2024).

8. Laurinavichyute, A., Yadav, H. & Vasishth, S. Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *J. Mem. Lang.* **125**, 104332 (2022).

9. Wang, S. V., Sreedhara, S. K. & Schneeweiss, S. Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat. Commun.* **13**, 5126 (2022).

10. Culina, A., van den Berg, I., Evans, S. & Sánchez-Tójar, A. Low availability of code in ecology: A call for urgent action. *PLOS Biol.* **18**, e3000763 (2020).

11. Minocher, R., Atmaca, S., Bavero, C., McElreath, R. & Beheim, B. Estimating the reproducibility of social learning research published between 1955 and 2018. *R. Soc. Open Sci.* **8**, 210450 (2021).

12. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).

13. Hardwicke, T. E. *et al.* Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. 18.

14. Stockemer, D., Koehler, S. & Lenz, T. Data Access, Transparency, and Replication: New Insights from the Political Behavior Literature. *PS Polit. Sci. Polit.* 1–5 (2018) doi:10/gdsnhv.

15. Eubank, N. Lessons from a Decade of Replications at the Quarterly Journal of Political Science. *PS Polit. Sci. Polit.* **49**, 273–276 (2016).

16. Trisovic, A., Lau, M. K., Pasquier, T. & Crosas, M. A large-scale study on research code quality and execution. *Sci. Data* **9**, 60 (2022).

17. Roche, D. G., Kruuk, L. E. B., Lanfear, R. & Binning, S. A. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biol.* **13**, e1002295 (2015).

18. Roche, D. G. *et al.* Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution. *Proc. R. Soc. B Biol. Sci.* **289**, 20212780 (2022).

19. Breznau, N. The reliability of computational replications: a study in computational reproductions. *R. Soc. Open Sci.* https://doi.org/10.1098/rsos.241038 (2025) doi:10.1098/rsos.241038.

20. Brodeur, A. *et al.* Promoting Reproducibility and Replicability in Political Science. *Res. Polit.* **11**, 20531680241233439 (2024).

21. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).

22. Gabelica, M., Bojčić, R. & Puljak, L. Many researchers were not compliant with their published data sharing statement: a mixed-methods study. *J. Clin. Epidemiol.* **150**, 33–41 (2022).

23. Gabelica, M., Cavar, J. & Puljak, L. Authors of trials from high-ranking anesthesiology journals were not willing to share raw data. *J. Clin. Epidemiol.* **0**, (2019).

24. Khan, N., Thelwall, M. & Kousha, K. Data sharing and reuse practices: disciplinary differences and improvements needed. *Online Inf. Rev.* **47**, 1036–1064 (2023).

25. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).

26. Alipourfard, N. *et al.* Systematizing Confidence in Open Research and Evidence (SCORE). (2021).

27. Vines, T. H. *et al.* The Availability of Research Data Declines Rapidly with Article Age. *Curr. Biol.* **24**, 94–97 (2014).

28. Tyner, A. H. *et al.* Extracting claims from empirical publications for the SCORE Program. *Preprint* (2025).

29. Abatayo, A. L. *et al.* Overview of the SCORE Program Methodology and Reporting. *Preprint* (2025).

30. Wood, B. D. K., Müller, R. & Brown, A. N. Push button replication: Is impact evaluation evidence for international development verifiable? *PLOS ONE* **13**, e0209416 (2018).

31. Vilhuber, L. & Cavanagh, J. Report of the AEA Data Editor. *AEA Pap. Proc.* **115**, 944–957 (2025).

32. Fišar, M. *et al.* Reproducibility in Management Science. Preprint at https://doi.org/10.31219/osf.io/mydzv (2023).

33. Huntington-Klein, N. *et al.* The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* **59**, 944–960 (2021).

34. Silberzahn, R. *et al.* Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).

35. Aczel, B. *et al.* Investigating the analytical robustness of the social and behavioural sciences. *Nature* (2026).

36. Clemens, M. A. The Meaning of Failed Replications: A Review and Proposal. *J. Econ. Surv.* **31**, 326–342 (2017).

37. Brown, A. W., Kaiser, K. A. & Allison, D. B. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proc. Natl. Acad. Sci.* **115**, 2563–2570 (2018).

38. Gelman, A. & Carlin, J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect. Psychol. Sci.* **9**, 641–651 (2014).

39. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

40. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).

41. Cummins, J. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. Preprint at https://doi.org/10.48550/arXiv.2509.13397 (2025).

42. Nuijten, M. B., Bakker, M., Maassen, E. & Wicherts, J. M. Verify original results through reanalysis before replicating. *Behav. Brain Sci.* **41**, e143 (2018).

43. Weissgerber, T. L. *et al.* Understanding the provenance and quality of methods is essential for responsible reuse of FAIR data. *Nat. Med.* **30**, 1220–1221 (2024).

44. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

45. Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R. & Debonnel, E. Certify reproducibility with confidential data. *Science* **365**, 127–128 (2019).

46. Vilhuber, L. Report of the AEA Data Editor. *AEA Pap. Proc.* **114**, 878–890 (2024).

47. Morehouse, K. N., Kurdi, B. & Nosek, B. A. Responsible data sharing: Identifying and remedying possible re-identification of human participants. *Am. Psychol.* https://psycnet.apa.org/record/2024-80872-001 (2024).

48. Nab, L. *et al.* OpenSAFELY: A platform for analysing electronic health records designed for reproducible research. *Pharmacoepidemiol. Drug Saf.* **33**, e5815 (2024).

49. Quintana, D. S. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* **9**, e53275 (2020).

50. Boedihardjo, M., Strohmer, T. & Vershynin, R. Covariance's Loss is Privacy's Gain: Computationally Efficient, Private and Accurate Synthetic Data. *Found. Comput. Math.* **24**, 179–226 (2024).

51. Khan, A. R. & Kabir, E. Resampling methods for generating continuous multivariate synthetic data for disclosure control. *J. Data Inf. Manag.* **3**, 225–235 (2021).

52. Commission, E. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. (2016).

53. Landi, A. *et al.* The "A" of FAIR – As Open as Possible, as Closed as Necessary. *Data Intell.* **2**, 47–55 (2020).

54. Abatayo, A. L. *et al.* Assessments of Credibility in the Social and Behavioral Sciences. *Preprint* (2025).

55. Tyner, A. H. *et al.* Investigating the replicability of the social and behavioral sciences. *Preprint* (2025).

Supporting Information for "Investigating the reproducibility of the social and behavioral sciences"

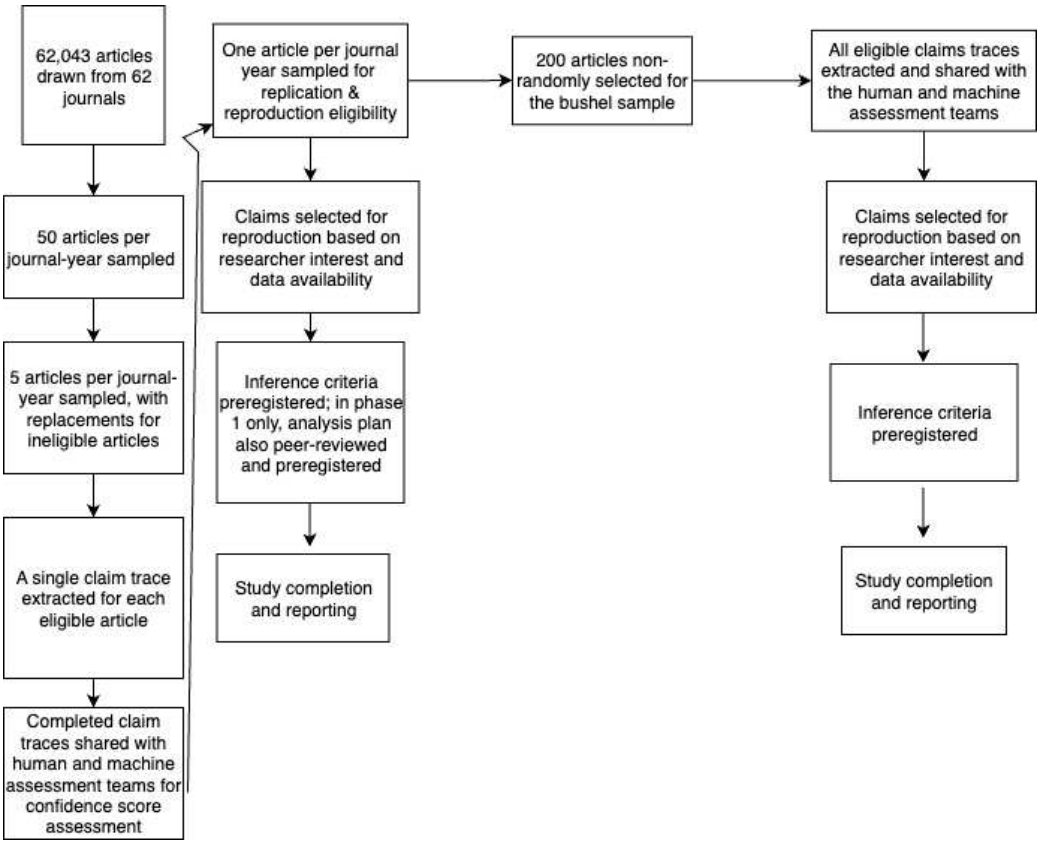
Olivia Miske, Anna Lou Abatayo, Mason Daley, Mirka Dirzo, Nicholas Fox, Noah Haber, Krystal M. Hahn, Melissa Kline Struhl, Brinna Mawhinney, Priya Silverstein, Theresa Stankov, Andrew H. Tyner, Matúš Adamkovič, Shilaan Alzahawi, Saule Anafinova, Eli Awtrey, Erick Axxe, James Bailey, Bert N. Bakker, Akshaya Balaji, Gabriel Banik, František Bartoš, Henk Berkman, Zachariah Berry, Felix S. Bethke, Timothy F. Brady, Nate Breznau, Sara Capitan, Tabaré Capitán, Kent Jason Cheng, William J. Chopik, Gwen-Jiro Clochard, Tom Coupé, Jamie Cummins, Elif Gizem Demirag Burak, Jianhua Duan, Kevin M. Esterling, Thomas R. Evans, Nathan Fiala, James Field, Victor Gay, Jing Geng, Johanna Gereke, Ilka Helene Gleibs, Amélie Gourdon-Kanhukamwe, Dmitry Grigoryev, Nicholas Gunby, Paul H. P. Hanel, Sanghyun Hong, Sean Dae Houlihan, Nick Huntington-Klein, Kamil Izydorczak, Kristin Jankowsky, Michalak Johannes, Kai Jonas, Pavol Kačmár, Hansika Kapoor, Sebastian Karcher, Marta Kołczyńska, David Kretschmer, Ljiljana Lazarevic, Katelin E. Leahy, Jessica C. Lee, Christopher Limnios, An-Chiao Liu, John Wills Lloyd, Ruben Lopez-Nicolas, Nigel Mantou Lou, Richard E. Lucas, Maximilian Maier, Daniel J. Mallinson, Marcel Martončik, Michael C. McCall, Nikita Mehta, Esteban Méndez, Johannes Michalak, Daniel C. Molden, Faisal Mushtaq, Claudia Neuendorf, Austin Lee Nichols, Gustav Nilsonne, Ernest O'Boyle, Jeewon Oh, Thomas Ostermann, Abiola Oyebanjo, Radoslaw Panczak, Yuri G. Pavlov, Zoran Pavlović, Noemi Peter, Kim Peters, Nathaniel D. Porter, Mariah Purol, Arathy Puthillam, Marco Ramljak, Arran T. Reader, W. Robert Reed, Jan Philipp Röer, Ivan Ropovik, Alexander O. Savi, Kathleen Schmidt, Landon Schnabel, Eric L. Sevigny, Samuel Shaki, Shishir Shakya, Andrew Soh, Angela Somo, Fatih Sonmez, Eirik Strømland, Jordan W. Suchow, Anna Szabelska, Anirudh Tagat, Melba Verra Tutor, Karolina Urbanska, Pieter Van Dessel, Elisabeth Julie Vargo, Diem Thi Hong Vo, Victor Volkman, Ke Wang, Aaron L. Wichman, Jamal R. Williams, Fabian Winter, Ferdinand Wintermantel, Nan Zhang, Ignazio Ziano, Cristina Zogmaister, Zorana Zupan, Brian A. Nosek, and Timothy M. Errington

Table of Contents

Overview

This supplement provides details on the methodology and additional results for the reproductions attempted during the Systematizing Confidence in Open Research and Evidence (SCORE) program funded by DARPA. Reproduction studies were just one component of the SCORE program. Background on the design and approach of the

whole program is available in Abatayo et al. (2026). Reproductions were conducted on claims extracted from a sample of papers published in social and behavioral science journals. Details about the identification and extraction of claims and the methodology of the overall SCORE program is available in Abatayo et al. (2026). This supporting information provides details on the methods for the reproduction process specifically. The image below provides an overview of the workflow for sampling and conducting reproduction studies.



Methods

In the SCORE program, we conducted reproductions of published claims as complementary empirical evidence to replication studies that served as ground truth for human and AI predictions about the credibility of social and behavioral science claims. This dataset was produced in a coordinated, distributed effort of researchers across the globe. Individual researchers and small teams provided their substantive expertise to conduct high-quality, good faith reproduction attempts, and a coordinating team provided the operational, financial, and logistical support to maintain the timeline; facilitate an internal review process to ensure projects were rigorously conducted; and conduct training in best practices for data sharing, preregistration, and outcome reporting for consistency across the project.

The priorities of this reproduction effort were rigor, transparency, and efficiency. Accordingly, this document focuses on those aspects of the process that facilitated these goals and gives relatively less attention to specific operational steps that are less relevant to understanding SCORE's reproduction evidence, such as grantmaking and managing the cooperative agreement with DARPA. The procedures documented below evolved over the three years of the program. When relevant to interpreting SCORE evidence, we highlight changes to our processes and indicate when in the project's timeline the changes were implemented.

Some of the methods reported here are similar to those reported in the supporting information for the SCORE replication paper because of shared methodology (Tyner et al., 2026). For some sections, the same initial description was used and then edited separately for the features unique to the replication and reproduction attempt methods. As a consequence, there is substantial overlap in text between the supporting information documents for this paper and Tyner and colleagues (2026).

Key Terms

Abatayo and colleagues (2026) contains a glossary of key terms for the program. A few are particularly relevant for this paper.

Reproductions involved testing the same claim as an original paper with the same analysis and same data. *Author-provided data* refers to the data made available by authors to conduct the reported analyses. For research in which the authors conducted their own data collection, there usually is not another type of data that could be available. For research in which the authors gathered data from existing sources, the original sources are noted as *source data* (or occasionally, *original data* or *raw data*) in this report. In secondary data research, source data are usually cleaned, subsetted, transformed, or joined with other secondary data by the authors to prepare the data for analysis. Authors may have included source data, derived data, or a combination in what they made available.

Reproducibility refers to successfully reproducing the original findings. *Push-button reproductions* refer to attempting a reproduction with the same analysis code and data with minimal modifications. *Source data reproductions* refer to attempting a reproduction by constructing a replica of the author dataset from "raw" source data files when the authors did not make data available. *Full reproducibility* refers to both successfully obtaining author data to attempt a reproduction and to successfully reproducing the outcomes. It is used in the supporting information when measuring reproducibility rates in the original sample of all papers that could, in principle, be reproduced if data were obtained and reanalyzed.

Sample and Data

Claims for potential reproduction were drawn from >27,000 social and behavioral science papers published from 2009 to 2018. Additional details about the journal selection process, paper selection process, and claim extraction process are reported in

Abatayo et al. (2026). Here, we provide information to supplement the main text for understanding selection of papers and claims for reproduction attempts.

The project was conducted in two phases following standard practices for DARPA programs. All of the reproduction attempts were conducted on a sample created during Phase 1 of the program, though many of the reproduction attempts occurred during Phase 2. This differs from other evidence-gathering parts of the program. For example, replications were conducted on papers selected during Phase 1 and Phase 2, and have a larger total sample as a consequence (Tyner et al., 2026). The report on reproductions focuses only on the sample of papers from Phase 1 as none of the papers selected during Phase 2 were subjected to reproduction attempts.

A stratified random sample of 3,000 papers was selected from the >27,000 papers comprising the full dataset. Those 3,000 papers, sometimes referred to as the Annotation dataset, each had a single claim extracted for evaluation by human and machine teams. A random sample of 600 papers was drawn from that dataset that maintained representativeness across journals, disciplines, and year of publication. This dataset, sometimes referred to as the Evidence dataset, was the basis of reproduction attempts. All 600 papers were examined for data availability and a subset were examined for reproducibility.

During the program, a subset of 200 papers from that 600 was created non-randomly, focusing first on papers for which replication or reproduction attempts had been conducted, second on papers likely to be amenable to conducting replication and reproduction attempts, and third on retaining relative representatives of papers across disciplines in the subset of 200 papers. The 200 papers went through an additional claim-extraction process during Phase 2 to code all eligible claims from those papers instead of just a single claim. Reproduction attempts on this dataset, sometimes referred to as the Bushel dataset, could have tested the reproducibility of multiple claims in a single paper.

Data Availability Assessment

Overview

We conducted an assessment of data availability for all 600 papers in the Evidence dataset from the SCORE program, along with 100 empirical preprints from a COVID-19 preprints dataset. Because of their distinct origins, reproductions of the COVID-19 papers were not examined or included in this report. See Marcoci and colleagues (2024) for some reporting on the research outcomes with the COVID-19 papers. Each paper went through a coding process to assess the availability of data and code from within the paper itself and through an online search for publicly available data and code. We also conducted outreach to the original authors to request original data and code when either could not be found publicly. The sections below describe these processes.

Paper Assessment

Before reaching out to authors to request data and code, the coordinating team conducted an assessment of each paper’s publicly available data and code. This

included a review of the original paper, the journal's website, the authors' website(s), and any other relevant sources for material availability that could facilitate a reproduction attempt. Four team members conducted this coding for each of the 600 papers in the evidence dataset over the course of five months from November 2022 to April 2023. This comprised the data availability assessment reported in the main text. There were two parts: paper review and online search.

Paper Review

When coding for data availability, each coder first reviewed the paper for any information on data and code availability. Coders first assessed whether the paper relied on new or secondary data. Coders considered "new data" studies to be any study that generated new data for the research, such as the authors administering a survey or conducting an experiment. Coders considered "existing data" studies, also called secondary data studies, to be any studies that gathered data from existing data sources (e.g., census bureau data, financial databases, firm data). If an original paper used both new and secondary data, then coders identified them as a "combination" study.

Coders next reviewed the paper for data and code availability statements and any links or references to shared data or code. If coders did not find a statement, they then checked the methods or data sections for links or descriptions of how the data were obtained. If coders could not find links or references to the data origins, they would search for key terms in the paper that were associated with data or code availability including "supplement," "online," "material," "http," "data," "code." Coders documented whether the paper had links to data or code, or statements about how to access the data or code including that the data was available on request. Coders followed available links to check if they resolved at a location that appeared to provide the data or code.

Online search

After completing the paper review, if coders had not found information about accessing data or code, then they conducted a web search. Coders limited web searches to about 10 minutes per paper. Coders used a search engine to search for the paper title, the paper title with author names, and author names. Coders looked for search results related to the publisher/journal website of the original paper, common online data repositories (e.g., ICPSR, Harvard Dataverse, OSF), and websites of the original authors. Any discovery of data or code was documented with information about the location, a link, and a brief description of whether the data or code appeared to be complete, accurate, and accessible — directly or via restricted access procedures. Coders concluded the search after finding data and code, or after 10 minutes of searching, whichever came first.

Author Outreach

We reached out to authors of the papers several times during the program. Authors of all 3,000 papers in the Annotation dataset from Phase 1 were informed about the project, told that their paper had been selected randomly, and asked for their feedback about the claim extracted from their paper.

For the 600 papers eligible for data availability assessment, if either data or code were not located from the steps outlined above, we requested the missing material (either data, code, or both) directly from the authors.

Coding Data and Code Availability

Data and code was documented as being available online (i.e., we found publicly accessible data, or we were able to access the data after consenting to a data access agreement/terms of use or after logging in with institutional credentials), privately shared (e.g., provided via email in response to our request for materials), or not available.

Limitations

Data or code that was evaluated as “not available” could have received that designation for reasons beyond the authors’ control.

For secondary data research, access to the data could be restricted by the originator. We did not try to access restricted data beyond steps of consenting to data access agreements or logging in with institutional credentials. There were 11 documented cases in which restricted data were identified in our sample that we did not obtain, and were thus coded as not available. Below are the notes describing the circumstances and identifiers to access the full documentation.

Paper ID	PDF name	Quote providing circumstances of restricted data
LmA2	Carrell_AmEcoJourn_2010_LmA2	The data are proprietary as noted by the readme (included in the reproduction package stored on ICPSR). The readme includes instructions for how to acquire the dataset - from readme.pdf file: "Data for AEJApp-2009-0014, "Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids," is proprietary and owned by the Alachua County, Florida School District. The corresponding author, Mark Hoekstra (markhoek@pitt.edu), holds the deidentified dataset (AEJApp-2009-0014_rawdata.dta) used for the project and will provide copies to authors who receive written permission from the Alachua County Public Schools..."
RqP7	Jacob_AmEcoJourn_2009_RqP7	The data are restricted and cannot be shared directly, however the authors provide specific instructions in the readme for how to acquire the original data and prepare it in order to reproduce the analyses from the original paper. From READ_ME_AEJApp-2007-0053.pdf: “We obtained the underlying student-level data used in this analysis from the Chicago Public Schools (CPS) on a restricted-use basis that prohibits us from sharing the data with others. However, other researchers are able to apply to the CPS to obtain access to this data. For information on how to apply for such data, please see the CPS website: http://research.cps.k12.il.us/cps/accountweb”
GOAw	Kline_AmEcoRev_2016_GOAw	None of the existing MDRC Jobs First data is available and needs an application to request access. The readme pdf included in the additional materials describes the files that the code is expecting to use once a researcher gains access to the Jobs First data and how the author's gained access themselves: "In these notes and associated files, we provide computer code that allows a user to replicate our results given access to these data. Specifically, the Stata code assumes that the sub-directory MDRC contains the file “ctadmrec.dta”, the main administrative-file from the Jobs First Experiment. We obtained this file by following the application process described at http://www.mdrc.org/available-public-use-files#bookmark4 and then converting to Stata dataset format the MDRC’s file “ctadmrec.sas7bdat”."
p3r0	Sarvimäki_JournLabEco_2016_p3r0	"The file “integration_estimation.do” includes Stata code used for producing the tables and figures. The data have been obtained by combining several administrative registers collected and held by Statistics Finland and the Ministry of Employment and the Economy. These data can be accessed through Statistics Finland’s remote access system. Details on data access policy and application procedure are discussed below."
wNKW	Hendricks_QuartJournEco_2018_wNKW	"Before proceeding, it is important to note that the main data set for this paper is the restricted access version of the New Immigrant Survey. Since it is restricted, it is not provided with these replication materials. Instead, if you wish to replicate the empirical results you must apply for and be granted access to this data separately. In the language of the NIS, you need access the restricted data version 1, which is somewhat less difficult to apply for. You will need to submit a proposal and work with your institution's IRB, primarily on a data protection plan to insure that the data remain secure. The details are available here: http://nis.princeton.edu/data_restricted2.html"
ADPO	Fevang_JournLabEco_2014_ADPO	"The data used in the paper comprise complete longitudinal administrative records on employment and physician-certified sickness absence 2001-2006, merged with information on firms and workers on the basis of encrypted identification numbers. Due to the sensitive nature of the data, the authors’ of this paper are not allowed to pass the data on to other users. Researchers working at certified research institutions can apply for access through Statistics Norway; see http://www.ssb.no/a/mikrodata/main.shtml (in Norwegian);" "Researchers interested in replicating our analysis are advised to contact one of the authors in order to obtain more information."

BbNg	Persson_AmEcoRev_2018_BbNg	The study relies on sensitive data from the Swedish National Board of health, all of which requires permission to access. The authors detail the steps to take to apply for access to the data in their "readme", which is hosted on ICPSR (https://www.openicpsr.org/openicpsr/project/113015/version/V1/view), and linked to on the journal's website (https://www.aeaweb.org/articles?id=10.1257/aer.20141406).
w8Zp	Mayer_AmEcoRev_2014_w8Zp	From the read-me-file: 'Due to confidentially agreements with the data providers, the data used in this analysis is not available for public distribution. The dataset we use in the paper is currently commercially offered by BlackBox Logic and Equifax. The other researchers could purchase and access such data from these data providers.'
DDj2	Goodman_JournLabEco_2013_DDj2	"The data used for the project are proprietary and cannot be freely posted online. Other researchers can gain access to these data for replication purposes by contacting the College Board here: The College Board 250 Vesey Street New York, NY 10281 212-713-8088 Or through the direct online portal here:http://research.collegeboard.org/data/request"
A1QY	Björnerstedt_AmEcoJourn_2016_A1QY	"The main dataset consists of confidential product level data for the Swedish analgesics market for the period January 1995-December 2009, which can be obtained from Apoteket AB, and for the period January 2010-May 2011, which can be obtained from Apotekens Servicebolag AB."
QAB1	Imbert_AmEcoJourn_2015_QAB1	"This paper draws on several data sets that are either commercially or freely available. Copyright restrictions do not allow these data to be redistributed, however interested researcher can purchase the data and reproduce the results of the paper using the attached dofiles.1. Schedule 10 data from National Sample Surveys Rounds 55 61 64 and 66. Data can be purchased from http://mospi.nic.in/Mospi_New/upload/nsso/ratelist_UnitData.pdf 2. Data on population, literacy, labor force participation, area and irrigation from 2001 Census: http://censusindia.gov.in/2011-common/censusdataonline.html 3. Yearly data on output and harvest prices published by the Ministry of Agriculture:http://eands.dacnet.nic.in/ 4. Rainfall data from the Tropical Rainfall Measuring Mission (TRMM): http://trmm.gsfc.nasa.gov/ 5. 2009 ARIS REDS Data on household hired labor: http://adfdell.pstc.brown.edu/arisredsdata/readme.txt 6. Data on elections from the Electoral Commission of India: http://www.eci.nic.in/ecimain1/index.aspx 7. Only reports on roads built under the Pradhan Mantri Gram Sarak Yojna (PMGSY): http://pmgsy.nic.in/ 8. Rural Laborers Consumer Price Index from the Labour Bureau: http://labourbureau.nic.in/indexes.htm"

The outreach to authors occurred sporadically over a long period of time between 2020 and 2023. Willingness to make data available might have been undermined by the periodic correspondence or insufficient follow-up by the coordinating team.

The coordinating team followed a process for searching for public data and code, but may have made errors or failed to search sufficiently. Authors may not have received the requests because of spam filters, changing institutions, or other barriers.

Our definition of available data is constrained to author-provided datasets. In some cases, the authors did not provide data, but there exist public source data (e.g., Census data) that might be sufficient to assess reproducibility. The main text reports evidence that reproducibility was much weaker for cases in which we found and obtained source data and attempted to reconstruct the authors' preparation and analysis pipeline. Nevertheless, our definition of data availability informs the interpretation of the findings in this research.

Finally, key parts of the SCORE program occurred during the pandemic, sometimes during the lockdown periods during which researchers may not have been able to

access the computers or files where data or code was stored. “Not available” is a statement of the outcome of the process, not an attribution of responsibility.

Reproducibility Assessment

Overview

There were three ways in which reproduction attempts could occur: push-button, Phase 1 process, and Phase 2 process. The process changed between Phase 1 and Phase 2 to simplify components that were deemed burdensome and unnecessary for conducting reproductions. Some features of the process were common across the three ways of conducting reproductions. Also, on several occasions, a reproduction was planned, reviewed, and conducted by the same analyst doing a replication of the same paper as reported in Tyner and colleagues (2026). Shared and unique features of the reproduction processes are identified in the sections below.

Sourcing Reproduction Analysts

The recruitment of teams to attempt reproductions, referred to as sourcing, was facilitated by construction of a dataset of expert individuals and laboratories that represents the collective resources of the Center for Open Science (COS) through its several large-scale replication and reproduction projects and COS’s partners: the Psychological Science Accelerator (mostly psychology and behavioral economics laboratories; https://psysciacc.org/), the Berkeley Initiative for Transparency in the Social Sciences (BITSS, an extensive network of economists, sociologists, political scientists, psychologists, and other social scientists; http://bitss.org/), , and the International Initiative for Impact Evaluation (3ie, which has access to a global team of researchers from a variety of social sciences, particularly developmental economics; http://www.3ieimpact.org). Each of these groups has substantial experience conducting replications or reproductions and had expressed interest in participating in this program. Leveraging these networks meant that most researchers in the database had experience with replication or reproduction studies. The replication studies are reported in Tyner et al. (2026).

Potential contributors to the SCORE program responded to calls for collaborators by completing a short survey about their analytical expertise and available resources (e.g., analytic software/coding languages, computing power). Survey respondents, along with individuals who were recruited through social media and word of mouth, comprised the SCORE collaborators email Google group. More than 200 researchers participated in the replication and reproduction efforts, many of whom completed multiple projects. Sourcing projects to repeat performers reduced the onboarding cost and positively contributed to the scalability of the program.

We employed a self-selection method in which analysts selected projects using Google sheets that provided the original paper title, DOI, relevant metadata (e.g., discipline, journal, year published), the key claim(s) and result(s), and a link to any original materials available that would be relevant to conducting a reproduction attempt (e.g.,

author-generated datasets, analysis code or instructions). The sheets were distributed via email to collaborators who signed up to be reproduction analysts.

Onboarding Teams to Attempt Outcome Reproductions

Once an analyst was matched with a reproduction project, Project Coordinators sent onboarding email and additional instructions. Project coordinators confirmed that analysts could complete the reproduction with the available resources and time. Once confirmed, the coordinators created a unique ID number for the reproduction attempt, then created and shared an OSF project, the relevant preregistration and reporting templates, and additional instructions with the analyst. As needed, the coordination team provided guidance to reproduction analysts regarding using the OSF and adhering to preregistration and documentation standards for the project.

Preregistration

Reproduction teams preregistered their attempts. In Phase 1, reproduction teams articulated the claim to be evaluated and described their analysis plan before conducting any analyses following standardized reporting formats. These protocols were scrutinized by an independent editor who evaluated the strength of the proposed methods and appropriateness of the design for testing the same question as the original study. For secondary datasets, two questions about the dataset needed to be answered "yes": [1] Is the final reproduction dataset that the research team constructed suitable for performing a high-quality, good-faith reproduction of the focal claim selected from the original study?; [2] Is the procedure for constructing the final reproduction dataset sufficiently documented that an independent researcher could construct the same dataset following the procedures and code they provide? For new and secondary datasets, two further questions and a final assessment needed to be answered "yes": [1] Is the analysis plan (including code) that's documented in the preregistration consistent with a high-quality, good-faith reproduction of the focal claim selected from the original study?; [2] Has the data analyst demonstrated that the analysis code works as expected on a random 5% of the final reproduction dataset?; and, [3] I have reviewed all sections of this preregistration, and I believe it represents a good-faith reproduction attempt of the original focal claim. After reviewing the preregistration, if the protocol met the criteria, the editor approved the reproduction study to move forward.

In Phase 2, reproduction teams preregistered the inference criteria and not the analysis plans. These were reviewed by project coordinators rather than by a recruited editor and peer reviewers.

Push-button reproductions were preregistered following the processes described in the push-button section below.

Phase 1 Process

Drafting preregistrations

In Phase 1 of the program, reproduction teams described their research plan using a Source Data Reproduction preregistration template or an Author Data Reproduction

preregistration template. The specific claim was provided by the coordinating team to the reproduction team. The preregistration forms were based on the standard OSF preregistration template. They included SCORE-specific instructions to guide a researcher through each step. The coordination team provided guidance and answered questions as needed.

Recruiting editors

During recruitment of collaborators, researchers could indicate interest in conducting studies and interest in serving in editorial roles. Program leaders conducted personal outreach to researchers with some experience in editorial or reviewer roles at disciplinary journals across the social-behavioral sciences to participate as editors for SCORE. Table S1 identifies the Editors that reviewed one or more reproduction studies. Reproduction studies for which they served as editor are identified by their OSF ID which can be found by replacing “abcde” with the five-character ID in the following URL schema: https://osf.io/abcde. Editors were responsible for reviewing and approving the submitted preregistrations. Editors could engage independent reviewers if needed, but rarely did so for reproduction studies.

Table S1. Editors for SCORE reproduction studies peer review process.

Name	Institution	Title	OSF Links
Amélie Gourdon-Kanhukamwe	Kingston University	Lecturer	8btme
Anna Szabelska	Psychological Science Accelerator		p45bu, uyzc4
Bert Bakker	University of Amsterdam (Amsterdam School of Communication Research)	Associate Professor	y5uwg, jp7tr, 7p5tw
Bill Chopik	Michigan State University	Associate Professor	anfk6, 6ye5m, tqpb5
Eli Awtrey	University of Cincinnati	Assistant Professor	kzpf8
Elisabeth Julie Vargo	Institute for Globally Distributed Open Research and Education		h72nm
Gustav Nilsonne	Karolinska Institutet (Department of Clinical Neuroscience)	Associate Professor	8sae9, mshda
Hansika Kapoor	Monk Prayogshala and University of Connecticut		7ak4n, ve9tx, 5ywth
Ignazio Ziano	University of Geneva	Assistant Professor	2a3fx
Kai Jonas	Maastricht University (Work and Social Psychology)	Professor	wv2gh, pkwgx
Michael Mullarkey	Aiberry	Senior Data Scientist	vufm2, c8u5q
Nathaniel Porter	Virginia Tech (University Libraries)	Assistant Professor	q5szk, 4rjbf
Onurcan Yilmaz	Kadir Has University	Associate Professor	2vust

Engaging original authors

For reproductions, we decided not to include original authors during the review process because the data was accessible to the authors creating circumstances in which original authors' could conduct reanalysis during review, and potentially introducing unwelcome influence on the review process. This is different from replication attempts, in which original authors were invited to participate in the peer review process (Tyner et al., 2026).

Phase 2 process

Transparency Trail

In Phase 2, we simplified the preregistration and review process after determining it was more burdensome than necessary. Analysts preregistered their inference criteria by filling out a Reproduction Criteria template rather than a full analysis plan, and these preregistrations were reviewed internally rather than going through the full independent peer review. This allowed analysts to work more flexibly with the materials they collected or were provided rather than being constrained to the single analysis plan they preregistered. This was simpler and better aligned with the goal of determining whether the original materials can be used to reproduce the original findings, where constraining researcher degrees of freedom is less of a concern. In lieu of preregistered analysis plans, analysts were required to report in a 'transparency trail' each of the analyses they performed before they determined whether or not they were able to reproduce the claim.

Push-button reproduction process

In both Phases 1 and 2, if we had both data and code available for a reproduction attempt, then we could achieve *push-button reproducibility*. Push-button reproducibility is achieved if the paper's outcomes are reproduced with minimal effort by the independent analyst other than applying the original code to the original data.

We used a standardized process for attempting push-button reproducibility that preceded the workflow described above. Analysts filled out a push-button reproduction preregistration template (Phase 1) or a reproduction criteria template (Phase 2) to specify the criteria that would be used to evaluate the reproduction outcomes. The reproduction criteria were then uploaded to OSF and registered. Then analysts were instructed to spend up to 30 minutes to conduct the push-button attempt, not including computation time if that was intensive. However, we did not monitor or hold analysts strictly to this timeframe. As such, whether analysts could reproduce the findings with minimal revisions to the data or code was effectively a discretionary assessment of the analyst. If the outcomes were reproduced successfully, then it was considered a successful push-button reproduction. If not, then the same analyst or a different analyst could attempt a reproduction effort as described in the prior section, with the flexibility and time to revise and adapt the author-provided code. Failed push-button reproduction attempts could also be picked up by other analysts and put through the "regular" reproduction process.

Inferential criteria for reproduction success

Reproducibility was coded as one of four possible outcomes: push-button reproducibility, precise reproducibility, approximate reproducibility, and not reproduced.

Claims were rated as *push-button reproducible* (a subset of *precisely reproducible* for reporting purposes in the main text) if [1] original data and code were available, [2] the code could be executed on the data with minimal revisions to the code, and [3] the observed outcomes precisely matched the reported outcomes. Claims were rated as *precisely reproducible* if [1] original data were available, [2] code written by the analyst, or the original code, could be executed on the data after some revision, and [3] the observed outcomes precisely matched the reported outcomes. Claims were rated as *approximately reproducible* if [1] original data were available, [2] the original code or code written by the analyst could be executed on the data, and [3] the observed outcomes were within 15% of the continuous reported outcomes and within .05 of the reported p-value. Claims were rated as *not reproduced* if the original data were available, but the other criteria were not met. Note that “not reproduced” is not synonymous with “not reproducible.” It is possible that some of the claims could be reproduced if issues confronted by the analyst could be resolved. However, in some cases, this is unlikely because of problems that have no obvious means of resolution, such as different sample sizes and reported analysis strategies that cannot be conducted with the original data.

Outcome reporting

Reproduction teams authored reports of their observed results and a comparison with the original study. In Phase 1, reporting templates specific to each reproduction type were provided to analysts. In Phase 2, a transparency trail reporting template was provided to reproduction teams to report their reproduction outcomes and deviations or additional steps that occurred during the reproduction attempt. Analysts also received instructions for uploading relevant files to the respective OSF project. After verifying that the written report was complete, a project coordinator filled out a Variable Form on behalf of the analysts to incorporate key variables and outcomes into the dataset. Once the reproduction variables were complete, a coordinating team member and statistical consultant would assess the reporting and calculate any missing variables from original studies, reproduction studies, and those used to evaluate whether reproductions were successful. Those results were then reported in a standardized format for extraction to the database and for referencing to the study code and data.

Table S2 provides links to all OSF projects for reproduction attempts that were started. Paper ID and Project ID columns provide the project-specific identifiers used for tracking and project management. For any given project, replace “abcde” in the URL schema https://osf.io/abcde with the five characters in the OSF column to find the plans, materials, data, and reporting on OSF. The “completed and reported” column is marked “yes” if the project met inclusion criteria and outcomes are reported in this paper.

Table S2. Identifiers and links to reproduction attempts

Paper ID	Project ID	OSF	Completed & Reported
0P4r	28884	jcbfw	Yes
0P4r	5066	ysncd	Yes
0PZl	g241	zfxk9	Yes
0PZl	41y2	5chvj	No
0a3Z	6okm6	nhecx	Yes
0a3Z	y401	b6n9x	No
0qar	6my96	yh25j	Yes
0qar	756g	fhczq	No
1574	2g5g	y5uwg	Yes
1574	2w9go	3a2r5	No
1Zx7	2w8w6	my426	No
2GKO	4142	nse8t	Yes
2lb5	2g781	vk2a3	Yes
2lb5	174z	m4hqw	Yes
3B4j	6zzok	wv65j	Yes
3WmY	675ko	hyqpr	Yes
3aPw	05g8	qczeu	Yes
3zRW	6797	swkur	Yes
4XLv	3gg3	ks6ut	Yes
4q0L	3z5z	m4yse	Yes
5Awm	2wkk2	kmqp2	Yes
5KrD	1y52	td3kh	Yes
5PyD	2k4w6	sc87r	Yes
5PyD	600k	g9aqj	Yes
7R9G	6m796	tk7y4	No
7R9G	2yk82	kh6rp	No
7WjP	2g9z2	vh5u6	Yes
7WjP	927	p45bu	No
7WjP	67m16	gnrvf	No
7X54	93k7	7qnrs	Yes
7X54	21k52	9gauh	Yes
7X54	65ym6	bxt5n	No
7d4J	5g68	ezx2k	Yes
7ybJ	k97z	8xyhb	Yes
88xa	6zzyw	jfxnt	Yes
88xa	6168	4fvzq	Yes
88xa	21444	frwyc	No
8R9d	69yok	yrh4v	Yes
8R9d	23312	8kyfe	No
8R9d	285g	tq2z6	No
8Wy0	1564	vfyxz	Yes
9DZl	21474	3mr7g	Yes
9DZl	0008	5kj8c	No
9Gkl	21752	vfqt8	No
9OK1	247gz	m9tej	Yes

9OK1	6oy46	qs827	Yes
9OK1	y486	ka496	No
9XrX	m93	fspwc	Yes
9ey	69y31	jzfs2	Yes
9ey	281k5	253zd	Yes
9lBL	2kg19	zsptj	Yes
9lBL	g45m	z9epb	No
9wya	9k2y	4w5g2	No
9wya	k8z7	38zs9	Yes
9wya	288zk	gb46p	Yes
AQgj	6m3zm	u3y9w	Yes
AQgj	o08	m98g4	No
AXBY	2w9ko	scqv3	Yes
AYQG	6g7k	cfra4	No
AgO1	9y8y	2vust	Yes
AqDO	65z92	7msd8	Yes
AvOr	21352	vndky	Yes
AvOr	mzk9	ecmtp	No
BebG	42k8	9n8uh	Yes
BlRQ	28894	xbvs6	No
Blxd	6778	ve9tx	Yes
Blxd	6om46	9avck	No
BrGp	6zz1k	wpk3g	Yes
BrGp	24783	6mv2x	No
ByBk	g931	rpc54	Yes
D2LY	k637	c8u5q	Yes
DDj2	kzmz	k9pvw	Yes
DEmL	6zzzk	7fvtn	Yes
DEqr	2g486	5vx27	Yes
DEqr	3g03	zdpy2	No
E0Q3	z789	j6c8g	Yes
E4Am	69y39	axtg4	Yes
E4Am	93mg	8tkwe	Yes
E5qr	95y	cs8y2	Yes
EAa	675g9	akubt	Yes
EKBZ	191z	a7mys	Yes
EQxa	3z3k	7ak4n	Yes
EZ3x	69yw9	8t3dy	No
EdQy	2y9w2	gpyu7	Yes
EdQy	6g28	sv3gb	Yes
G0Kb	g2z	5k8m4	Yes
G1Lr	2zg7	hbrwf	Yes
G4mp	67519	s57jr	Yes
G55r	302k	ujmvw	Yes
GJe4	2g73y	f3tp2	No
GJe4	y4m6	mkr6p	Yes

GOYb	50y8	5vntp	Yes
GQvr	6m37m	2v5q9	Yes
Gv3O	65996	563d4	Yes
Gv3O	mz17	yjxce	No
J0Yv	2k7w2	frc3x	Yes
J7Z2	2kg39	rv724	No
J999	6zzgw	35tgu	Yes
J999	8zgg	k26gr	No
J999	6g08	qmk7t	No
JRpA	6m516	fpuh3	Yes
JRpA	8297	kh2dp	No
JxXe	67o46	qd3rm	Yes
JxXe	969y	eahxw	No
JxXe	2k8g6	ndfe3	No
KRgk	7my5	xue5d	Yes
Kj9d	5196	8btme	Yes
L22B	2g182	gk6mh	Yes
L22B	96g	hbzu3	No
La9x	0036	tgkhv	Yes
LbEB	21mg2	u8zk5	Yes
LbEB	69736	uc9ny	No
LbEB	y041	jpu9q	No
LmA2	2kgk8	3axzu	No
LyWB	2w93z	qu3p2	Yes
LyWB	g28z	qw3e4	No
Njqj	6o5m6	c5adw	Yes
Njqj	g1m	4axu6	Yes
Nv99	6ow1o	2uxrt	Yes
Nv99	2816	guf6v	No
OY3B	2gwz2	qnvt2	Yes
OY3B	k2m7	h72nm	No
OYX0	2y4km	pndku	Yes
OeGv	2y3w2	q5ka2	Yes
OeGv	4980	yndwz	No
OeGv	241w6	2n3c7	No
Ovkm	78gg	4x3b9	Yes
P1rY	6z8o6	g2jua	Yes
P1rY	328k	dkr2s	No
P8az	32z3	p7tb4	Yes
PVQK	69y19	v2yaq	Yes
Peaa	m7k3	nekdp	Yes
PkXJ	288y4	eu2yr	Yes
Pxp7	288kk	4y8ja	Yes
Pxp7	7226	u3th5	No
Q1dl	2w94o	c4ar6	Yes
Q1dl	05k6	dk58a	No

QYNq	2zk7	8wf3h	Yes
R0ak	17y4	6wfmz	Yes
RYKv	2k5g2	df36m	Yes
Rjp9	5yg9	tmg3z	Yes
RqVE	9k3g	4p892	Yes
Ryq7	2yz86	ezxr5	No
Ryq7	95my	jegqs	Yes
V0PA	57g6	jp3qb	Yes
VB9K	6m3gm	3c84w	No
VBx1	6m31g	5fc4n	Yes
VDJV	6z582	8eqk6	Yes
VDJV	g6yz	n8wgd	No
VRKK	6mw96	d482q	No
VRKK	2zmg	2z6j4	Yes
Vx4e	0937	d4p59	Yes
W0GN	g7z1	adxj2	Yes
WLkV	gy3z	mjb97	Yes
WLpV	214k4	u6bea	No
Wre	316k	f5m4b	Yes
Wre	2y4om	zsq2u	No
Wre	67316	edgj6	No
YOXl	g8m	yfvzr	Yes
YRvg	3g4k	b48am	Yes
YRvg	yy01	axds7	No
YabW	2y474	57pn2	Yes
YabW	24773	ktx9s	No
YeQg	m4m7	6e9ka	Yes
YpZZ	2kgyw	v9ykq	Yes
YpZZ	1012	vufm2	Yes
Z0ma	8m1	f3p2m	Yes
ZaZK	23w7z	tzbfh	No
ZdgL	2637	uyzc4	Yes
a2Yx	28m96	qk9v3	Yes
a2Yx	96ky	yswvd	Yes
amYY	235w2	3a5u7	Yes
amYY	m9k3	sfwqd	No
bLe8	24733	h2u96	Yes
d2O3	658w7	cgz3v	No
e2pq	2kgw8	yv5k4	Yes
e5rW	67746	namvy	Yes
e5rW	1962	37jd4	Yes
eg1q	2w9oz	x9pd3	Yes
eg1q	80yg	2a3fx	No
exBp	38y3	6ye5m	Yes
g0XQ	6owm3	drcnw	Yes
gdlO	3z03	qnmrj	Yes

jDWN	2w1k2	axfzm	Yes
jaK4	6727	jhe6q	Yes
k7wj	2kgg8	urfmp	Yes
kXp8	kzyz	925nz	Yes
ky28	zz11	7p5tw	Yes
l22v	68	psmq5	Yes
lxXV	y791	q5szk	Yes
mrZ	5916	xgz2r	Yes
mxyQ	214z4	9wp8d	No
pqzK	231w2	tcrp9	Yes
q4X2	675z9	jxdme	Yes
q8xv	mkk9	jdb7q	No
q8xv	23g12	ub4c6	Yes
q8xv	g4ym	eqyhz	Yes
q8xv	2w97z	hrmsx	No
q8xv	21yg2	97eq4	No
qNvQ	69516	597r2	No
qQ9Z	28z92	8zxjq	Yes
qXX2	yy60	4rjbf	Yes
qXX2	k1yz	ezhcs	No
qYr7	69y8k	vjcsa	No
qg47	21gg2	dcez6	Yes
qg47	247ww	nruap	No
qgWj	215g2	2h47w	Yes
qggQ	2g7gy	5a7qp	Yes
qzGw	69m36	f73z9	No
rWbG	6ow93	zw4xy	Yes
rjb	67116	7ea8k	Yes
rjb	y730	v8n4x	No
rym8	949y	xshrn	Yes
vaWE	2go82	6mdxr	Yes
vaWE	1m02	4n7ef	Yes
vmxO	69936	n4v2w	No
wRvv	65wm6	e48gd	Yes
wRvv	312k	jp7tr	No
xGGO	214w4	ez53u	No
xYbO	67442	dxmrb	Yes
xYbO	k0z	q75tz	No
y2DG	2y44m	qgvzh	No
yAPR	y436	kup8x	Yes
yJwG	2kgy8	vc8hs	Yes
yQeR	65396	sv9tm	Yes
yQeR	g4k1	yhq5d	No
yjkQ	6oww3	wnv8y	Yes
yypJ	zmg9	anfk6	Yes
yypJ	21952	mr6fs	No

yypJ	2k3w2	smuat	No
yzgG	5z18	g83kx	Yes
z0v1	y410	5ywth	Yes
z0v1	0056	u5r47	No
z4dO	6m17	kzpf8	Yes
zK2	245w6	w8n3s	Yes
zK2	5698	yjghe	No
zV1O	9yzy	wb524	Yes
zb3Y	67539	pukd8	Yes
zlBL	7515	tqpb5	Yes
zlm2	675wo	y47pr	Yes
zlw	zg66	yujtc	Yes
zmYY	g5m	a6ksz	No
zmYY	2w9mo	3j7gq	Yes
zqwm	2w7w2	z7sjh	Yes

Audit of reproduction analyses and outcomes and preparation for public release

The audit and revisions process consisted of multiple parts. Project coordinators reviewed the final reports on each OSF project to check if the outputs matched the reported outcomes in the dataset. If the report did not match or was missing outcomes, then the auditor would check the output from the code of the project. If there continued to be a discrepancy, then the issue was flagged for further review. In addition, coordinators completed checks of code to ensure the code contained within each OSF project ran without error using the data that the lab provided.

We also audited the final reports and output of each reproduction with project team members that were not involved in conducting that reproduction. These auditors reviewed values in the final report and output to check if they matched the dataset. These auditors also completed specific claim reproductions. The majority of reproduction outcomes received a computational reproduction check, with priority given to reproduction analyses that were *not* conducted in R (since that was the language used for the reproduction checks). After that process, another group of auditors conducted a final review on the data and the output in the report and code within each OSF project. They also provided a holistic assessment of the reproduction analysis and provided their feedback. Project coordinators resolved any open issues themselves or in coordination with the analysts.

Following the initial submission of this manuscript and before public release of the data, we conducted an audit of the OSF projects housing the reproduction outcomes to verify completeness and appropriateness of shared information. This audit included an internal check for sensitive or proprietary materials and email correspondence with each lab. Each lab received a checklist that contained a step-by-step guide to check that their OSF project was ready to be made public. Steps within the checklist included: a check for the presence of the final report, removal of the PDF of the original paper, a check for other proprietary or sensitive materials, and steps for documenting and citing original

data sources. When labs completed the checklist, they emailed the coordinating team to confirm.

A communal tracking sheet was available to all labs as they marked completion of their checklist and observed others' completing their own. Coordinators confirmed all labs completed their checklist. This included spot-checks of the content of projects to confirm labs' reports of handling sensitive and proprietary materials properly.

For projects that were started but never completed, the coordinators followed a similar process with the individual labs and provided extra support for checking for sensitive or proprietary materials. Coordinators conducted further review for any cases in which a completed checklist was not returned.

Attrition of reproductions that were started but not completed

A reproduction attempt was defined as starting upon initiating a preregistration draft or preparing or posting content for the reproduction in its OSF project. Note that 10 papers with completed reproductions occurred following the replication workflow reported in Tyner et al (2026) because the data was available as part of gathering and conducting a replication study. These reproductions do not conform to the definition of starting an attempt and are removed from consideration of attrition. 148 of 155 (95.5%) of papers with started reproductions were completed.

Results

Claims-level Summary of Reproducibility

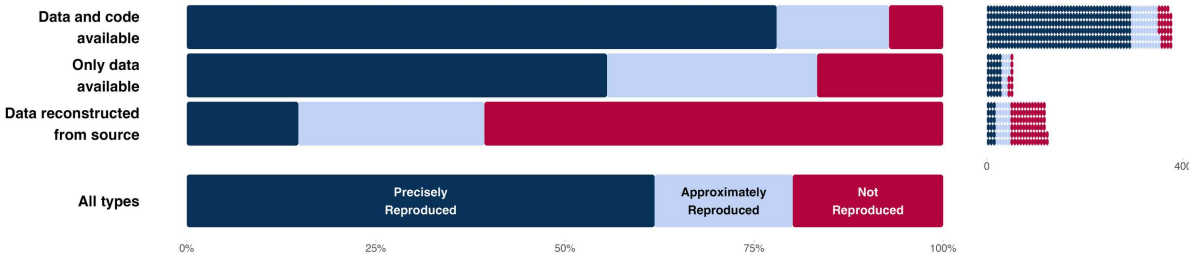
In the main text, we focused on paper-level reporting of reproducibility. If there were multiple claims reproduced in a paper, they were weighted so that each paper contributed equally to the overall findings. Here we report the claims-level outcomes for comprehensiveness. Note that in this analysis, multiple claims from the same paper are treated equally, so that papers with more claims have more impact on the overall results. It is possible that claims from the same paper are functionally independent if they are tested with different variables or data, but they are inevitably interdependent in that they came from the same authors and project. In this section, we duplicate some of the text and figures from the reproducibility results in the main text, but now report the data at the levels of claims.

Of the 553 claims that were assessed for reproducibility, we observed approximate or precise reproducibility for 443 claims (80.1% [95% CI 76.6 - 83.2%]) and precise reproducibility for 342 claims (61.8% [95% CI 57.7 - 65.8%]).

Figure S1 shows reproducibility results separately for different circumstances of conducting the reproduction. When code and data were available, we attempted to execute the original code or adapt it if necessary. We observed approximate or precise reproducibility for 350 of the 377 claims (88.1% [95% CI 81.4 - 94.6%]) and precise reproducibility for 294 of the 377 claims (78.0%). For 251 (66.6%) of these claims, we

were able to reproduce the findings with minimal effort other than executing the code on the data, a high standard known as *push button reproducibility*.

Figure S1. Reproducibility by whether data and code were available, only data were available, or when the paper's data were reconstructed from available source data for all claims



Caption: Reproducibility success rates as a percentage of attempts (left), and reproducibility success rates as counts (right).

When only data were available, we attempted to reproduce the findings by generating new code following the analyses described in the paper. Of these, we observed approximate or precise reproducibility for 45 of the 54 claims (83.3%) and precise reproducibility for 30 of the 54 claims (55.6%).

When author-provided data were unavailable, but source data were available, we attempted to reproduce the findings by preparing the data and generating new code. Of these, we observed approximate or precise reproducibility for 48 of the 122 claims (39.3%) and precise reproducibility for 18 of the 122 claims (14.8%). In summary, reproducibility rates were comparatively high when data and code were both available, and comparatively low when needing to reconstruct the data and code.

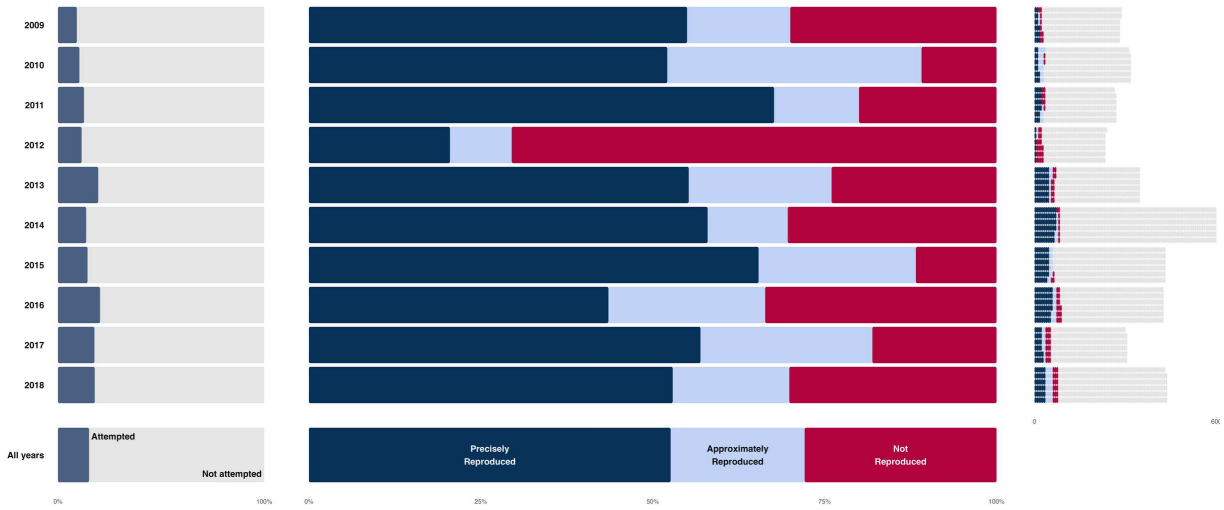
We also consider *full reproducibility* as the rate of both obtaining data and successfully reproducing the outcomes for all papers investigated. SCORE's sampling procedures for assessing data availability and reproducibility were partly but not completely dependent on each other, so a direct measure combining the two to estimate full reproducibility is not possible. However, we can approximate full reproducibility by multiplying our estimates of data availability (i.e. the proportion of the literature which were implied to be reproducibility-assessable) with our estimate of reproducibility among those that were assessed.

Across papers, 24.3% [95% CI 21.1 - 27.9%] had data available and were therefore assessable for reproducibility. Among those assessed, 72.1% [95% CI 64.5 - 79.0%] of papers were approximately or precisely reproduced, and 52.6% [95% CI 44.7 - 59.9%] of papers were precisely reproduced. Together, this implies that 17.5% [95% CI 15.1 - 20.1%] of papers were approximately or precisely reproducible, and 12.8% [95% CI 10.7 - 15.1%] of papers were precisely reproducible in this sample. Full reproducibility is a minimum estimate as it can only increase with additional effort to obtain author data, reconstruct datasets from original sources, or troubleshoot reanalysis challenges.

Across claims, 24.3% [95% CI 21.1 - 27.9%] proportion of our sample had data available and were therefore assessable for reproducibility. Among those assessed for reproducibility, 80.1% [95% CI 76.6 - 83.2%] of claims were observed to have approximate or precise reproducibility, while 61.8% [95% CI 57.7 - 65.8%] of claims were able to be precisely reproduced. Together, this implies that 19.5% of claims were approximately or precisely reproducible, while 15.0% of claims were precisely reproducible in this sample. Our estimate of full reproducibility is a lower bound that can only increase with additional effort to obtain author data, reconstruct datasets from original sources, or troubleshoot reanalysis challenges.

Figure S2 presents reproducibility success by year. For some years, the number of outcome reproduction attempts is small. Considering only claims with an attempt, the prevalence of precise reproducibility was not significantly associated with time (Spearman's ρ = -0.040 [95% CI -0.127 - 0.046]). The prevalence of approximate or precise reproducibility was also not significantly associated with time (Spearman's ρ = 0.007 [95% CI -0.082 - 0.101]). These findings are consistent with the results across papers as reported in the main text.

Figure S2. Reproducibility by year of publication for all claims

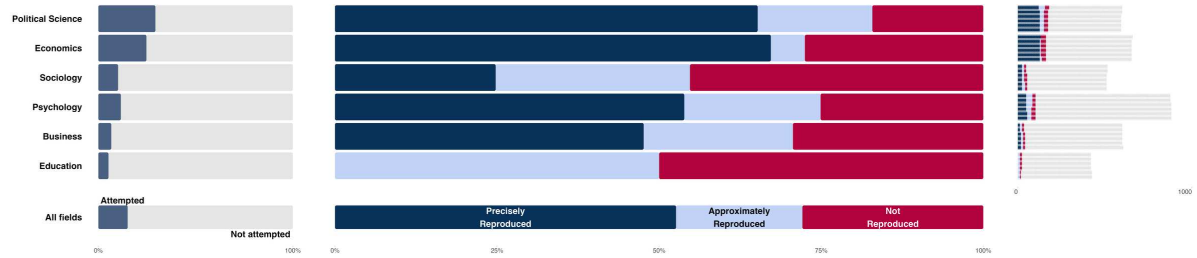


Caption: The left column illustrates the proportion of outcome reproduction attempts from the sample of claims. The middle column illustrates reproducibility as a percentage of the attempts. The right column illustrates reproducibility as counts compared with the sample of claims.

Figure S3 presents reproducibility by discipline. Political Science and Economics had much higher rates of reproduction attempts than other fields due to greater data availability. Considering only claims with a reproduction attempt, we observed approximate or precise reproducibility for 151 of 175 (86.3%) Political Science claims and 133 of 162 (82.1%) Economics claims. We observed precise reproducibility for 125 of 175 (71.4%) Political Science claims and 127 of 162 (78.4%) Economics claims. Combining the data across the other four disciplines, we observed approximate or precise reproducibility for 159 of 216 (73.6%) claims and precise reproducibility for 90 of

216 (41.7%) claims. These findings are consistent with the results across papers as reported in the main text.

Figure S3. Reproducibility by discipline for all claims

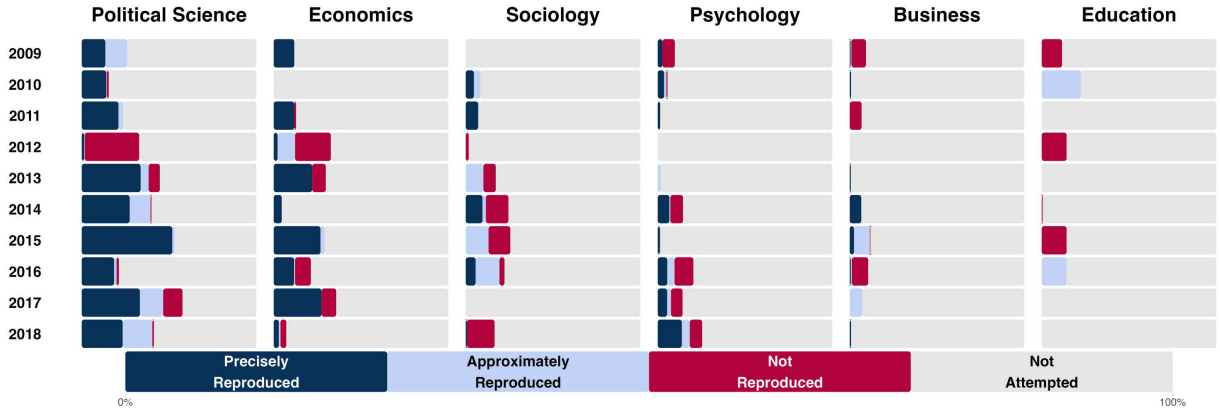


Caption: The left column illustrates the proportion of outcome reproduction attempts from the sample of claims. The middle column illustrates reproducibility as a percentage of the attempts. The right column illustrates reproducibility as counts compared with the sample of claims.

Reproducibility by discipline and year

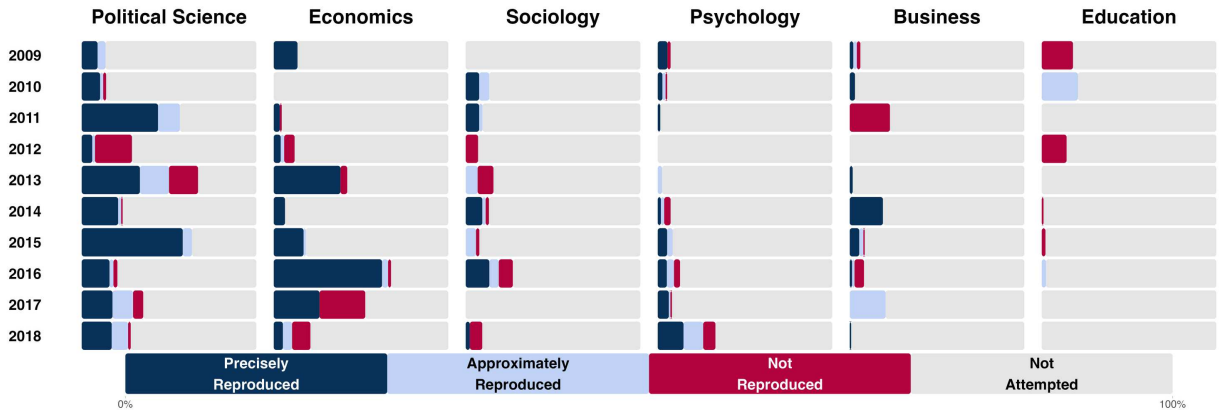
Figure S4 presents data availability and reproducibility attempts by discipline and by year across papers. Figure S5 provides the same information across claims. These figures provide a visualization of the descriptive findings that more data were available to assess reproducibility in more recent years and in Political Science and Economics compared with other fields (more color than gray), and that reproducibility success tended to be higher in Political Science and Economics than other fields (more light and dark blue than red).

Figure S4. Reproducibility by discipline and year by paper as a proportion of the sample



Caption: Reproducibility as a percentage of the sample of papers from each year and each discipline.

Figure S5. Reproducibility by discipline and year by claim as a proportion of the sample



Caption: Reproducibility as a percentage of the sample of claims from each year and each discipline.

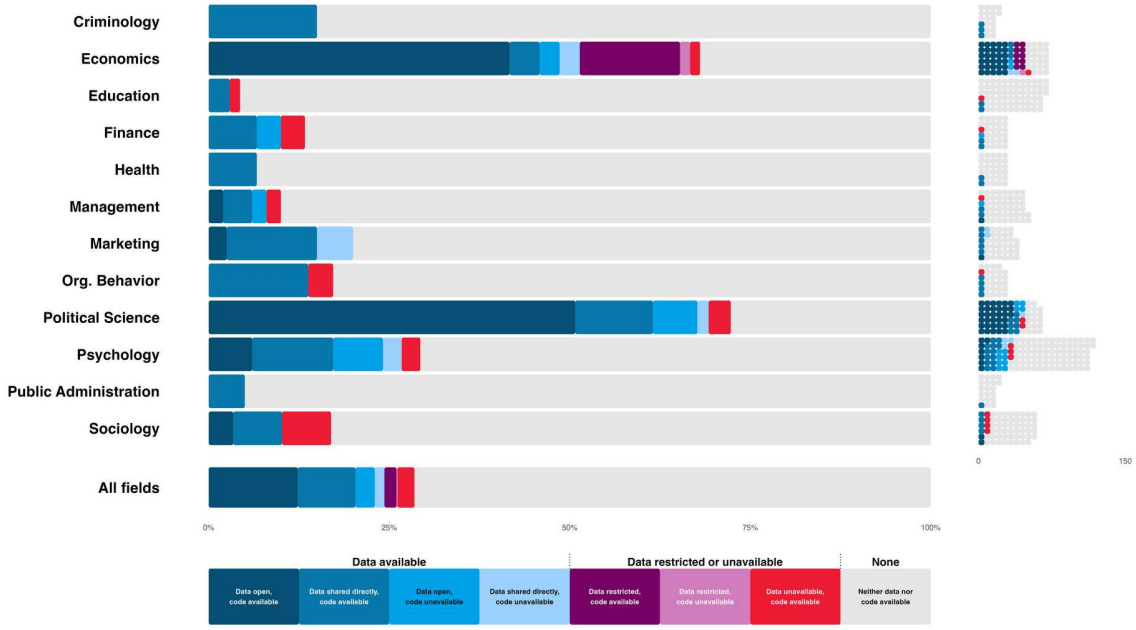
Reproduction Outcomes by Subfield

The selection of journals was done considering representation from 12 subdisciplines that were aggregated to 6 disciplines for expository purposes. Social-behavioral subdisciplines have fuzzy boundaries, and journals do not necessarily abide by those boundaries in the content that they publish. Nevertheless, the selection of journals was based on considering nominations of journals that were representative of these subdisciplines to ensure diverse representation across subdisciplines. Here, we summarize the primary reproduction outcomes separating the 62 journals into their originally identified subdiscipline.

An obvious caution is that the sample sizes for some of these subsets are small leading to highly imprecise results. There is not a strong basis for interpreting variation across subdisciplines as indicative of meaningful differences in reproduction rates.

Figure S6 illustrates that higher data availability is even more pronounced for Political Science and Economics after separating them from Public Administration and Finance, respectively. Likewise, Psychology's data availability is somewhat stronger after separating it from Health.

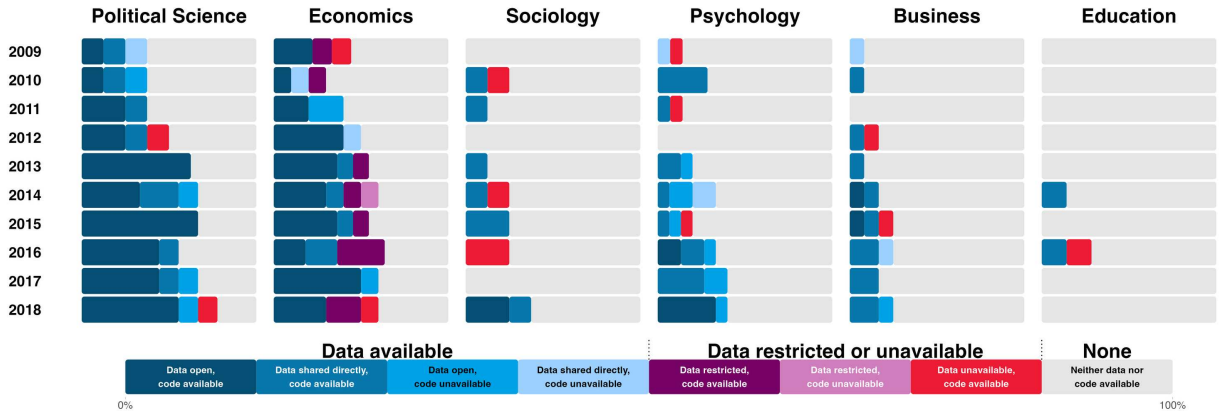
Figure S6. Data availability rates by 12 subdisciplines



Caption: The left panel shows data and code availability as a percentage of papers; the right panel shows raw counts of papers with data and code available and not available. Note that purple reflects restricted data, which did not count as available data, but might be accessible in principle.

We plotted data availability by discipline and year in Figure S7 for the 6 disciplines. Combined, Political Science and Economics were 33.4 [95% CI 18.1 - 93.0] times as likely to have open data and code (44.5% [95% CI 36.5 - 52.9%]) than the other disciplines 1.3% [95% CI 0.7 - 2.6%]). In Figure S8, we report the same outcomes separated by the 12 subdisciplines. Again, the high performance of Political Science and Economics is more pronounced after separating them from Public Administration and Finance. The Figure also highlights that all observed instances of restricted data were from Economics journals.

Figure S7. Data availability rates by year of publication for all disciplines



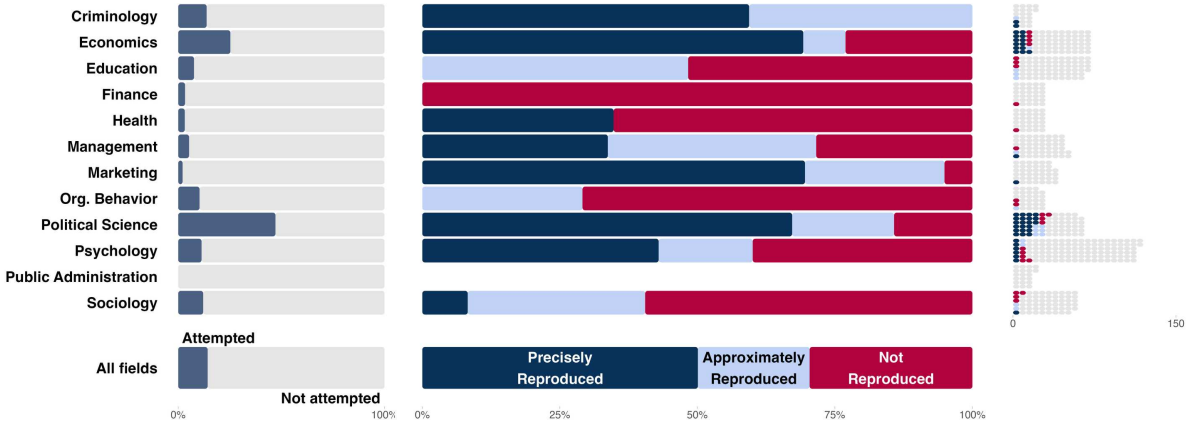
Caption: Smallest sample sizes per cell were in Education (n's from 6 to 7 per year). Note that purple reflects restricted data, which did not count as available data, but might be accessible in principle.

Figure S8. Data availability rates by year of publication for 12 subdisciplines

Caption: Smallest sample sizes per cell were in criminology and public administration, each having an n of 2 per year. Note that purple reflects restricted data, which did not count as available data, but might be accessible in principle.

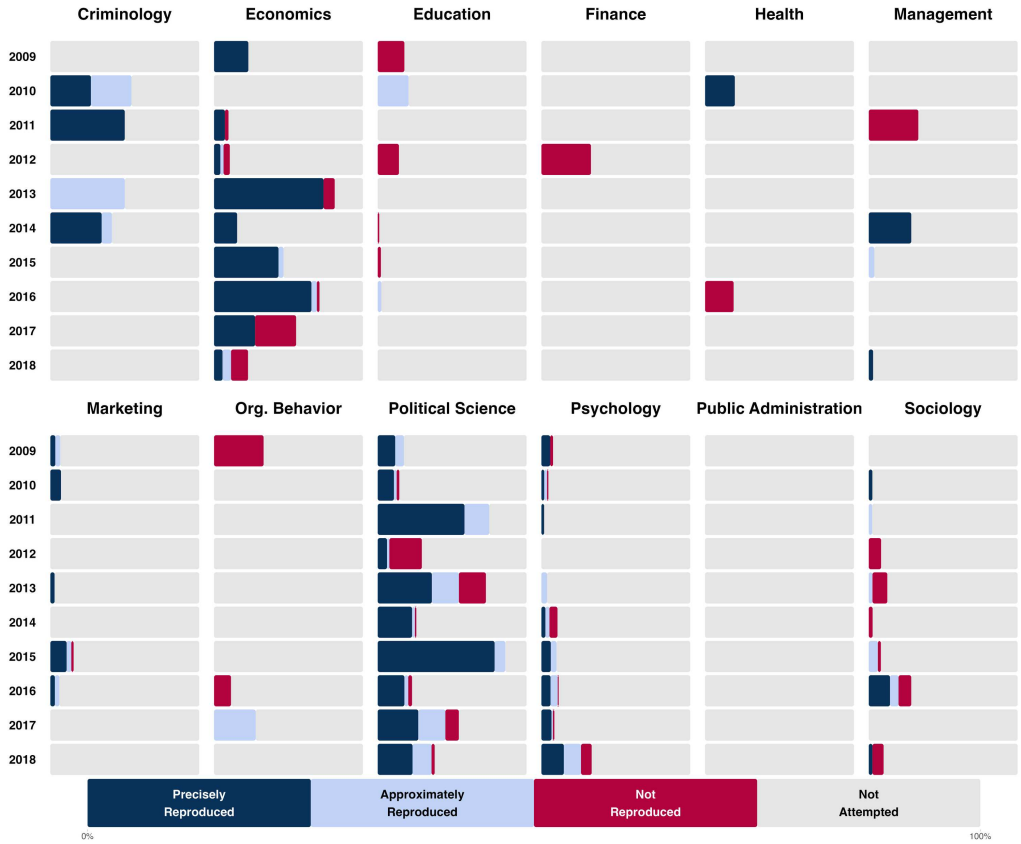
Figure S9 presents reproducibility by the 12 subdisciplines. It complements Figure 5 from the main text that showed the same data across 6 disciplines. Figure S10 separates those same outcomes by year. The higher performance for Political Science and Economics after separating the subdisciplines is less dramatic than for data availability because no Public Administration and few Finance papers were subjected to outcome reproduction attempts because of lack of data availability.

Figure S9. Reproducibility by 12 subdisciplines



Caption: The left column illustrates the proportion of outcome reproduction attempts from the sample of papers. The middle column illustrates reproducibility as a percentage of the attempts. The right column illustrates reproducibility as counts compared with the sample of papers.

Figure S10. Reproducibility by 12 subdisciplines and by year



Caption: Reproducibility as a percentage of the sample of papers from each year and each discipline.

Reproducibility assessments in comparison to the whole sample by year of publication

Table S3 presents the distribution of papers across research progress milestones by year. As with the discipline comparison in the main text, the requirement that data needed to be available produced the most substantial divergence from representativeness. In general, papers from more recent years became a larger proportion of the sample. Papers published from 2014–2018 were 51.2% of papers eligible for reproduction and were 63.0% of the sample with data available. Papers published from 2009–2013 were 48.8% of papers eligible for reproduction and were 37.0% of the sample with data available. Subsequent steps of initiating and completing reproducibility assessments produced less variation in representation by year.

Table S3. Number of papers at each stage of the selection process and number and percentage of papers and claims reproduced by year that the paper was published.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
						n (%)					
Papers with claims	287 (9.6%)	297 (9.9%)	292 (9.7%)	295 (9.8%)	305 (10.2%)	304 (10.1%)	309 (10.3%)	305 (10.2%)	305 (10.2%)	301 (10.0%)	3000 (100%)
Papers eligible for reproduction	56 (9.3%)	59 (9.8%)	58 (9.7%)	59 (9.8%)	61 (10.2%)	61 (10.2%)	62 (10.3%)	62 (10.3%)	61 (10.2%)	61 (10.2%)	600 (100%)
Papers with multiple claims	17 (8.5%)	18 (9.0%)	16 (8.0%)	13 (6.5%)	20 (10.0%)	26 (13.0%)	22 (11.0%)	25 (12.5%)	18 (9.0%)	25 (12.5%)	200 (100%)
Papers with single claim	39 (9.8%)	41 (10.2%)	42 (10.5%)	46 (11.5%)	41 (10.2%)	35 (8.8%)	40 (10.0%)	37 (9.2%)	43 (10.8%)	36 (9.0%)	400 (100%)
Papers with source or author data available	9 (4.9%)	15 (8.2%)	11 (6.0%)	13 (7.1%)	18 (9.8%)	25 (13.7%)	23 (12.6%)	23 (12.6%)	24 (13.1%)	22 (12.0%)	183 (100%)
Papers with reproduction started	10 (6.1%)	14 (8.5%)	10 (6.1%)	14 (8.5%)	15 (9.1%)	20 (12.1%)	19 (11.5%)	20 (12.1%)	21 (12.7%)	22 (13.3%)	165 (100%)
Papers with reproduction completed	10 (6.8%)	12 (8.1%)	10 (6.8%)	13 (8.8%)	13 (8.8%)	15 (10.1%)	18 (12.2%)	20 (13.5%)	16 (10.8%)	21 (14.2%)	148 (100%)
Total reproductions of claims	30 (4.8%)	43 (6.9%)	36 (5.8%)	28 (4.5%)	73 (11.7%)	93 (14.9%)	67 (10.8%)	103 (16.5%)	61 (9.8%)	89 (14.3%)	623 (100%)
Reproductions of unique claims	26 (4.7%)	34 (6.1%)	34 (6.1%)	27 (4.8%)	68 (12.2%)	85 (15.3%)	62 (11.1%)	88 (15.8%)	54 (9.7%)	79 (14.2%)	557 (100%)

Table S4. Data availability for papers by Journal

	Data open, code available	Data shared directly, code available	Data open, code unavailable	Data shared directly, code unavailable	Data restricted, code available	Data restricted, code unavailable	Data unavailable , code available	Neither data nor code available	Total
Business									
Academy of Management Journal	0	1	0	0	0	0	1	8	10
Journal of Business Research	0	0	0	0	0	0	0	10	10
Journal of Consumer Research	0	2	0	2	0	0	0	6	10
Journal of Management	0	0	0	0	0	0	0	10	10
Journal of Marketing	1	0	0	0	0	0	0	9	10
Journal of Marketing Research	0	3	0	0	0	0	0	7	10
Journal of Organizational Behavior	0	2	0	0	0	0	1	7	10
Journal of the Academy of Marketing Science	0	0	0	0	0	0	0	10	10
Management Science	1	1	1	0	0	0	0	7	10
Organization Science	0	0	0	0	0	0	0	9	9
Organizational Behavior and Human Decision Processes	0	2	0	0	0	0	0	8	10
The Leadership Quarterly	0	0	0	0	0	0	0	10	10
Economics									
American Economic Journal: Applied Economics	5	0	0	0	4	0	0	0	9
American Economic Review	6	0	1	0	3	0	0	0	10
Econometrica	4	0	1	0	0	0	0	0	5
Experimental Economics	1	2	0	1	0	0	0	6	10
Journal of Financial Economics	0	0	1	0	0	0	0	9	10
Journal of Labor Economics	5	0	0	1	2	1	1	0	10
Journal of Political Economy	8	0	0	0	0	0	0	1	9
Review of Financial Studies	0	1	0	0	0	0	0	9	10
The Journal of Finance	0	1	0	0	0	0	1	8	10
The Quarterly Journal of Economics	1	1	0	0	1	0	0	7	10
World Development	0	0	0	0	0	0	0	9	9
Education									
American Educational Research Journal	0	0	0	0	0	0	0	10	10
Computers & Education	0	0	0	0	0	0	0	10	10
Contemporary Educational Psychology	0	0	0	0	0	0	0	10	10
Educational Researcher	0	0	0	0	0	0	1	8	9
Exceptional Children	0	0	0	0	0	0	0	10	10
Journal of Educational Psychology	0	2	0	0	0	0	0	8	10
Learning and Instruction	0	0	0	0	0	0	0	10	10
Political Science									
American Journal of Political Science	7	0	0	0	0	0	0	3	10
American Political Science Review	5	1	0	0	0	0	1	3	10
British Journal of Political Science	4	1	2	0	0	0	0	3	10
Comparative Political Studies	2	3	1	1	0	0	1	2	10
Journal of Conflict Resolution	9	0	1	0	0	0	0	0	10
Journal of Experimental Political Science	3	1	0	0	0	0	0	1	5
Journal of Public Administration Research and Theory	0	1	0	0	0	0	0	9	10
Public Administration Review	0	0	0	0	0	0	0	10	10
World Politics	3	1	0	0	0	0	0	6	10
Psychology									
Child Development	0	2	0	0	0	0	0	8	10
Clinical Psychological Science	1	1	0	0	0	0	0	4	6
Cognition	0	0	4	0	0	0	0	6	10
European Journal of Personality	1	2	0	0	0	0	1	6	10
Evolution and Human Behavior	0	0	3	0	0	0	0	7	10
Health Psychology	0	1	0	0	0	0	0	9	10
Journal of Applied Psychology	0	1	0	0	0	0	0	9	10
Journal of Consulting and Clinical Psychology	0	0	0	0	0	0	0	10	10
Journal of Environmental Psychology	0	0	0	1	0	0	0	9	10
Journal of Experimental Psychology: General	1	1	0	1	0	0	0	7	10
Journal of Experimental Social Psychology	2	3	1	0	0	0	1	3	10
Journal of Personality and Social Psychology	1	1	0	0	0	0	0	8	10
Psychological Medicine	0	0	0	0	0	0	0	10	10

Psychological Science	1	2	0	1	0	0	1	5	10
Social Science & Medicine	0	1	0	0	0	0	0	9	10
American Journal of Sociology	1	1	0	0	0	0	1	7	10
American Sociological Review	1	1	0	0	0	0	1	7	10
Sociology									
Criminology	0	3	0	0	0	0	0	7	10
Demography	0	1	0	0	0	0	0	8	9
European Sociological Review	0	0	0	0	0	0	2	8	10
Journal of Marriage and Family	0	0	0	0	0	0	0	10	10
Law and Human Behavior	0	0	0	0	0	0	0	10	10
Social Forces	0	1	0	0	0	0	0	9	10

Reproduction outcomes by journal

Here we provide process (Table S4) and outcome (Table S5) reproducibility results for papers by journal. Sample sizes are too small for generating confident inferences about variation across journals. Nevertheless, these data may be useful for generating hypotheses or exploring potential associations between journal policies and reproducibility outcomes across the sample.

Table S5. Reproducibility for papers by journal

	Not attempted	Not Reproduced	Approximately Reproduced	Precisely Reproduced	Total
Business					
Academy of Management Journal	9	0	0	1	10
Journal of Business Research	10	0	0	0	10
Journal of Consumer Research	7	0	2	1	10
Journal of Management	9	0	1	0	10
Journal of Marketing	9	1	0	0	10
Journal of Marketing Research	8	0	0	2	10
Journal of Organizational Behavior	8	1	1	0	10
Journal of the Academy of Marketing Science	10	0	0	0	10
Management Science	9	0	0	1	10
Organization Science	8	1	0	0	9
Organizational Behavior and Human Decision Processes	9	1	0	0	10
The Leadership Quarterly	9	1	0	0	10
Economics					
American Economic Journal: Applied Economics	4	1	1	3	9
American Economic Review	6	1	0	3	10
Econometrica	0	2	0	3	5
Experimental Economics	7	1	0	2	10
Journal of Financial Economics	9	1	0	0	10
Journal of Labor Economics	5	2	1	2	10
Journal of Political Economy	1	1	1	6	9
Review of Financial Studies	10	0	0	0	10
The Journal of Finance	10	0	0	0	10
The Quarterly Journal of Economics	10	0	0	0	10
World Development	7	2	0	0	9
Education					
American Educational Research Journal	8	1	1	0	10
Computers & Education	10	0	0	0	10
Contemporary Educational Psychology	10	0	0	0	10
Educational Researcher	8	0	1	0	9
Exceptional Children	10	0	0	0	10
Journal of Educational Psychology	7	1	2	0	10
Learning and Instruction	8	2	0	0	10
Political Science					
American Journal of Political Science	4	1	0	5	10
American Political Science Review	4	3	2	1	10
British Journal of Political Science	2	3	2	3	10
Comparative Political Studies	6	1	0	3	10
Journal of Conflict Resolution	0	1	2	7	10
Journal of Experimental Political Science	2	1	1	1	5
Journal of Public Administration Research and Theory	10	0	0	0	10
Public Administration Review	10	0	0	0	10
World Politics	5	2	2	1	10
Psychology					
Child Development	9	0	1	0	10
Clinical Psychological Science	5	0	0	1	6
Cognition	6	2	1	1	10
European Journal of Personality	8	1	0	1	10
Evolution and Human Behavior	8	0	0	2	10
Health Psychology	9	0	0	1	10
Journal of Applied Psychology	9	0	1	0	10

Journal of Consulting and Clinical Psychology	9	1	0	0	10
Journal of Environmental Psychology	9	1	0	0	10
Journal of Experimental Psychology: General	8	0	2	0	10
Journal of Experimental Social Psychology	5	2	1	2	10
Journal of Personality and Social Psychology	9	0	1	0	10
Psychological Medicine	10	0	0	0	10
Psychological Science	5	2	0	3	10
Social Science & Medicine	9	1	0	0	10
American Journal of Sociology	8	1	1	0	10
American Sociological Review	7	1	1	1	10
Sociology					
Criminology	6	0	3	1	10
Demography	6	1	1	1	9
European Sociological Review	7	2	1	0	10
Journal of Marriage and Family	7	3	0	0	10
Law and Human Behavior	10	0	0	0	10
Social Forces	8	2	0	0	10

Relationship between journal policies and reproducibility

This section reports the details of an exploratory investigation into the effect of journal policies on reproducibility as a possible explanation for variation in reproducibility across disciplines. A summary of this investigation is reported in the main text.

Background

In the SCORE program, we conducted hundreds of reproduction tests of findings from papers from 62 journals in the social and behavioral sciences. We observed substantial variation in reproducibility success rates across disciplines. A plausible explanation for that variation is the fact that policies for data and code sharing and conducting reproducibility checks vary across journals, with some fields -— notably Economics and Political Science — having adopted such policies at higher rates than other disciplines.

With the help of the TOP database (https://cos.io/top), we had documentation of relevant current policies, but not the history of policy adoption across the 62 journals. To more confidently assess the correlation between journal policies and reproducibility rates, we needed to create a dataset with the date of policy adoption of data sharing, code sharing, and reproducibility checks for the journals that have adopted such policies. With this dataset, we could conduct exploratory analyses examining the association of journal policies with observed reproducibility rates across the sampled papers published from 2009 to 2018.

Approach

The goal of the data collection was to create a complete, accurate history of journal policies of the SCORE journals related to data sharing, code sharing, and conducting

reproducibility checks. The goal of the analysis with the data was to assess the relationship between journal policies and likelihood of reproducibility success.

Practical considerations included the following:

- Initial assessments of available historical data — such as the Wayback Machine for reviewing journal websites — suggested that the record is incomplete. We determined that we would need to conduct outreach to journal staff and editors for their assistance in constructing a historical record.
- Journal staff are busy. We needed to keep the request for information as simple as possible to maximize the response rate and minimize burden.
- It is very easy to make the assessment of journal policies very complex. We stayed focused on a few high-priority assessments and pointed the way for future investigations to gather additional evidence that could illuminate more detail about policy particulars and their impact.
- Transparent documentation of the limitations of the data would help us document the limitations of the data analysis and interpretation and facilitate additional analyses of these data by others post publication.

Method

Measures

We measured the following with minor variations across data-sourcing teams:

- Does the journal have a policy REQUIRING data sharing? [Yes/No]
    - If yes, in what year did papers first appear in print that were subject to the policy? [year]
    - Is the year response provided with certainty, a high confidence estimate, or a low confidence estimate? [Certain, high confidence, low confidence]
    - What is the source of data sharing policy information? [open-ended]
- Does the journal have a policy for independently verifying whether data sharing occurred? [Yes/No]
    - If yes, in what year did papers first appear in print that were subject to the policy? [year]
    - Is the year response provided with certainty, a high confidence estimate, or a low confidence estimate? [Certain, high confidence, low confidence]
    - What is the source of verifying data sharing policy information? [open-ended]
- Does the journal have a policy REQUIRING code sharing? [Yes/No]

    - If yes, in what year did papers first appear in print that were subject to the policy? [year]
    - Is the year response provided with certainty, a high confidence estimate, or a low confidence estimate? [Certain, high confidence, low confidence]
    - What is the source of code sharing policy information? [open-ended]
- Does the journal have a policy for independently verifying whether code sharing occurred? [Yes/No]
    - If yes, in what year did papers first appear in print that were subject to the policy? [year]
    - Is the year response provided with certainty, a high confidence estimate, or a low confidence estimate? [Certain, high confidence, low confidence]
    - What is the source of verifying code sharing policy information? [open-ended]
- Does the journal have a policy to conduct independent reproducibility checks of reported results before publication? [Yes/No]
    - If yes, in what year did papers first appear in print that were subject to the policy? [year]
    - Is the year response provided with certainty, a high confidence estimate, or a low confidence estimate? [Certain, high confidence, low confidence]
    - What is the source of independent reproducibility checks policy information? [open-ended]
- Flag and explain if the journal may have had one or more of these policies in the past but removed them. [open-ended]
- Provide any context needed that might qualify understanding of individual ratings. [open-ended]

Deliberate constraints:

- We ignored policies that only **recommended** or **encouraged** data and code sharing, but did not require it.
- We ignored variation in the nature of the requirements such as how exceptions were handled. The only criterion of interest for this investigation is whether data sharing was **required**.
- We aimed for precision to the correct **year** of the policy implementation.

Data Sourcing

We pursued three data-sourcing methods in parallel: [1] Literature review of existing papers that documented journal policies, [2] surveying journal staff and editors, and [3] TOP database and internet search. Parallel data sourcing had the benefits of maximizing coverage and enabling cross-checking of sources for accuracy. These benefits outweighed, and partially address, the costs of combining data across potentially distinct coding methods.

- **Literature review**: It is conceivable that other researchers have done similar policy reviews and we could rely on their records.
    - Advantages: Easier fit with standard evidence synthesis methods
    - Disadvantages: We defer to the other authors' accuracy. It is unlikely that other reviews answer precisely the same questions that we are trying to answer. It is unlikely that other reviews cover all the journals that we need to cover.
- **Outreach to journals**: A very simple survey to journal staff and/or editors.
    - Advantages: Standardized instrument. The publishers maintain the journal websites and so update with new policies, and many publishers presumably maintain records of their own journals' policies.
    - Disadvantages: They might not know and have to collect second hand information. They might take a very long time to respond. Journals from large publishers have often changed publishers one or more times and records may be incomplete. Journal archives may be considered confidential.
- **TOP Factor database and Internet search**: Current journal policies are documented in the TOP Factor database. The Wayback Machine and possibly other historical sources to review journal websites for their stated policies.
    - Advantages: TOP is well-coded for our purposes. Search methods rely on public records of policies.
    - Disadvantages: Limited historical data and incomplete journal coverage. Lots of missing historical data.

Procedure

Three teams independently constructed as much of the dataset as possible using only their assigned sources, but with a similar coding rubric and documentation process to maximize consistency, efficiency, and transparency. A static version of the dataset and working sheets for historical purposes is available at: https://osf.io/axh2b. We had initially planned to code policies for requiring data and code verification, but observed that coding the existence of such policies was difficult, so we abandoned coding data

and code verification policies. Some coding of those is available from individual teams in the working sheets.

Team 1 (Literature Review) searched for published papers that coded journal policies from our sample of journals. They translated codings from those sources into our dataset and documented the source and any potential qualifiers for interpretation of the source and coding rubric.

Team 2 (Outreach to Journals) constructed a very brief survey for journal staff to clarify the history of their journal policies. They conducted personal outreach to those staff and editors with an objective of a 100% response rate.

Team 3 (TOP and Internet search) used all available information from the TOP Factor database and internet searches, such as the Wayback Machine and journal editorials about new policies, to complete the dataset. Some of this work had already been completed and was reported in an earlier draft of this paper. Also, some of the effort to determine the year of policy implementation with the Wayback Machine was attempted previously, with limited success.

The shared spreadsheet with separate sheets for each team supported oversight of progress and identification of gaps or inconsistencies. Periodic check-ins among contributors were performed to troubleshoot problems or calibrate coding decisions.

The three independently constructed datasets were combined into a single dataset to maximize coverage and identify potential discrepancies in coding between data sources. Obvious errors or discrepancies were resolved collaboratively. The final dataset was used for exploratory analyses relating journal policies to reproducibility success.

Team 1: Literature Review

Literature search was done systematically, using the query:

("data shar" OR "reproduc*" OR "replicat*") AND ("journal policy" OR "journal policies") AND ("social science*")*

Below are the results from this systematic search:

Database Name	Number of Articles Found
Web of Science	9
Google Scholar	61
Scopus	9
PubMed	11

In addition to these results, we included 16 articles and 1 dataset identified outside the systematic search. These were discovered through other teams, citation tracking, or alternate search strategies. A complete list of the articles found, whether they were downloaded, and other related information can be found in this spreadsheet.

After removing duplicates (i.e., articles appearing in more than one database), we obtained a total of 75 unique articles. Each article was reviewed to determine whether it addressed data-sharing and code-sharing policies for the journals in our list, as well as the timing of any such policy implementation. Two of the articles required obtaining a spreadsheet from the Harvard Dataverse, and another article required contact with the authors to obtain all materials.

We manually entered the relevant data codes into the spreadsheet and provided a citation to the article. We increased the certainty rating when we found multiple articles that referred to the same data point. Only those articles that contained relevant data were cited. Table S6 presents the studies that provided coding information on journal policies.

Table S6. Studies used for coding journal policies

Fink, L. & Marcus, J. Replication code availability over time and across fields: Evidence from the German Socio-Economic Panel. Economic Inquiry 63, 357–386 (2025).
Brodeur, A., Cook, N. & Neisser, C. p-Hacking, Data type and Data-Sharing Policy. The Economic Journal 134, 985–1018 (2024).
Prosser, A. M. B. et al. When open data closes the door: A critical examination of the past, present and the potential future for open data guidelines in journals. British Journal of Social Psychology 62, 1635–1653 (2023).
Askarov, Z., Doucouliagos, A., Doucouliagos, H. & Stanley, T. D. The Significance of Data-Sharing Policy. Journal of the European Economic Association 21, 1191–1226 (2023).
McAuliff, B. D. et al. Further action toward valid science in Law and Human Behavior: Requiring open data, analytic code, and research materials. Law and Human Behavior 46, 395–397 (2022).
Freedland, K. E. Health Psychology adopts Transparency and Openness Promotion (TOP) Guidelines. Health Psychol 40, 227–229 (2021).
Christensen, G., Dafoe, A., Miguel, E., Moore, D. A. & Rose, A. K. A study of the impact of data sharing on article citations using journal policies as a natural experiment. PLOS ONE 14, e0225883 (2019).
Christensen, G. & Miguel, E. Transparency, Reproducibility, and the Credibility of Economics Research. Journal of Economic Literature 56, 920–980 (2018).
Höffler, J. H. Replication and Economics Journal Policies. American Economic Review 107, 52–55 (2017).
Fidler, F. et al. Metaresearch for Evaluating Reproducibility in Ecology and Evolution. BioScience 67, 282–289 (2017).
Crosas, M. et al. Replication Data for: Data policies of highly-ranked social science journals. Harvard Dataverse https://doi.org/10.7910/DVN/CZYY1N (2017).
Gleditsch, N. P. & Janz, N. Replication in International Relations. International Studies Perspectives 17, 361–366 (2016).
O'Reilly, R., Herndon, J. & O'Reilly, R. Data Sharing Policies in Social Sciences Academic Journals: Evolving Expectations of Data Sharing as a Form of Scholarly Communication. UNC Dataverse https://doi.org/10.15139/S3/12157 (2015).
Herndon, J. & O'Reilly, R. Data Sharing Policies in Social Sciences Academic Journals: Evolving Expectations of Data Sharing as a Form of Scholarly Communication. in Databrarianship: The Academic Data Librarian in Theory and Practice (eds. Kellam, L. & Thompson, K.) (American Library Association, Chicago, IL, 2015).

Team 2: Outreach to journals

We gathered contact information for journal editors and/or staff. A researcher from our collaborative team contacted the representatives from each journal personally after adapting the template email below for the particular journal. We sent up to two reminders, one from the original sender, and a second from the team's principal investigator who was cc'ed on the original email.

Email Template

Dear [recipient],

I am writing to request information about the transparency policies at *[journal name].* I am part of a team led by Brian Nosek, Executive Director of the Center for Open Science. We are in the final stages of a large-scale investigation examining how journal policies are related to research reproducibility. As part of this project, we are collecting information on relevant practices at top journals.

Could you please take a minute to answer the seven questions below? You can do that by replying to this email and either ticking the associated boxes or supplying the requisite information in the space provided. I greatly appreciate your assistance.

Question 1 (RE: Data Availability Policy)

How does the journal handle availability of **data**?
(Tick one option.)

☐ The journal requires authors to submit data, which are then hosted directly on the journal's website alongside the article.

☐ The journal requires authors to deposit data and materials in a third-party public repository (e.g., OSF, Dataverse, Dryad) and provide a link.

☐ The journal encourages but does not require authors to make data available.

☐ The journal has no stated policy regarding availability of data.

Question 2 (RE: Data Availability Policy)

When did the current policy go into effect?

(Fill in the blanks below)

YEAR = _____ (Best guess if you are unsure)

% CONFIDENCE = _____

Question 3 (RE: Code Availability Policy)

How does the journal handle availability of **code**?
(Tick one option.)

☐ The journal requires authors to submit code, which is then hosted directly on the journal's website alongside the article.

☐ The journal requires authors to deposit code in a third-party public repository (e.g., OSF, Github, Dataverse, Dryad) and provide a link.

☐ The journal encourages but does not require authors to make code available.

☐ The journal has no stated policy regarding availability of code.

Question 4 (RE: Code Availability Policy)

When did the current policy go into effect?

(Fill in the blanks below)

YEAR = _____ (Best guess if you are unsure)

% CONFIDENCE = _____

Question 5 (RE: Reproducibility Checks)

Does the journal conduct independent reproducibility checks of the reported results before publication? (Select one option.)

☐ Yes

☐ No

Question 6 (RE: Reproducibility Checks)

When did the current policy go into effect?

(Fill in the blanks below)

YEAR = _____ (Best guess if you are unsure)

% CONFIDENCE = _____

Question 7 (Open ended)

Is there anything else that would be helpful for us to know about the journal's transparency policies over time?

(Write in the space below)

It would be very helpful if we could hear a reply from you within the week.

Please let me know if I can answer any questions about this request. Thank you for your help in providing this important information.

Sincerely,

[sender]

Team 3: TOP and Internet Search

We started our work from a TOP Factor Score datasheet, which was prepared earlier by Macie Daley in 2023 as part of the activities within the SCORE project. This sheet drew data from the TOP Factor database (https://www.topfactor.org/) that included transparency policies for thousands of journals coded for their adherence to the TOP Guidelines (Nosek et al., 2015).

The dataset provided us information whether the journal required data submission from authors. First, we checked the websites of the 62 journals to verify whether the policy reported in the TOP Factor Score datasheet was changed or updated. If the journal required authors to submit data/code/replication packages, we searched the website for the information about the year when the new policy was introduced.

Typically, the information about the journal policies was provided as part of submission guidelines or author instructions. Some journals have dedicated web pages, describing

their data policy or research transparency policy. However, sometimes the information about journal editorial policies was provided in a news section. Additionally, some information about the adoption of new editorial policies was provided by academic publishers, like the Oxford University Press.

Some journals provided clear information to authors about the history and development of their editorial policies, including the dates of policy start or update. For example, the American Economic Association has kept an archive of its editorial policies dating back to 2005. The journals of the Cambridge University Press have a dedicated Research Transparency page on their websites. However, other journals provided only the then-current requirements on their website. Instead, these journals reflected on their history in dedicated editorials. These editorials often indicated the year when the journal adopted any new requirements for authors.

The editors of many political science journals in the dataset joined the Journal Editors' Transparency Statement (JETS) in 2014, effective from January 15, 2016. This statement obliged the journals to require data and analytic materials from authors. This statement also shed light on the development of changes in the journals in our study.

As a final step, if the journal website and editor’s notes did not provide a year of adopting new requirements for authors, we used the Wayback Machine to trace the changes in the journal website and identify the year when new policies were introduced.

For the majority of the journals, sufficient information was provided on the journal website or in the editorials or editor’s reports. However, the Wayback Machine was used in the case of a few journals.

Combining Data

The three teams worked independently to construct datasets using different information sources, with recognition that each source might not provide complete information, or might not provide the same information.

For the purposes of this project, we sought to create a single dataset combining the results across the data sources. We provide the data from all sources so that future investigations can examine them independently.

We followed the following process for combining data across sources:

- If data for a journal was available from only one source, we used the data from that source.
- If data for a journal was available from multiple sources and the sources were in agreement, we used the data from the source with the least missing data.
- If data for a journal was available from multiple sources and the sources were not in agreement, then we discussed which data is likely to be most reliable and documented the rationale for our selection. Sometimes, this involved additional literature or Internet searches to verify journal policies.

In Table S7, we summarize the coverage for each method and the combined data. We were able to identify policies for all journals with use of the TOP Guidelines database and internet searches, but the literature review and outreach to journal staff were helpful nonetheless, particularly for identifying dates of policy adoption and to raise discrepancies for discussion and resolution.

We provided a subjective rating for whether the final policy and year coding was made with certainty, high confidence, or low confidence. And, we explicitly labeled whether there were discrepancies between the data sources on either the policy or the year of implementation. In addition to the final dataset, all of the historical coding from individual teams, aggregation, discrepancy resolution, and coder notation is available in a series of spreadsheets.

Table S7. Percentage of 62 journals for which each data source had information.

	Data Sharing		Code Sharing		Reproducibility Checks	
Data Source	**Policy**	**Year**	**Policy**	**Year**	**Policy**	**Year**
Literature Review	45%	42%	44%	42%	5%	5%
Outreach to Journals	77%	71%	77%	73%	77%	77%
TOP & Internet Search	100%	100%	100%	100%	100%	100%
Combined Data	100%	100%	100%	100%	100%	100%

Note: If the data source confirmed that the journal did not have a policy, then the year that the policy was established was also considered confirmed (i.e., never).

Analysis

Preliminary Analysis: What combination of policies are present when there are 1 or 2 policies?

- Starting assumption: Journals will almost always follow this pattern: If one policy, it is requiring data sharing; If two policies, it is requiring data and code sharing.

If it is the case that this order is universal, or near universal, then explain it in text. If it is more complex, then there could be value in deeper exploratory analyses of the contributions of each of these policies independently and in conjunction with one another. The rest of the analysis plan is drafted on the assumption that it would be universal during the key time period (2009 to 2018).

Primary research question: Are more stringent reproducibility policies associated with greater success on reproducibility?

- The primary analysis will be at the paper level, a secondary analysis will be at the claims level. The secondary analysis will be presented in supplementary information like other claims-level analyses from the paper.

- Treatment variable: Was the paper published under a regime of 0, 1, 2, or 3 reproducibility policies (none, required data sharing, required code sharing, independent reproducibility check)?
- Outcome variable: Reproducibility (0 = not reproduced, 1 = approximately reproduced, 2 = precisely reproduced)
- Analysis: The analysis will proceed in two parts.
    - 1) A 4x3 crosstab with a chi-square test pooling the data across years for both the paper-level and the claims-level tests.
    - 2) An ordered logit model with year fixed effects, including right-hand side indicator variables for regimes 1, 2 and 3, using regime 0 as the baseline. The year fixed effects adjust for changes in reproducibility over time. We note however that we cannot interpret any significant results as a causal effect in this descriptive analysis, in that it is possible that journals with higher standards also adopt more stringent transparency policies. The claims level ordered logit analysis would replicate this setup but adding paper-level fixed or random effects given the dependence that is likely to occur within papers. The estimation method could be either frequentist MLE or Bayesian MCMC with flat priors – whichever the visualization team prefers.
- Presentation
    - Visualize the proportion of outcomes that were not, approximately, and precisely reproduced for each level of the treatment variable: 0, 1, 2, or 3 policies in place.
    - Visualize this relationship with the addition of the time (year) dimension. No formal analysis is needed of the relationship with time, the descriptive visualization provides further insight to the existing figure highlighting improvement in reproducibility over time.

Follow-up visualization: When did policies get introduced across the journals, illustrating by discipline variation and including through 2025 to offer a hypothesis about how reproducibility may be changing since 2018. This descriptive visualization does not need formal inferential tests, but will likely benefit from calculating proportions of journals with the policies at different time points in the visualization to illustrate that adoption of these policies is increasing over time--including beyond our sample period.

Secondary research question: Are more stringent reproducibility policies associated with greater data availability?

- This research question borders on trivial -- i.e., does requiring data sharing increase data sharing? However, it is useful to have the empirical evidence for two reasons:

    - The existence of a policy does not necessarily translate to the implementation of a policy
    - It is likely that implementation is imperfect. The reasons could be justified (e.g., can't implement it consistently for good reasons) or unjustified (e.g., didn't implement the policy because of laziness, disorganization, or other non-substantive reasons). This analysis will not unpack the reasons, but it will offer some evidence for future interrogation.

- Data availability is coded at the paper-level, so this is only a paper-level analysis.
- Treatment variable: Was the paper published under a regime of 0, 1, 2, or 3 reproducibility policies (required data sharing, required code sharing, independent reproducibility check)?
- Outcome variable: Data availability (0 = failed data availability check, 1 = passed data availability check)
- Analysis: The analysis will proceed in two parts. 1) A 4x2 crosstab with a chi-square test pooling the data across years at just the paper-level. 2) A logit model with year fixed effects, including right-hand side indicator variables for regimes 1, 2 and 3, using regime 0 as the baseline. The year fixed effects adjust for changes in reproducibility over time. We note however that we cannot interpret any significant results as a causal effect in this descriptive analysis, in that it is possible that journals with higher standards also adopt more stringent transparency policies. The claims level logit analysis would replicate this setup but adding paper-level fixed or random effects given the dependence that is likely to occur within papers. The estimation method could be either frequentist MLE or Bayesian MCMC with flat priors – whichever the visualization team prefers.
- Presentation
    - For space considerations, most of these analyses and visualization may appear in supplementary information.
    - Visualize the proportion of outcomes that were process reproducible for each level of the treatment variable: 0, 1, 2, or 3 policies in place.
    - Visualize this relationship with the addition of the time (year) dimension. No formal analysis is needed of the relationship with time, the descriptive visualization provides further insight to the existing figure highlighting improvement in reproducibility over time.
- Cautions
    - The definition of the treatment and outcome variables is important for understanding the meaning of the observed relationship. The text will need to contextualize this.

- If it is the case that the presence of 1 policy always means that data sharing was required, and 2 policies always meant data and code sharing, then the interpretation will be straightforward. However, if there is variation in what policies were present with 1 and 2 policies in place, then the discussion of the meaning is more complex. A simple follow-on analysis could be done that exclusively checks the relationship between the data sharing policy and whether data sharing occurred. Likewise for code sharing.
- Early indicators suggest incompleteness in our coding of whether there are independent checks of whether data sharing and code sharing occurred as required. So, this analysis plan leaves out those variables -- though they are plausibly important for a policy implementation to be successful. Future investigations will need to look more closely at those variables.

All statistical tests in this section use two tailed statistical measures.

Results

Considering policies adopted within or before the sampling frame of our investigation (2009 to 2018), policy adoption followed a predictable order. If the journal had just one policy requirement, it was data sharing. If the journal had two policy requirements, it was data and code sharing. Because of this predictability, we reported on the relationship between policy adoption and reproducibility outcomes with four levels: no policies; data sharing policy; data and code sharing policy; and data, code, and reproducibility check policies.

Notably, more recent policy adoption outside the SCORE sample time frame has mostly followed this order, but imperfectly. Three journals introduced a code-sharing requirement without a data sharing requirement (Health Psychology [2021], Journal of Experimental Psychology: General [2020 to 2022], Review of Financial Studies [2020]), and two journals introduced a reproducibility check policy without public data or code sharing requirements (Journal of Marketing [2023], Journal of Marketing Research [2023]). For visualization of journal policies from 2004 to 2025, we presented the policies independently (Figure 6 in main text).

In an earlier draft of this manuscript, we reported an exploratory analysis of current data and code sharing journal policies that prompted reviewers to suggest a deeper investigation. On entering this investigation, we were aware of the distribution of current journal policies, had general awareness of historical policies, and could reasonably infer possible relationships with reproducibility outcomes. While we developed an analysis plan before conducting the analysis, those analysis plans cannot be properly interpreted as being independent of the data. As such, these results should be cautiously interpreted as the outcomes of an exploratory investigation.

Reproducibility

The primary question was whether more stringent reproducibility policies were associated with greater success on reproducibility. The analysis was conducted at both the paper-level and claim-level and proceeded in two parts, [1] 4 (no policy; data required; data and code required; data, code, and reproducibility checks required) x 3 (not reproduced, approximately reproduced, precisely reproduced) crosstab with a chi-square test pooling the data across years, and [2] an ordered logistic regression model.

The crosstabs are presented in Table S8 across papers and Table S9 across claims. Descriptively, papers were precisely reproduced more frequently with any transparency policy (70.5%, 71.9%, and 65.0%) than with no policy (40.3%); likewise, when comparing across claims (80.9%, 77.3%, and 62.5% with a policy; 50.9% no policy). However, the chi-square test provided suggestive evidence across papers ($p = 0.023$) and significant evidence across claims ($p < .001$).

Table S8. Paper-level reproducibility by journal policy (N = 145)

	No policy	Data required	Data and code required	Data, code & repro check required	Total
Precisely reproduced	35.0 (40.3%)	9.9 (70.5%)	28.8 (71.9%)	2.6 (65.0%)	76.3 (52.6%)
Approximately reproduced	22.2 (25.6%)	1.1 (8.1%)	3.7 (9.2%)	1.2 (30.0%)	28.3 (19.5%)
Not reproduced	29.7 (34.1%)	3.0 (21.4%)	7.6 (18.9%)	0.2 (5.0%)	40.5 (27.9%)
Total	87.0 (60.0%)	14.0 (9.7%)	40.0 (27.6%)	4.0 (2.8%)	145.0 (100.0%)

Note: *First three rows present column-wise percentages. Total row (bottom row) presents row-wise percentages. Paper-level outcomes are weighted counts.*

Chi-squared test: $\chi^2(6) = 14.6, p = 0.023$

Table S9. Claims-level reproducibility by journal policy (N = 553)

	No policy	Data required	Data and code required	Data, code & repro check required	Total
Precisely reproduced	166.0 (50.9%)	38.0 (80.9%)	133.0 (77.3%)	5.0 (62.5%)	342.0 (61.8%)
Approximately reproduced	84.0 (25.8%)	3.0 (6.4%)	12.0 (7.0%)	2.0 (25.0%)	101.0 (18.3%)
Not reproduced	76.0 (23.3%)	6.0 (12.8%)	27.0 (15.7%)	1.0 (12.5%)	110.0 (19.9%)
Total	326.0 (59.0%)	47.0 (8.5%)	172.0 (31.1%)	8.0 (1.4%)	553.0 (100.0%)

Note: *First three rows present column-wise percentages. Total row (bottom row) presents row-wise percentages.*

Chi-squared test: *$\chi^2(6) = 46.4, p < .001$*

Table S10 presents the ordered logistic regression model across papers and Table S11 presents the same analysis across claims.

Across papers, the ordered logistic model in Table S10 suggests that stronger journal policies are associated with reduced risk of a Not Reproduced outcome (the baseline). The odds of Not Reproduced are 1.78 times higher than an Approximate outcome if a journal has no policy (the baseline); however, if journals have any level of policy, the odds of being in a lower reproducibility outcome category decreases by around 70% (reflecting the three odds-ratios 0.33, 0.30, and 0.33 respectively). The odds of being in a Precise or Approximate reproducible outcome category are not noticeably different from one another (95% odds-ratio CI [0.48, 1.09]), this applies specifically to being less likely to have a Not Reproduced outcome than both. The effect is most obvious for Policy level 2 with a 95%CI of [0.14, 0.67] which remains far below an odds-ratio of 1.00. It is mostly below 1.00 for Policy level 1 with a 95% CI of [0.10, 1.12], but not very reliable for Policy level 3 which has a very wide CI at [0.04, 2.42] presumably due to very few cases (See Table S8).

Table S10. Ordered Logistic Regression Models. Paper-Level Reproducibility.

Term	Coeff.	Std. Error	Odds Ratio	95%CI
Threshold: Precisely Reproduced|Approximately Reproduced	-0.33	0.21	0.72	[0.48, 1.09]
Threshold: Approximately Reproduced|Not Reproduced	0.58	0.21	1.78	[1.17, 2.71]
Policy level 1	-1.10	0.62	0.33	[0.10, 1.12]
Policy level 2	-1.19	0.40	0.30	[0.14, 0.67]
Policy level 3	-1.12	1.02	0.33	[0.04, 2.42]

Note: Weighted N=145. Each claim outcome was weighted as a fraction of all outcomes for that paper. Policy level 1 = data sharing requirement, policy level 2 = data and code sharing requirements, policy level 3 = data, code, and reproducibility check requirements.

Across claims, the ordered logistic model in Table S11 shows that stronger journal policies are associated with reduced risk of a Not Reproduced outcome (the baseline). Not Reproduced has a 3.57 times larger odds than an Approximate outcome if a journal has no policy (the baseline); however, if journals have any level of policy, these odds decrease by about 90% and this is statistically robust for journals with a Policy level 1 policy (OR = 0.07, 95% CI = [0.01, 0.72] and journals with a Policy level 2 policy (OR = 0.06, 95% CI = [0.01, 0.29]. For journals with Policy level 3 this appears to be the case (OR = 1.96 but the confidence intervals are far above and below 1.00 ([0.00, 4.41], presumably due to very low case numbers (See Table S9).

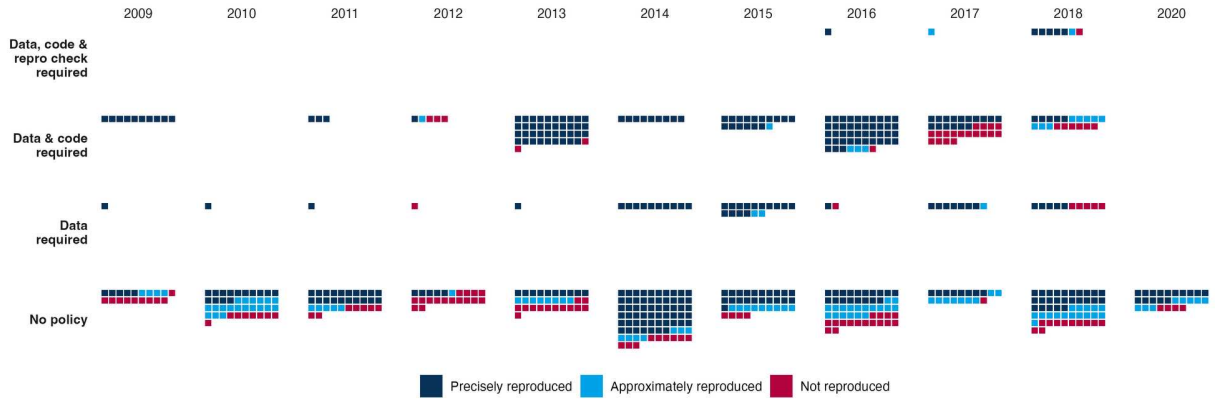
Table S11. Ordered Logistic Regression Models. Claims-Level Reproducibility.

Term	Coeff.	Std. Error	Odds Ratio	95%CI
Threshold: Precisely Reproduced|Approximately Reproduced	-0.71	0.40	0.49	[0.22, 1.07]
Threshold: Approximately Reproduced|Not Reproduced	1.27	0.40	3.57	[1.62, 7.83]
Policy level 1	-2.67	1.19	0.07	[0.01, 0.72]
Policy level 2	-2.81	0.81	0.06	[0.01, 0.29]
Policy level 3	-2.36	1.96	0.09	[0.00, 4.41]

Note: N=553. Includes random-intercepts at the paper-level (for N = 145 papers). Policy level 1 = data sharing requirement, policy level 2 = data and code sharing requirements, policy level 3 = data, code, and reproducibility check requirements.

Figure S11 provides a visualization of the relationship between journal policy and reproduction success over time across claims. More assertive policies are represented toward the top of the y-axis, and time is represented on the x-axis. 41.0% of the sample of claims were subjected to any of the three policy requirements in total, and only a small proportion of those were in the earlier years of this sample.

Figure S11. Reproducibility outcomes by journal policy and year across claims.



Caption: Data points are claims.

Also, note that there is an important dependency in these data that qualify interpretations of the relationship between journal policies and reproducibility success. For us to conduct an outcome reproduction test, we needed to have access to the study data. As a consequence, to the extent that journal policies requiring data sharing enhance data sharing, the make-up of the sample of papers and claims from journals with a data sharing policy could be distinct in a variety of ways from the make-up of the sample of papers and claims from journals without a data sharing policy. The observed exploratory evidence is suggestive that journal transparency policies are associated with greater reproducibility, but further investigation is needed to provide clear evidence of a causal relationship.

Data Availability

A secondary research question concerned whether more stringent reproducibility policies were associated with greater data availability. We conducted the same series of analyses as for data availability on the sample of 600 papers that were subjected to data availability tests. The analysis was conducted at the paper-level in two parts, [1] 4 (no policy; data required; data and code required; data, code, and reproducibility checks required) x 2 (did not succeed, succeeded) crosstab with a chi-square test pooling the data across years, and [2] an ordered logistic regression model.

The crosstabs are presented in Table S12. Descriptively, papers achieved data availability more frequently with any transparency policy (14 (87.5%), 45 (66.2%), and 5 (100.0%)) than with no policy (82 (16.0%)), with a significant chi-square test ($p < .001$). This aligns with the definition of data availability for this investigation: success more likely if data were accessible for reproduction. The pattern would likely be different if we had adopted a more stringent definition of data availability that included code.

Table S12. Data availability outcomes by journal policy (N =600)

	No policy	Data required	Data and code required	Data, code & repro check required	Total
Data available	82 (16.0%)	14 (87.5%)	45 (66.2%)	5 (100.0%)	146 (24.3%)
Not available	429 (84.0%)	2 (12.5%)	23 (33.8%)	0 (0.0%)	454 (75.7%)
Total	511 (85.2%)	16 (2.7%)	68 (11.3%)	5 (0.8%)	600 (100.0%)

Note: *First two rows present column-wise percentages. Total row (bottom row) presents row-wise percentages.*

Chi-squared test: $\chi^2(3) = 99.8, p < .001$

For data availability, a logistic regression reported in Table S13 predicts Succeeded or Did Not Succeed. For the baseline condition when a journal has no policy, Did Not Succeed has an odds that are 5.23 times higher; however, when the journal has either a Policy level 1 or Policy level 2 policy then the same unsuccessful outcome is 97% and 90% less likely respectively (ORs = 0.03 [0.00, 0.10] and 0.10 [0.06, 0.17]). Data, code and reproducibility check has no discernable effect because the Did Not Succeed outcome has 0 cases in this matrix cell (see Table S12).

Table S13. Logistic Regression Model. Data Availability.

Term	Coeff.	Std. Error	Odds Ratio	95%CI
Baseline (No policy)	1.65	0.12	5.23	[4.16, 6.67]
Policy level 1	-3.60	0.77	0.03	[0.00, 0.10]
Policy level 2	-2.33	0.28	0.10	[0.06, 0.17]
Policy level 3	-17.22	650.87	0.00	NA

Note. N= 600 Policy level 1 = data sharing requirement, policy level 2 = data and code sharing requirements, policy level 3 = data, code, and reproducibility check requirements. NA appears for the 95% CI for Policy level 3 because it is essentially undefined given small sample size and 0 instances of not reproduced.

Journals with transparency policies had much higher rates of data availability than journals with no policies. Nevertheless, success rates were not perfect for journals with such policies. Future investigations could determine the reasons for this, which could be a mixture of legitimate exceptions to the policy, coding errors for papers meeting data sharing requirements by our investigation, and failures to enact the policy by the journal and authors. Also, while it is very plausible that journals having a policy requiring data sharing caused data sharing to occur, these analyses do not directly support a causal interpretation. Our exploratory investigation provides a basis for the hypothesis that journal policies increase process and reproducibility to be investigated further in future research.

References

Abatayo, A. L., Achakulvisut, T., Acuna, D., Aczel, B., Balaji, L., Bandrowski, A. E., Benjamin, D. M. … (2026). Empirical, Human, and Machine Assessments of Research Credibility in the Social and Behavioral Sciences.

Marcoci, A., Wilkinson, D. P., Vercammen, A., Wintle, B. C., Abatayo, A. L., Baskin, E., Berkman, H., Buchanan, E. M., Capitán, S., Capitán, T., Chan, G., Cheng, K. J. G., Coupé, T., Dryhurst, S., Duan, J., Edlund, J. E., Errington, T. M., Fedor, A., Fidler, F., Field, J. G., Fox, N., Fraser, H., Freeman, A. L. J., Hanea, A., Holzmeister, F., Hong, S., Huggins, R., Huntington-Klein, N., Johannesson, M., Jones, A. M., Kapoor, H., Kerr, J., Kline Struhl, M., Kołczyńska, M., Liu, Y., Loomas, Z., Luis, B., Méndez, E., Miske, O., Mody, F., Nast, C., Nosek, B. A., Parsons, E. S., Pfeiffer, T., Reed, W. R., Roozenbeek, J., Schlyfestone, A. R., Schneider, C. R., Soh, A., Tagat, A., Tutor, M., Tyner, A., Urbanska, K., & van der Linden, S. (2024). Predicting the replicability of social and behavioural science claims in a crisis: The COVID-19 preprint replication project. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01961-1

Tyner, A., Abatayo, A. L., Daley, M., Field, S., Fox, N., Haber, N., Hahn, K., Kline Strul, M., Mawhinney, B., Miske, O., Silverstein, P., … (2026). Investigating the replicability of the social and behavioral sciences.