**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 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

# Abstract

Published claims should be *reproducible*, yielding the same result when applying the same analysis to the same data. 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. We assessed whether originally reported claims could be reproduced using the same data and analysis for papers in which authors made data available or we obtained source data to reconstruct the dataset. 76.2 (52.6% [95% CI 45.8 - 60.3%]) papers were rated as precisely reproducible and 104.4 (72.1% [95% CI 65.0 - 78.7%]) papers as at least approximately reproducible (within 15% of the original effects or within .05 of original p-values) after weighting 553 claims from 144.9 papers. 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 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 error, and suggests alternative explanations to investigate, but an underlying assumption is that research 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 produced the reported outcomes. This paper investigates how close we are to this ideal.

Investigations in Economics, Finance, Political Science, Cognitive Science, Psychology, Social Sciences, Electronic Health Records research, Ecology, and elsewhere suggest that *outcome reproducibility,*[1,2](https://www.zotero.org/google-docs/?oleor1) defined as observing the same results using the same analysis on the same data, cannot be taken for granted.[3–16](https://www.zotero.org/google-docs/?zUNbxZ) 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 depend on the accessibility of the data, and preferably the analytic code, because descriptions of analytic methods may be incomplete or difficult to translate back to code.[17](https://www.zotero.org/google-docs/?oQGFSq) If data are inaccessible to everyone, then it is not possible to independently verify that reported results were reported precisely. For this investigation, we define *process reproducibility* as sharing the author-prepared data, which is a higher expectation than availability of source data (e.g., secondary data analysis of public government data) and a lower expectation than availability of data and code to enable an independent test of outcome reproducibility. Data sharing 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.[7,11,18–22](https://www.zotero.org/google-docs/?6uHAcw) Data inaccessibility is a threat to research credibility because it interferes with verifying outcome reproducibility.[23](https://www.zotero.org/google-docs/?TKEkgU)

As part of the DARPA-funded Systemizing Confidence in Open Research and Evidence (SCORE) program,[24](https://www.zotero.org/google-docs/?B5BVNz) we conducted a systematic investigation of process and outcome reproducibility in the social and behavioral sciences.

Table 1. 62 journals included in the sampling frame 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.*

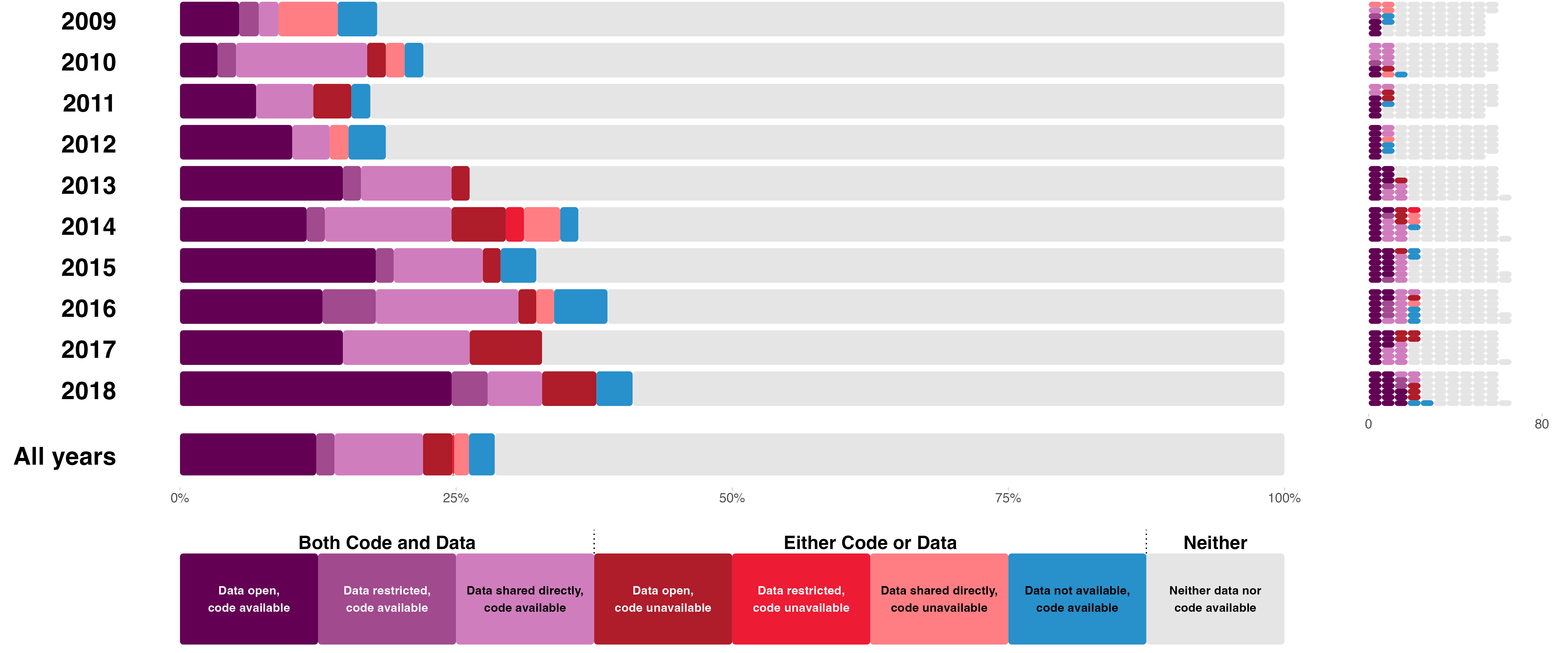
# Results

## Process reproducibility

We assessed process reproducibility 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-generated* data was considered shared if it was publicly available, or made accessible by the authors after we requested it. We examined the paper 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. We did not code whether *source* data were available (existing data sources not prepared specifically for the paper; e.g., Consumer Price Index, U.S. Bureau of Labor Statistics). We conducted a similar search for analytic code that appeared to produce 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%]).

Based on our criterion of obtaining at least data, this means that 146 achieved process reproducibility and 454 did not, 24.3% [95% CI 21.1 - 27.9%]. Using a criterion for process reproducibility based only on data availability means that it is, in principle, possible to reanalyze the data to assess outcome reproducibility. A more stringent criterion of data and code availability (20.3% [95% CI 17.3 - 23.7%]) adds the benefits of ease and direct examination of original code to evaluate outcome reproducibility.



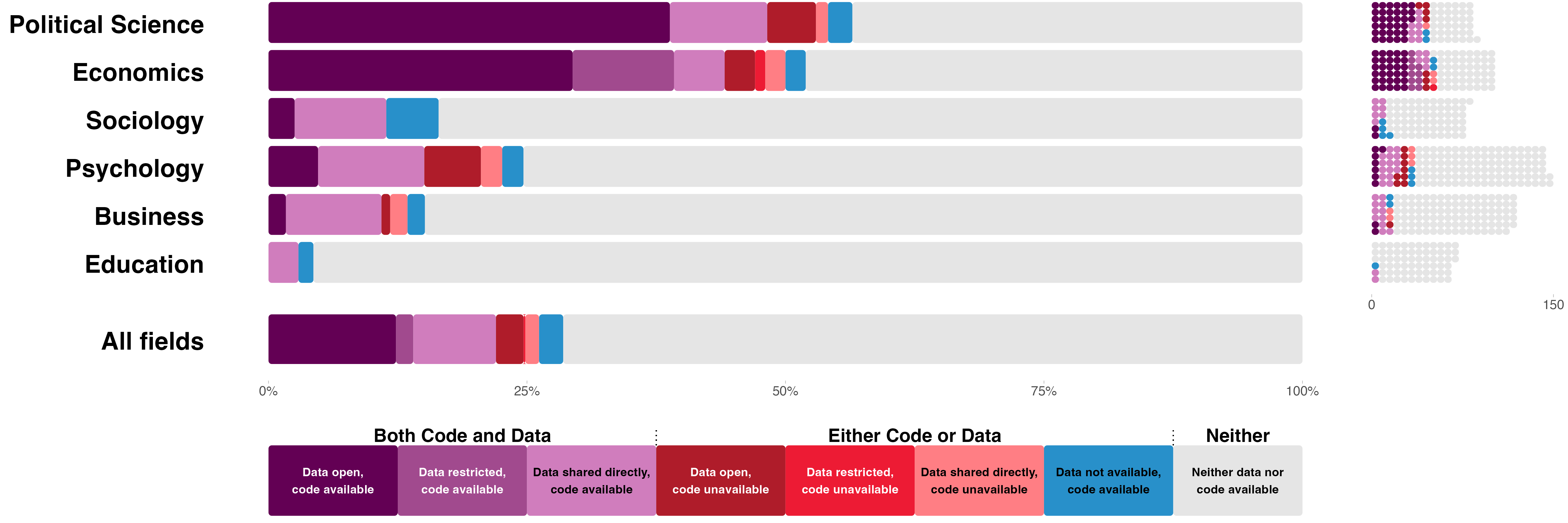
## Process reproducibility by year of original publication

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, and also to standards in data sharing improving over time.[11](https://www.zotero.org/google-docs/?uPaBCW) Our 10-year time span of papers involved both widespread discussion of reproducibility and changes to journals preproducibility 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.

## Process reproducibility 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 process reproducibility 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%]). In the supporting information, Figure S6 illustrates even higher process reproducibility 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 outcome reproducibility

Whereas we evaluated process reproducibility only at the paper level, outcome reproducibility could also be assessed for individual claims within papers. We mostly extracted and evaluated single key claims from papers, but for a subset of papers, multiple claims per paper were extracted.[25](https://www.zotero.org/google-docs/?OReo9x) 59 of 144.9 (40.7%) papers had >1 claim assessed for outcome reproducibility (mean claims per paper = 3.8, SD = 5.8, range = 1-37). In total, there were 553 claims from 144.9 papers assessed for outcome reproducibility.

The outcome reproducibility of multiple claims within a paper could be statistically independent, but in practice are likely to be dependent because they are from the same project and authors. We assessed outcome reproducibility at the (1) claim level and (2) paper level with weighting to account for interdependence among claims. As such, outcome reproducibility for papers could be a fraction based on the outcomes of multiple claims from the same paper. We report paper-level outcomes in the main text and claim-level outcomes in supporting information (Figures S1-S3).

## Outcome reproducibility assessments in comparison with the sample

Assessing outcome reproducibility depended on data availability, which varied by discipline and time. Table 2 presents the distribution of papers by discipline across events in the research process that could have shifted representativeness.[26](https://www.zotero.org/google-docs/?bfj6jD) The initial stages of selecting papers and identifying claims maintained representativeness by discipline (see also Tables S3 and S5).

Successfully obtaining author-generated data (n = 146) defined most of the sample for conducting outcome reproducibility tests. We supplemented that with 37 papers for which we did not have access to authors’ data but were able to obtain source data to recreate the authors’ datasets. The combination of these is represented in Table 2 as the “papers with source or author data available.” Political Science and Economics papers became a more substantial portion of the sample. Sociology was mostly unchanged, and the other fields were a less substantial portion of the sample. At the claims level, Economics had a notably higher proportion because of having a larger number of reanalyzed claims per paper. A similar analysis of representativeness by publication year is available in the supporting information.

## Observed outcome reproducibility

Analysts were matched with data to reanalyze and followed a structured protocol. Outcome reproducibility was investigated with three possible outcomes: precise reproducibility, approximate reproducibility, and not reproduced. Precise reproducibility was achieved if 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 one or more of the statistical 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%) |

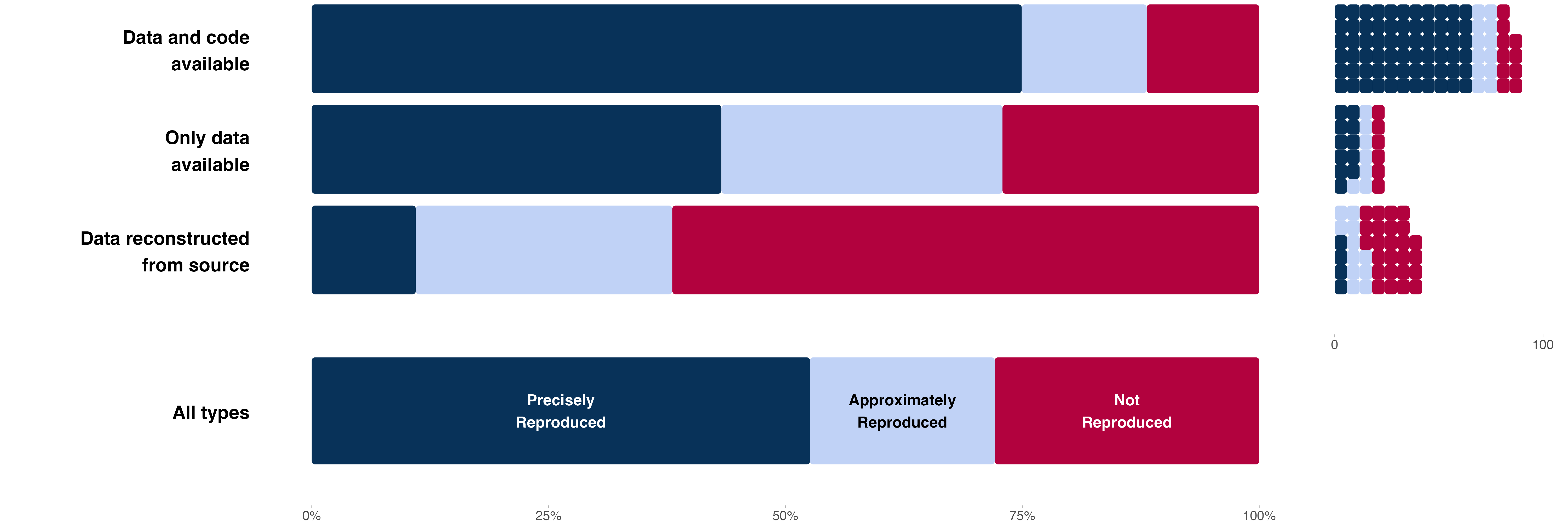
While 148 papers and 557 claims had at least one outcome reproduction attempt, 3.1 papers and 4 claims had none of our eligible statistical outcomes and were not counted for the quantitative assessment of outcome reproducibility. As such, 144.9 papers consisting of 553 claims were assessed for outcome reproducibility. For 7 papers, an outcome reproduction attempt began, but the analysts’ determined that the material they had was not sufficient to assess outcome 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 to complete a reproduction attempt. These were counted as outcome reproducibility failures.

Of the 144.9 papers that were assessed, we observed approximate or precise reproducibility for 104.4 papers (72.1% [95% CI 65.0 - 78.7%]) and precise reproducibility for 76.2 papers (52.6% [95% CI 45.8 - 60.3%]).

Figure 3 shows outcome 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 73.2 of the 83.1 papers (88.1% [95% CI 80.9 - 94.7%]) and precise reproducibility for 62.3 of the 83.1 papers (74.9% [95% CI 66.2 - 83.5%]). For 52.1 (62.6% [95% CI 53.5 - 72.2%]) of these papers, 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*.[27](https://www.zotero.org/google-docs/?SmPDkH)

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 54.0 - 87.7%]) and precise reproducibility for 9.5 of the 22.1 papers (43.2% [95% CI 26.0 - 60.4%]).

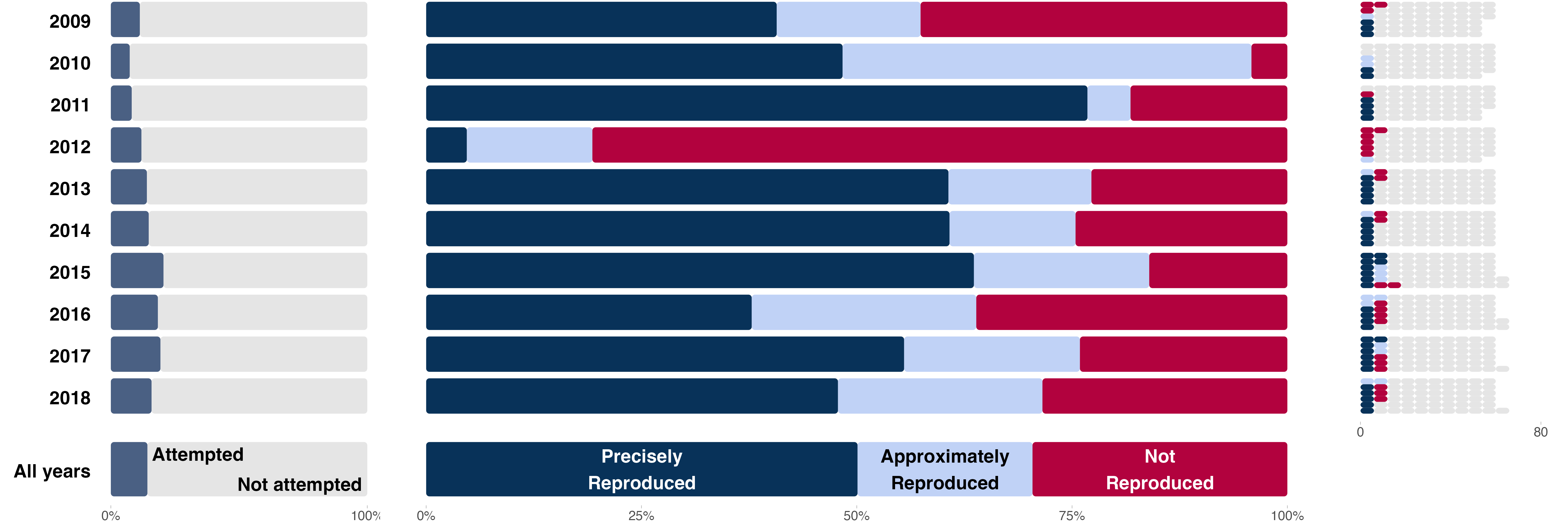
When author-prepared data were unavailable, but source data were available, we attempted to reproduce the findings by preparing the data and generating new code. Of these, we observed approximate or precise reproducibility for 15.1 of the 39.7 papers (38.1% [95% CI 23.9 - 52.9%]) and precise reproducibility for 4.4 of the 39.7 papers (11.0% [95% CI 3.0 - 20.0%]). In summary, outcome reproducibility rates were comparatively high when data and code were both available, and comparatively low when needing to reconstruct the data and code.



In addition to the empirically defined outcome 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%]).

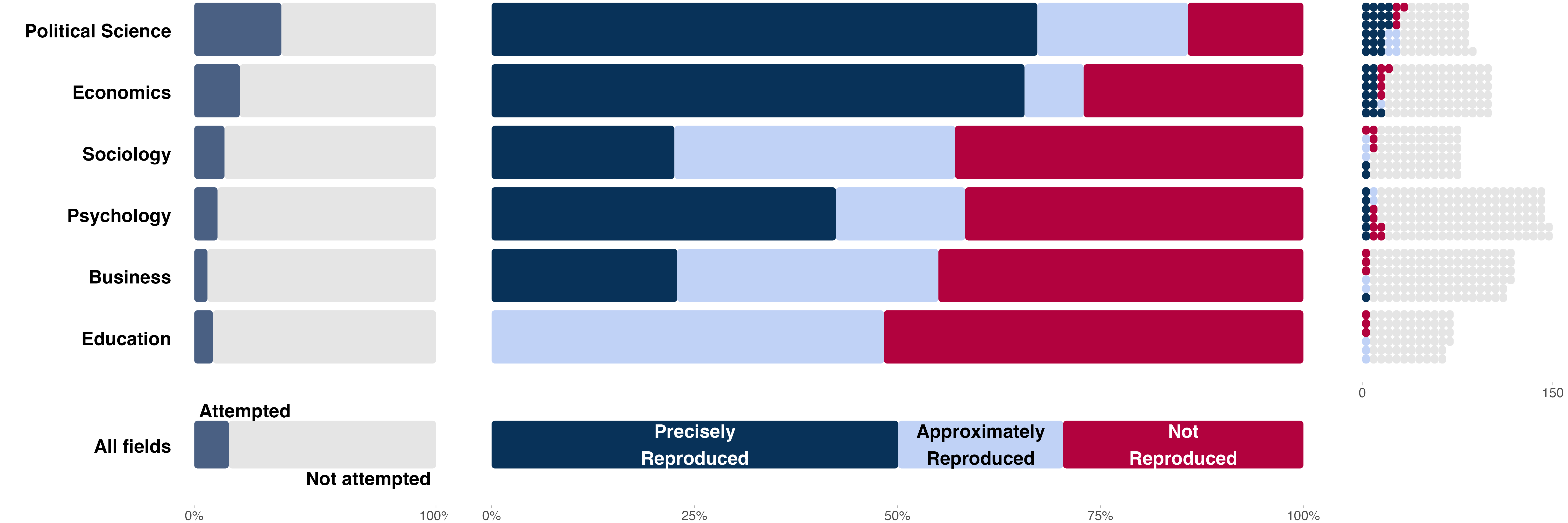
## Outcome reproducibility by year of original publication

Figure 4 presents outcome reproducibility by year. The number of outcome 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 - 0.153]) and the prevalence of approximate and precise reproducibility (Spearman’s *ρ* = 0.007 [95% CI -0.183 - 0.206]) were not significantly associated with time.



## Outcome reproducibility by discipline

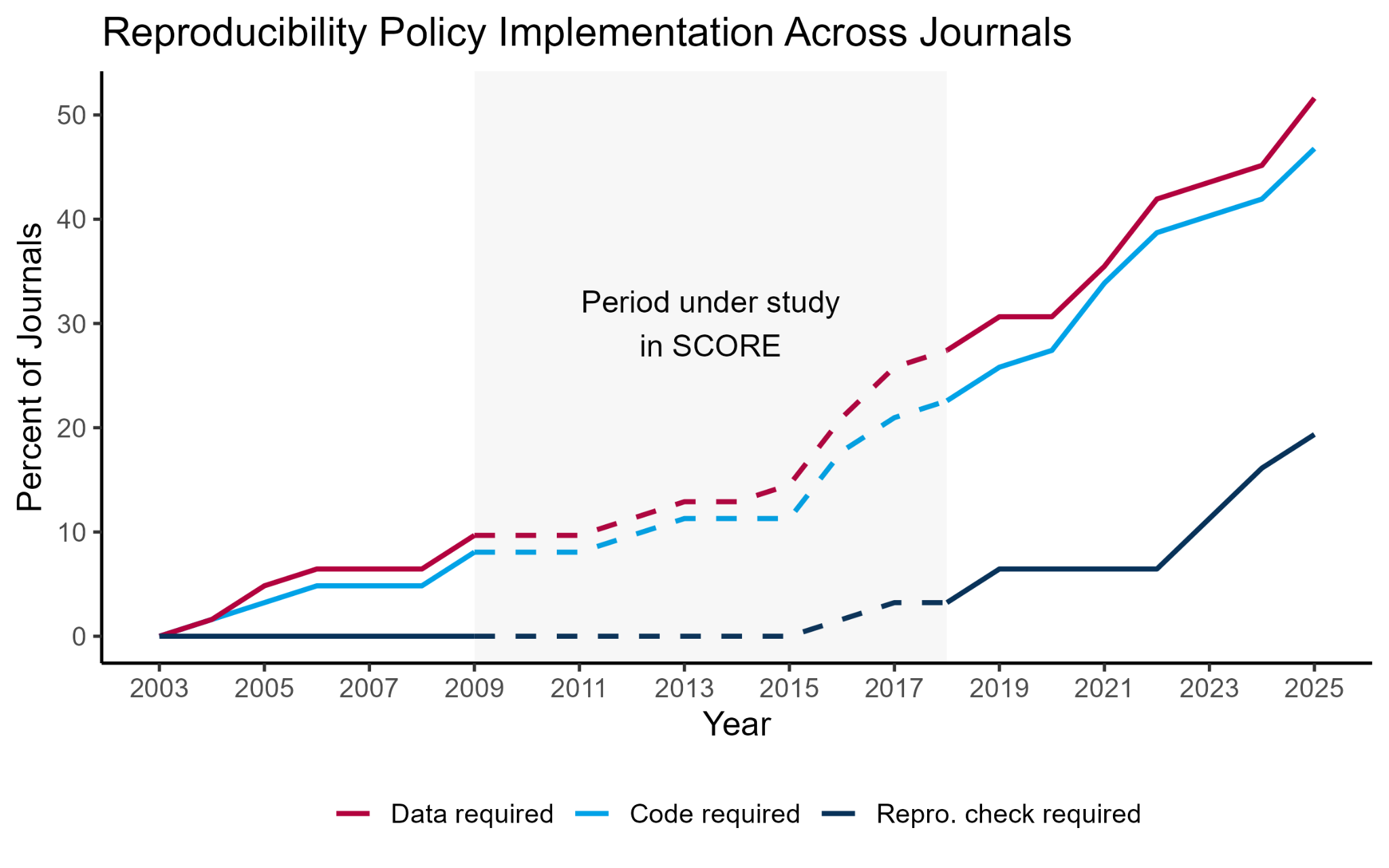
Figure 5 presents outcome reproducibility by discipline (see also Figures S4, S5, S9, S10)). Political Science and Economics had much higher rates of reproduction attempts than other fields due to greater data availability. Considering only papers with a reproduction attempt, we observed approximate or precise reproducibility for 34.7 of 41.9 (82.8% [95% CI 72.5 - 92.7%]) Political Science papers and 23.9 of 33.0 (72.5% [95% CI 57.6 - 86.3%]) Economics papers. We observed precise reproducibility for 27.3 of 41.9 (65.1% [95% CI 52.4 - 77.2%]) Political Science papers and 22.2 of 33.0 (67.2% [95% CI 51.8 - 81.6%]) 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 55.5 - 75.8%]) papers and precise reproducibility for 26.7 of 70.0 (38.2% [95% CI 28.1 - 48.6%]) papers.



## Process and outcome reproducibility by journal policies

We conducted a follow-up exploratory investigation of the relationship between journal policies and reproducibility success (Figure S11; Tables S7-S13). Process reproducibility success was observed more consistently for papers published in journals requiring data sharing (xx%), data and code sharing (xx%), or data and code sharing and reproducibility checks (xx%) than for papers published in journals with none of those policies (xx%; X²(3) = 240.93, p < .0001). Conditional on attempting an outcome reproduction, precise outcome reproducibility was observed more frequently for papers published in journals requiring data sharing (xx%), data and code sharing (xx%), or data and code sharing and reproducibility checks (xx%) than for papers published in journals with none of those policies (xx%) though the evidence was only suggestive (X²(6) = 14.81, p = 0.0218). Similar analyses at the claim level were consistent and significant (X²(6) = 46.44, p < 0.0001). Analyses using an ordered logistic regression model at both the paper and claim-levels suggested that the association between journal policy and outcome reproducibility was primarily observed in distinguishing papers and claims that were not reproduced from papers and claims that were approximately and precisely reproduced.

There have been incremental increases in journal policies requiring data sharing, code sharing, and reproducibility checks over time (Figure 6). 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 policy, and 6.8% of journals from other fields did so. As of mid-2025, those 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. 88.9% of journals from Economics and Political Science had at least one policy, and 36.4% of journals from other fields did so.



# Discussion

The most substantial barrier to observing reproducibility in our random 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 this might be due in part to the presence of journal policies requiring sharing data, code, or reproducibility checks. Looking forward, the proportion of journals requiring these policies has increased since 2018. These findings provide several insights about reproducibility in the social and behavioral sciences.

## Does a process reproducibility failure mean that the outcomes cannot be trusted?

No. A process reproducibility failure is a failure of transparency and verifiability. It is possible that the results would be perfectly reproducible if the data were available, but – because they are not – it is not possible to find out. The primary consequence of process reproducibility failures is uncertainty, we cannot know whether the results are reported precisely.

Our criterion for process reproducibility was the availability of author data that was prepared for the analyses reported in the paper. It is likely that more data or code could have been accessed if we had adopted more assertive methods to obtain it.

We could have relaxed the definition of what counted as data availability beyond author-prepared datasets, such as including occasions for which unprepared source data could be obtained, though we observed much less outcome reproducibility in such cases. There may be a tradeoff between process and outcome reproducibility. Greater leniency on what counts as sufficient data sharing may be associated with greater failures in reproducing the outcomes. A hypothesis to account for that tradeoff is that reporting on data analysis in papers can be insufficient to reconstruct the actual data preparation and analysis pipeline from source data if documentation of preparation steps is incomplete.[17](https://www.zotero.org/google-docs/?I0tYgt)

How should we handle process reproducibility failures then? The estimated outcome 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 outcome 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 the outcomes were reported precisely in the original paper, then the higher reproducibility estimate might be closer to reality. It is a near certainty that some of the papers with unavailable data would have reproduced successfully if the data could have been obtained.

Whatever one’s perspective, the benefits of achieving process reproducibility are clear. Imagine that a failure of process reproducibility says nothing about outcome reproducibility and our observed approximate or precise reproducibility rate of 72% is true for quantitative social and behavioral sciences. Then, the outcome reproducibility rate for papers with unavailable data is the same as papers with available data. For the papers without data available, the findings for 72% would be reproducible, we just would not know which 72%. Our findings suggest that reported results cannot be assumed to be precise, and a lack of process reproducibility prevents verification.

## Does an outcome reproducibility failure mean that the finding is wrong?

No. An outcome 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 correctly despite the independent failure to reproduce them. Prior evidence suggests imperfect consistency across reproducibility analysts, and perhaps more so when working from raw data rather than author-prepared datasets, suggesting that this plays a meaningful role. 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. We did not assess whether the finding would be interpreted the same way as the original interpretation. Regardless, we should not be sanguine about being almost correct. The similarity between original and reproduction outcomes could be coincidental, masking fundamental differences in the underlying code or data that change the meaning or interpretation; or it could be trivial, such as a rounding error, that has no impact on the interpretation.[28](https://www.zotero.org/google-docs/?LAK6RG) Without being able to precisely reproduce the original outcomes, the similarity is an ambiguous indicator of whether the underlying differences are meaningful.

Finally, an outcome reproducibility test can fail because of an error in the original research. Even the most experienced researchers will make errors in data management, analysis, recordkeeping, and transcription. Science is difficult, often messy. That does not excuse errors or obviate their consequences, but it does mean that establishing reproducibility requires skill and is an achievement. 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 trustworthy researchers will sometimes be unable to detect and correct mistakes.

## Does an outcome reproducibility success mean that the finding is correct?

No. Outcome reproducibility success means that the results are reported precisely. They are computationally reproducible. Precisely reported findings 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 a large number of analyses inflating the likelihood of exaggerated findings. Outcome reproducibility is a baseline assessment of credibility. Quantitative findings should be able to meet this standard. Those that do can then be productively interrogated on other dimensions of credibility and correctness.[29](https://www.zotero.org/google-docs/?F8Jl03)

## What accounts for the differences in process and outcome reproducibility?

Papers from Political Science and Economics were both more likely to achieve process and outcome reproducibility than papers from other fields. This prompted an exploratory investigation that observed an association between process and outcome reproducibility success and journal policies requiring data sharing, code sharing, and reproducibility checks. Economics and Political Science were more likely to have such policies between 2009 and 2018 from which papers were sampled for this investigation. These exploratory results prompt the hypothesis that journal transparency policies cause increases in reproducibility. We also observed that transparency policies have become more popular across our sample of social and behavioral science journals since 2018. This prompts a follow-on hypothesis that the process and outcome reproducibility success rates would be higher if this investigation were replicated with the same journals using a sample of papers from 2019 to 2025. 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 complex interactions between different causes. For example, a transparency policy could ironically cause an increase in process reproducibility and a decrease in outcome reproducibility if implementing the policy is highly burdensome and is done so without training and norms for preparing data and code for sharing in that research community. Data and code might be shared, but be unusable by independent researchers.

## 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 process reproducibility 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 outcome 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 process and outcome reproducibility rates have fallen short of perfection.[5,9,10,18](https://www.zotero.org/google-docs/?4k7qvx) Given that not all papers could be assessed for outcome reproducibility, it is unlikely that our outcome reproducibility estimate is precise for our full sample, or for the social and behavioral sciences generally. But, this evidence does suggest that reproducibility practices can improve in all disciplines investigated.

## Observed reproducibility may never reach 100%

Even if all findings are reported precisely, there are occasions in which reproducibility will not be easily verified because of barriers to data access. Principal challenges are privacy and proprietary concerns. 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. For example, some datasets can be anonymized to be publicly shareable for the purposes of demonstrating reproducibility of key findings.[31](https://www.zotero.org/google-docs/?yQDDcg) 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.[32](https://www.zotero.org/google-docs/?u0oQN4) 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.[33](https://www.zotero.org/google-docs/?3vs5Px)

## Conclusion

A credo of the open scholarship movement is “as open as possible, as closed as necessary.”[34,35](https://www.zotero.org/google-docs/?Bil6Oz) 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.

# References

[move references for main text here]

# Figure Legends

Figure 1. Process reproducibility success rates 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 data and code available and not available. Note that middle purple and middle red reflect restricted data, which did not count as available data for process reproducibility, but might be accessible in principle.

Figure 2. Process reproducibility success rates 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 middle purple and middle red reflect restricted data, which did not count as available data for process reproducibility, but might be accessible in principle.

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

Figure 4. Outcome reproducibility by year of publication. Left: Proportion of outcome reproduction attempts from the sample of papers. Middle: Outcome reproducibility as a percentage of the attempts. Right: Outcome 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.

Figure 5. Outcome reproducibility by discipline. Left: Proportion of outcome reproduction attempts from the sample of papers. Middle: Outcome reproducibility as a percentage of the attempts. Right: Outcome reproducibility as counts compared with the sample of papers. Note that papers with multiple claims may have only some claims that were reproducible according to these success measures (i.e. are partly reproducible), but color coding of paper counts is rounded to the nearest paper.

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 outcome reproduction attempts (Journal of Finance, Journal of Public Administration Research and Theory, and Law and Human Behaviour) but their policies are included here.

# Method

We examined two aspects of reproducibility: *process reproducibility* assessed whether author data was available so that a reproduction could be attempted, and *outcome* *reproducibility* assessed whether the same outcome as reported originally was 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.[24](https://www.zotero.org/google-docs/?rrjfLi) 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).[36,37](https://www.zotero.org/google-docs/?ttIYAv) Data, materials, code, and other outputs from the program that can be shared 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 using a Python script from 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 it was possible to extract a statistically significant test result that supported a claim made in the paper’s abstract.[25](https://www.zotero.org/google-docs/?USSoFZ) Papers that did not produce an eligible claim 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 reproductions of their findings, with some adjustments made for representativeness (see Abatayo and colleagues [2025] for details on the sampling and selection process).[26](https://www.zotero.org/google-docs/?b48FpP) Other papers and data were gathered during the SCORE program, but they did not include reproduction attempts and are not discussed in this paper.

## Process reproducibility

We assessed process reproducibility 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 may include data or code. If either data or code were not located from the paper, coders searched for publicly available materials online, checking specifically 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 outcome reproducibility assessment.

## Outcome reproducibility

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, 144.9 papers were assessed for outcome reproducibility. Here, random sampling is lost because selection for outcome 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 detailed above (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).

## 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 reproductions 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).

# Methods References

[references for methods here]

# 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 about ¼ of published papers could be assessed for reproducibility because of data unavailability. We also found that Economics and Political Science had higher reproducibility rates than other disciplines. Reproducibility failures add unnecessary uncertainty to the complex enterprise of knowledge production and 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 this 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](mailto: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 | Ashland University | 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 |  |
| 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 | 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 |  |
| Michalak | Johannes | 0000-0003-4701-5464 | Witten/Herdecke University |  |
| 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 process reproducibility 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 | 0 | 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). |
| 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 process reproducibility 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 | 0 | 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. |
| 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 | 0 | 0 | 1 | Conducted a computational replication of one study. writing and editing. |
| 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. |
| 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 | 0 | 0 | 0 | 0 | 0 | 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 think Olivia can look up how many I edited and which ones. |
| 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) |
| 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 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 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 |
| 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 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 |  |
| Michalak Johannes | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| 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 | 0 | 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. |
| 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). It is possible some of the other studies I reviewed or edited reviews for preregistration on were also reproductions but labelled as replications; I recall working on more than 3 reproductions. See list in comments below. |
| 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\_9DZl" (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**,.](https://www.zotero.org/google-docs/?ZUlbKd)

[2. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. (The National Academies Press, Washington, DC, 2019). doi:10.17226/25303.](https://www.zotero.org/google-docs/?ZUlbKd)

[3. Chang, A. C. & Li, P. Is Economics Research Replicable? Sixty Published Papers FromThirteen Journals Say “Often Not”. *Crit. Finance Rev.* **10**, (2021).](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[5. Vilhuber, L. Reproducibility and replicability in economics. *Harv. Data Sci. Rev.* **2**, 1–39 (2020).](https://www.zotero.org/google-docs/?ZUlbKd)

[6. Brodeur, A. *et al. Mass Reproducibility and Replicability: A New Hope*. https://econpapers.repec.org/paper/zbwi4rdps/107.htm (2024).](https://www.zotero.org/google-docs/?ZUlbKd)

[7. Pérignon, C. *et al.* Computational Reproducibility in Finance: Evidence from 1,000 Tests. *Rev. Financ. Stud.* **37**, 3558–3593 (2024).](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[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.](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[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.](https://www.zotero.org/google-docs/?ZUlbKd)

[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.](https://www.zotero.org/google-docs/?ZUlbKd)

[15. Eubank, N. Lessons from a Decade of Replications at the Quarterly Journal of Political Science. *PS Polit. Sci. Polit.* **49**, 273–276 (2016).](https://www.zotero.org/google-docs/?ZUlbKd)

[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).](https://www.zotero.org/google-docs/?ZUlbKd)

[17. Breznau, N. The reliability of computational replications: a study in computational reproductions. *R. Soc. Open Sci.* (2025) doi:10.1098/rsos.241038.](https://www.zotero.org/google-docs/?ZUlbKd)

[18. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).](https://www.zotero.org/google-docs/?ZUlbKd)

[19. 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).](https://www.zotero.org/google-docs/?ZUlbKd)

[20. 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).](https://www.zotero.org/google-docs/?ZUlbKd)

[21. Khan, N., Thelwall, M. & Kousha, K. Data sharing and reuse practices: disciplinary differences and improvements needed. *Online Inf. Rev.* **47**, 1036–1064 (2023).](https://www.zotero.org/google-docs/?ZUlbKd)

[22. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).](https://www.zotero.org/google-docs/?ZUlbKd)

[23. Brodeur, A. *et al.* Promoting Reproducibility and Replicability in Political Science. *Res. Polit.* **11**, 20531680241233439 (2024).](https://www.zotero.org/google-docs/?ZUlbKd)

[24. Alipourfard, N. *et al.* Systematizing Confidence in Open Research and Evidence (SCORE). (2021).](https://www.zotero.org/google-docs/?ZUlbKd)

[25. Tyner, A. H. *et al.* Extracting claims from empirical publications for the SCORE Program. *Preprint* (2025).](https://www.zotero.org/google-docs/?ZUlbKd)

[26. Abatayo, A. L. *et al.* Overview of the SCORE Program Methodology and Reporting. *Preprint* (2025).](https://www.zotero.org/google-docs/?ZUlbKd)

[27. 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).](https://www.zotero.org/google-docs/?ZUlbKd)

[28. Clemens, M. A. The Meaning of Failed Replications: A Review and Proposal. *J. Econ. Surv.* **31**, 326–342 (2017).](https://www.zotero.org/google-docs/?ZUlbKd)

[29. Nuijten, M. B., Bakker, M., Maassen, E. & Wicherts, J. M. Verify original results through reanalysis before replicating. *Behav. Brain Sci.* **41**, e143 (2018).](https://www.zotero.org/google-docs/?ZUlbKd)

[30. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).](https://www.zotero.org/google-docs/?ZUlbKd)

[31. Morehouse, K. N., Kurdi, B. & Nosek, B. A. Responsible data sharing: Identifying and remedying possible re-identification of human participants. *Am. Psychol.* (2024).](https://www.zotero.org/google-docs/?ZUlbKd)

[32. Nab, L. *et al.* OpenSAFELY: A platform for analysing electronic health records designed for reproducible research. *Pharmacoepidemiol. Drug Saf.* **33**, e5815 (2024).](https://www.zotero.org/google-docs/?ZUlbKd)

[33. Quintana, D. S. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* **9**, e53275 (2020).](https://www.zotero.org/google-docs/?ZUlbKd)

[34. Commission, E. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. (2016).](https://www.zotero.org/google-docs/?ZUlbKd)

[35. Landi, A. *et al.* The “A” of FAIR – As Open as Possible, as Closed as Necessary. *Data Intell.* **2**, 47–55 (2020).](https://www.zotero.org/google-docs/?ZUlbKd)

[36. Abatayo, A. L. *et al.* Empirical, Human, and Machine Assessments of Research Credibility in the Social and Behavioral Sciences. *Preprint* (2025).](https://www.zotero.org/google-docs/?ZUlbKd)

[37. Tyner, A. H. *et al.* Investigating the replicability of the social and behavioral sciences. *Preprint* (2025).](https://www.zotero.org/google-docs/?ZUlbKd)

Supporting Information for “Investigating the reproducibility of the social and behavioral sciences”

[add authors here]

# Table of Contents

[add toc here]

# Overview

This supplement provides details on the methodology and additional results for the reproductions attempted during the 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. (2025). 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 is available in Abatayo et al. (2025).

# Methods

In the "Systematizing Confidence in Open Research and Evidence" (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 which facilitated these goals and gives relatively less attention to specific operational steps which are less relevant to understanding SCORE’s reproduction evidence such as grant making and managing the cooperative agreement with DARPA. The procedures documented below evolved over the three years of SCORE. When relevant to interpreting SCORE evidence, we highlight changes to our processes and indicate when in the project’s timeline the changes were implemented.

Several parts of the methods for this supporting paper are very similar to the supporting information for SCORE replications because of a shared methodology (Tyner et al., 2025). For some sections, the same initial description was used and then edited separately for the features unique to the replication and reproduction attempt methods.

## Key Terms

Abatayo and colleagues (2025) 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. *Process reproducibility* refers to it being possible, in principle, to assess reproducibility because at least the author data is available. *Author data* refers to the dataset prepared and used by the 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 author dataset for analysis.

*Outcome 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 author dataset was not 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 comparing outcome reproducibility with 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. (2025). 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 actual 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., 2025). 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. Each of those 3,000 papers, sometimes referred to as the Annotation dataset, 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 process reproducibility, and a subset were examined for outcome 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 evidence had been gathered, second on papers likely to be able to produce replication and reproduction evidence, 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.

## Process Reproducibility (PR) Assessment

### Overview

We conducted an assessment of process reproducibility 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 process reproducibility assessment reported in the main text. There were two parts: paper review and online search.

#### Paper Review

When coding for process reproducibility 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 dataset from Phase 1 were informed about the project, that their paper had been selected randomly, and asked for their feedback about the claim extracted from their paper.

From the authors of the 600 papers eligible for process reproducibility assessment, if either data or code were not located from the prior steps outlined above, we requested the missing material (either data, code, or both) necessary to reproduce the analyses and results from the original paper.

### Coding Data and Code Sharing

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 shared. To count as being shared, the data needed to be the author-generated data for analysis, not source data.

### Limitations

Data or code that was evaluated as “not shared” could have received that designation for reasons beyond the authors’ control. For secondary data research, data could be restricted access by the originator. We did not try to access restricted data. The outreach to authors occurred sporadically over a long period of time between 2020 and 2023. Willingness to share 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 shared data is constrained to author-prepared datasets and availability of source data (e.g., Census data) did not qualify. The main text reports evidence that outcome reproducibility was much weaker for cases in which we obtained source data and attempted to reconstruct the author data and repeat the analysis. Nevertheless, this definition of data sharing importantly informs the interpretation of process reproducibility 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 shared” is a statement of the outcome of the end of this process, not an attribution of responsibility.

## Outcome Reproducibility (OR) 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 (2025). 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: Psychological Science Accelerator (mostly psychology and behavioral economics laboratories; <https://psysciacc.org/>), the Berkeley Initiative for Transparency in the Social Sciences (BITSS; <http://bitss.org/>) an extensive network of economists, sociologists, political scientists, psychologists, and other social scientists, and International Initiative for Impact Evaluation (3ie; <http://www.3ieimpact.org>) has access to a global team of researchers from a variety of social sciences, particularly developmental economics. 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. (2025).

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 an [onboarding email and additional instructions](https://osf.io/6gzqv?view_only=47ab47bd267e4435a0054ff73eaba101) specific to the reproduction type. Project coordinators confirmed that they 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 passed 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 recruiting an 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](https://osf.io/sx48f?view_only=47ab47bd267e4435a0054ff73eaba101) or an [Author Data Reproduction preregistration template](https://osf.io/xr63u?view_only=47ab47bd267e4435a0054ff73eaba101). 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 recruiting for 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 link: <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 often accessible to the authors and may have introduced unwelcome influence during review. This is different from replication attempts in which original authors were invited to participate in the peer review process (Tyner et al., 2025).

### 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](https://osf.io/geq3n?view_only=47ab47bd267e4435a0054ff73eaba101) rather than the 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 Phase1 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](https://osf.io/7tdea?view_only=47ab47bd267e4435a0054ff73eaba101) (Phase 1) or a [reproduction criteria template](https://osf.io/geq3n?view_only=47ab47bd267e4435a0054ff73eaba101) (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. 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 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 outcome reproduction success

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

Claims were rated as *push-button reproducible* (a subset of *precisely reproducible* for reporting purposes in the main text) if the authors’ data and code were available, and if the code could be executed on the data with minimal revisions to the code, and if the observed outcomes precisely reproduced 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 reproduced 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 data 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](https://osf.io/euxjc?view_only=47ab47bd267e4435a0054ff73eaba101) 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](https://osf.io/fj2b4?view_only=47ab47bd267e4435a0054ff73eaba101) 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 link <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 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 project authors.

Following the original 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 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 for confirmation.

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 started but were 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. 10 papers with completed reproductions occurred following the replication workflow reported in Tyner et al (2025) 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 Outcome Reproducibility

In the main text, we focused on paper-level reporting of outcome 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 this analysis means that 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 outcome reproducibility results in the main text, but now report the data at the claims-level.

Of the 553 claims that were assessed for outcome 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 outcome 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 80.9 - 94.7%]) 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*.

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%).

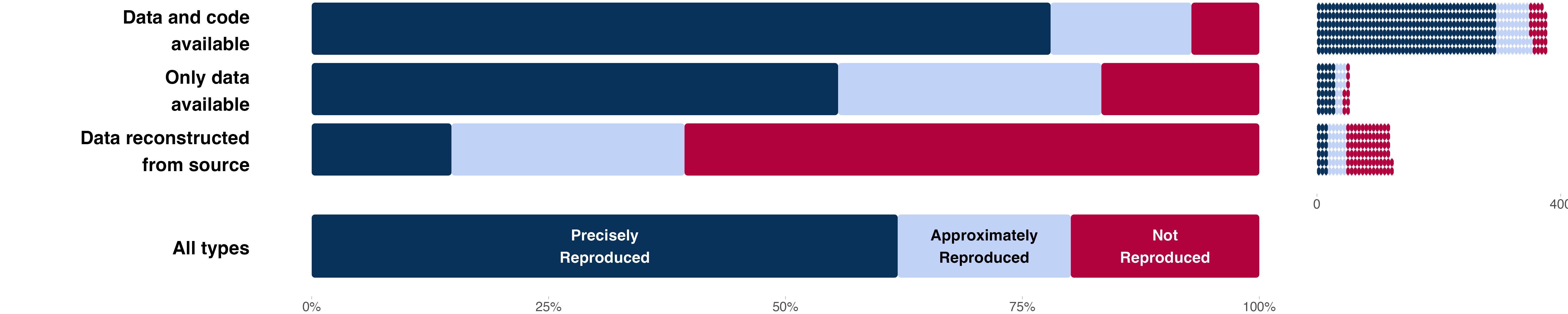
When author-prepared data were unavailable, but source data were available, we attempted to reproduce the findings by preparing the data and generating new code. Of these, we observed approximate or precise reproducibility for 48 of the 122 claims (39.3%) and precise reproducibility for 18 of the 122 claims (14.8%). In summary, outcome 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 process and outcome reproducibility were not directly 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 process reproducibility (i.e. the proportion of the literature which were implied to be outcome reproducibility-assessable) with our estimate of outcome reproducibility among those that were assessed.

Across papers, 24.3% [95% CI 21.1 - 27.9%] had data available and therefore assessable for outcome reproducibility. Among those assessed, 72.1% [95% CI 65.0 - 78.7%] of papers were approximately or precisely reproduced, and 52.6% [95% CI 45.8 - 60.3%] of papers were precisely reproduced. Together, this implies that 17.5% [95% CI 15.0 - 20.2%] of papers were approximately or precisely reproducible, and 12.8% [95% CI 10.5 - 14.9%] 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 therefore assessable for outcome reproduction. Among those assessed for outcome reproduction 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. Full reproducibility is a minimum baseline as it can only increase with additional effort to obtain author data, reconstruct datasets from original sources, or troubleshoot reanalysis challenges.

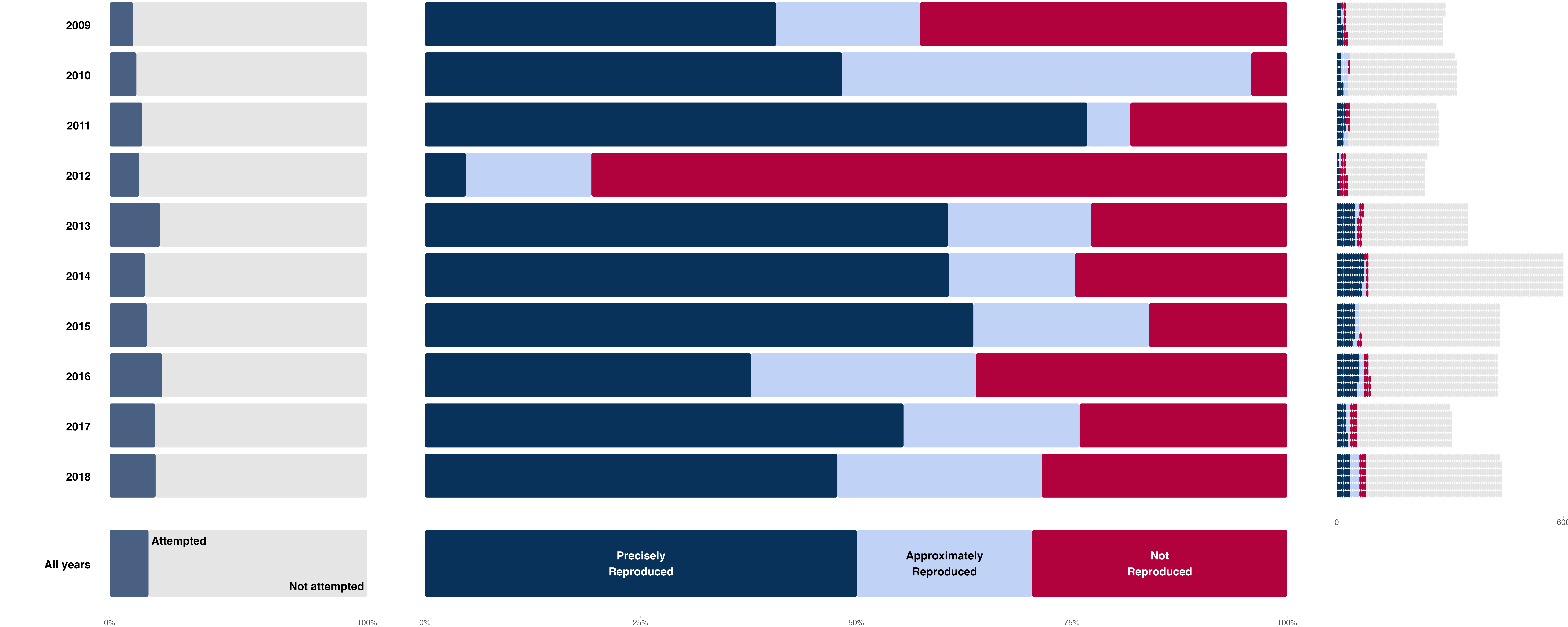
Figure S1. Outcome reproducibility by whether data and code were available, only data were available, or when the paper’s data were reconstructed from available source data for all claims



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

Figure S2 presents outcome reproducibility success by year. Sample sizes of outcome reproduction attempts within each year are 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 -0.125 - 0.053]). The prevalence of approximate or precise reproducibility was also not significantly associated with time (Spearman’s *ρ* = 0.007 [95% CI -0.080 - 0.098]). These findings are consistent with the results across papers as reported in the main text.

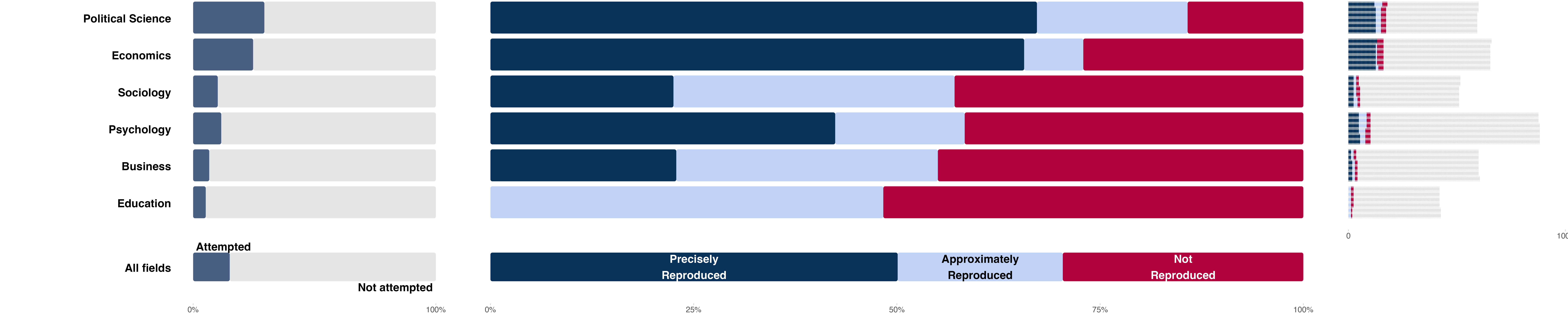
Figure S2. Outcome reproducibility by year of publication for all claims



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

Figure S3 presents outcome 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. Outcome reproducibility by discipline for all claims

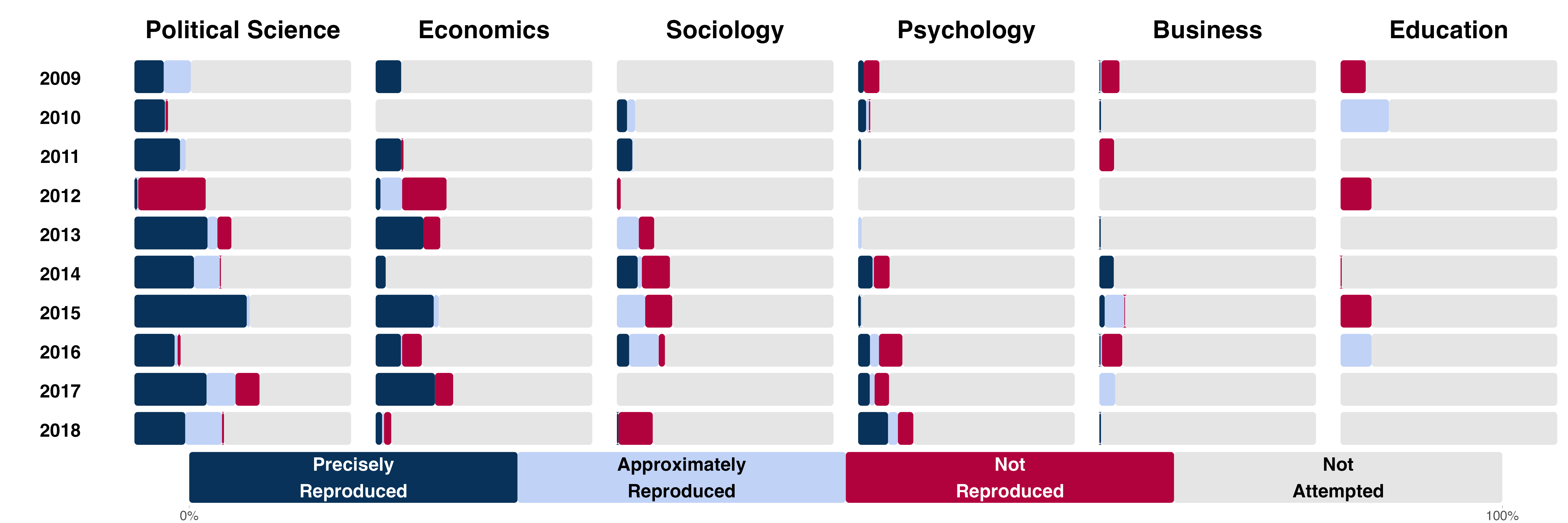


*Caption: The left column illustrates the proportion of outcome reproduction attempts from the sample of claims. The middle column illustrates outcome reproducibility as a percentage of the attempts. The right column illustrates outcome reproducibility as counts compared with the sample of claims.*

## Outcome reproducibility by discipline and year

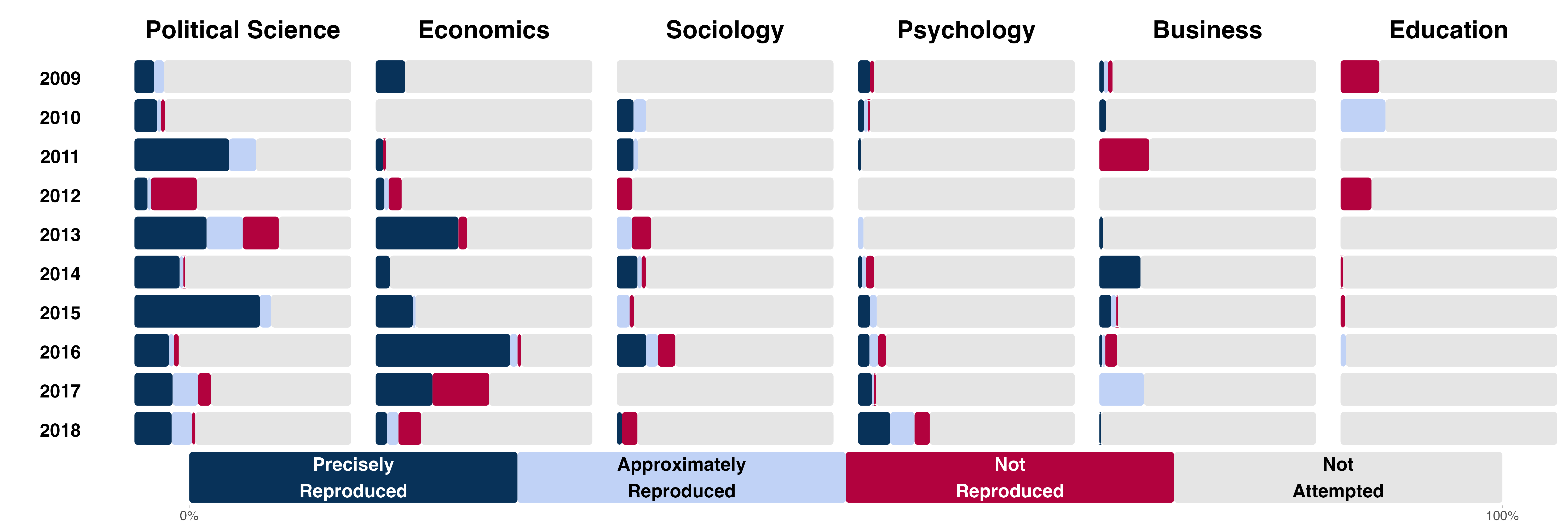
Figure S4 presents all of the process and outcome reproducibility attempts by discipline and by year across papers. Figure S5 provides the same information across claims. These figures provide a visualization of the descriptive findings that more data were available to assess reproducibility in more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

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



*Caption: Outcome reproducibility as a percentage of the sample of papers from each year and each discipline.*

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



*Caption: Outcome reproducibility as a percentage of the sample of claims from each year and each discipline.*

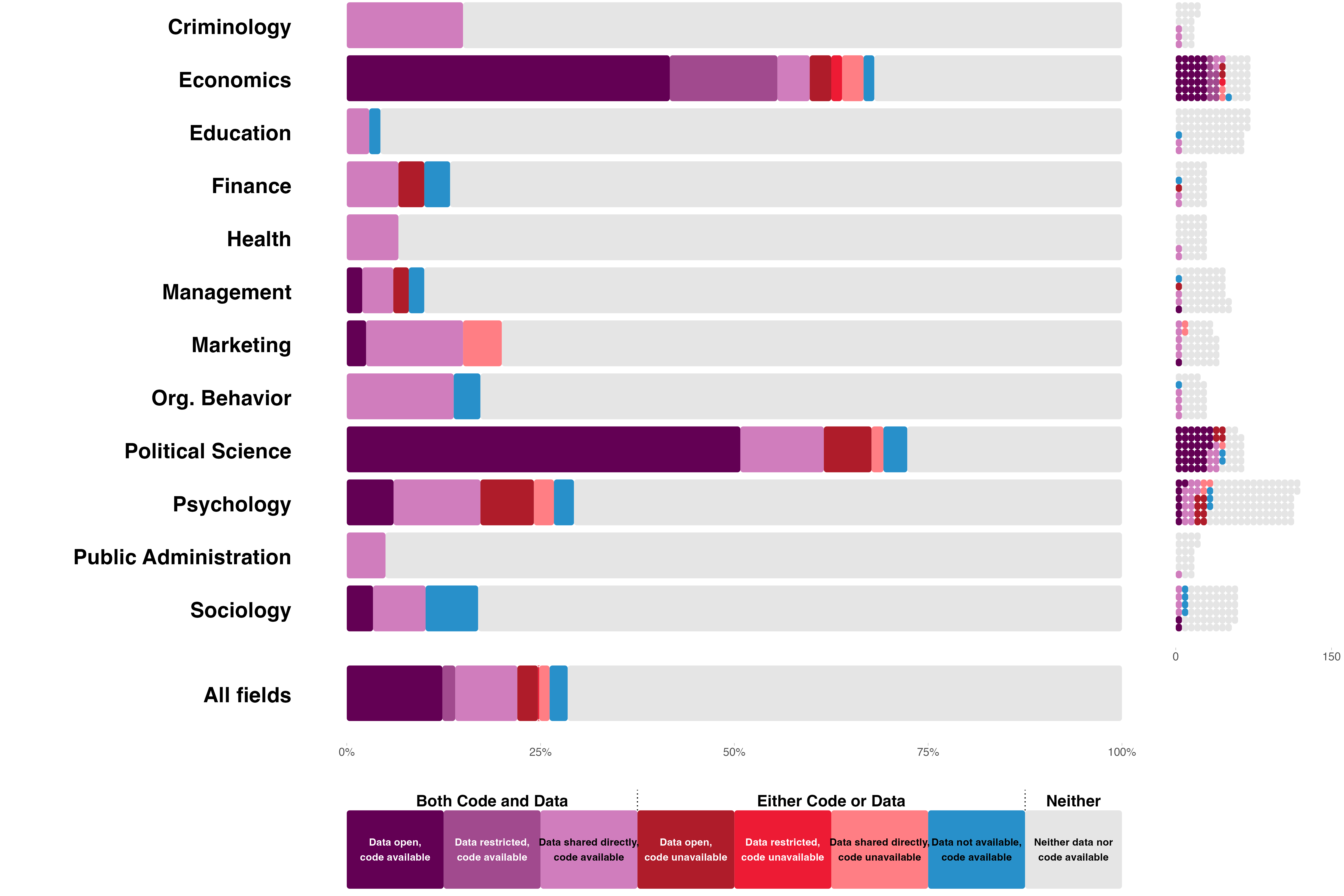
## Reproduction Outcomes by subfield

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 process reproducibility success is even more pronounced for Political Science and Economics after separating them from Public Administration and Finance, respectively. Likewise, Psychology’s process reproducibility performance is somewhat stronger after separating it from Health.

Figure S6. Process reproducibility success 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 middle purple and middle red reflect restricted data, which did not count as available data for process reproducibility, but might be accessible in principle.*

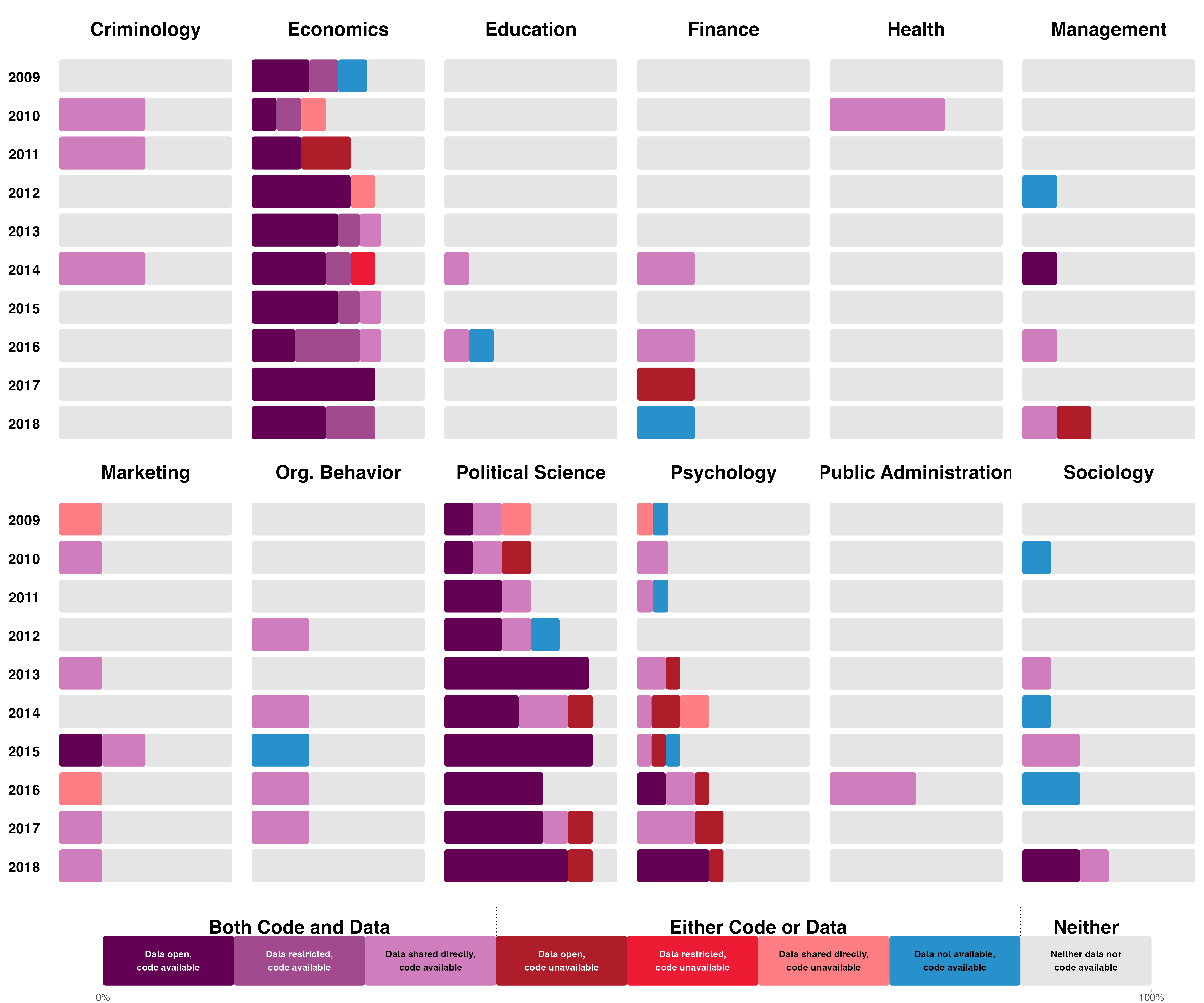
We plotted process reproducibility by discipline and year in Figure S7 for the 6 disciplines as reported in the main text. Combined, Political Science and Economics were 7.6 [95% CI 4.0 - 22.0] times as likely as to have open data and code (dark purple; 32.6% [95% CI 26.3 - 39.6%]) than the other disciplines 4.3% [95% CI 2.2 - 8.2%]). 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.

Figure S7. Process reproducibility success 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 middle purple and middle red reflect restricted data, which did not count as available data for process reproducibility, but might be accessible in principle.*

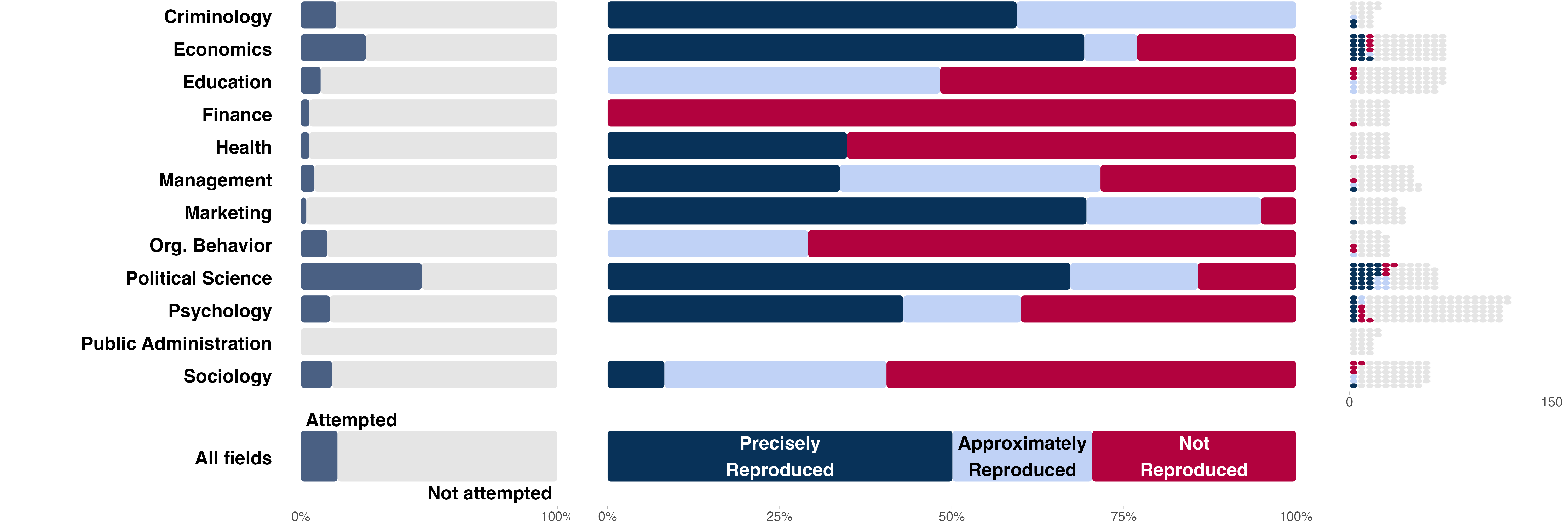
Figure S8. Process reproducibility success rates by year of publication for 12 subdisciplines



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

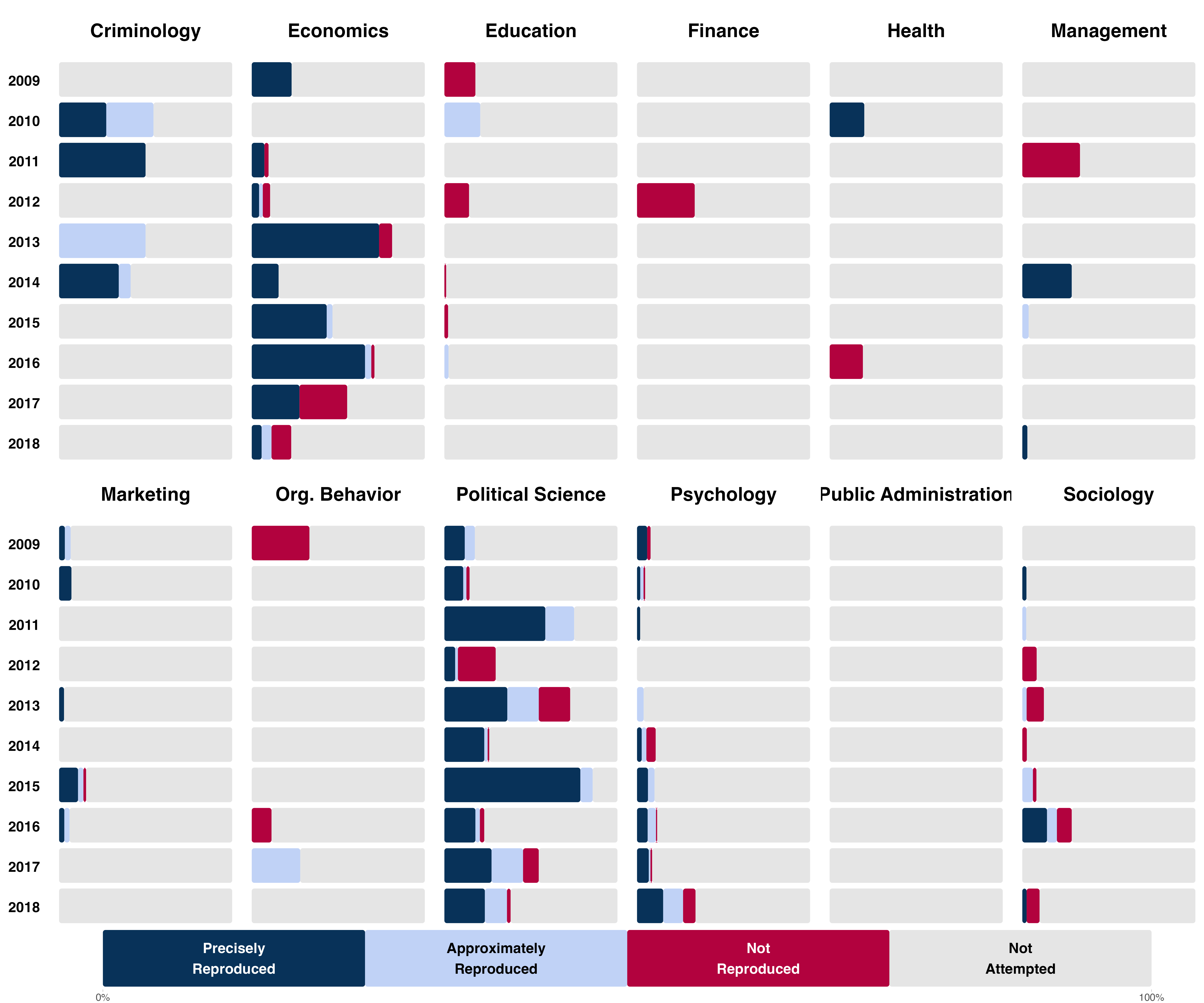
Figure S9 presents outcome reproducibility by the 12 subdisciplines. It complements Figure 5 from the main text that showed the same data across 6 disciplines. Figure S10 separates those same outcomes by year. The higher performance for Political Science and Economics after separating the subdisciplines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

Figure S9. Outcome reproducibility by 12 subdisciplines



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

Figure S10. Outcome reproducibility by 12 subdisciplines and by year



*Caption: Outcome reproducibility as a percentage of the sample of papers from each year and each discipline.*

## Outcome reproducibility assessments in the comparison with 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 outcome 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%) |

## 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 S4. Process reproducibility for papers by Journal

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

Table S5. Outcome reproducibility for papers by journal

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

# Relationship between journal policies and reproducibility

In the first version of this paper that was submitted for peer review, we included an exploratory analysis in the discussion illustrating that there is variation across disciplines in current (as of 2024) journal policies related to data and code sharing--with economics and political science journals having particularly strong standards compared with the other journals. With that data, we could not directly associate reproducibility with journal policies. Reviewers suggested that a closer look would be of substantial value. This section reports the details of our exploratory investigation conducted after that first round of peer review. 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, and some fields -- notably economics and political science -- have adopted such policies at higher rates than other disciplines.

With the help of the TOP database, we have documentation of relevant current policies, but we do not have sufficient documentation of 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. It was considered likely 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 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. For example, we collected confidence ratings on the provided dates of policy adoption. Also, we gathered converging evidence from [1] other researchers’ reviews of journal policies, [2] reports from journal staff, and [3] the Internet and TOP database.

## 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 **recommended** or **encouraged** data and code sharing.
* 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. This maximized coverage of the data we need to gather, and enabled cross-checking for accuracy across sources. These benefits outweigh, 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 already 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 completed as much of [the dataset](https://docs.google.com/spreadsheets/d/1KBdGqBUOz5zAMA-Mgv0XPgfzAkJrz28Ikd49jFyc2XA/edit?gid=1215742485#gid=1215742485) as possible using their sources using 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: XXXX. Once recognizing that verification of data verification and code verification was difficult and unreliable, we decided not to pursue them intensively or combine them into the final dataset. 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 finding journal editorials about new policies to complete the dataset. Some of this work was completed and already reported in the original version 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](https://docs.google.com/spreadsheets/d/1KBdGqBUOz5zAMA-Mgv0XPgfzAkJrz28Ikd49jFyc2XA/edit?gid=1215742485#gid=1215742485) with separate sheets for each team supported oversight of progress and identification of gaps or inconsistencies. Periodic check-ins among contributors occurred to troubleshoot problems or calibrate coding decisions.

The three independent 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](https://docs.google.com/spreadsheets/d/1KBdGqBUOz5zAMA-Mgv0XPgfzAkJrz28Ikd49jFyc2XA/edit?gid=335619581#gid=335619581).

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 the 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 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 requirements to 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 a 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 few journals.

### Combining Data

The three teams worked independently to complete the measures 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 source 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 very helpful, 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](https://docs.google.com/spreadsheets/d/1KBdGqBUOz5zAMA-Mgv0XPgfzAkJrz28Ikd49jFyc2XA/edit?gid=959450767#gid=959450767).

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% | 73% | 77% | 76% | 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 Planning

**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 outcome 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 (required data sharing, required code sharing, independent reproducibility check)?
* Outcome variable: Outcome 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 success on process reproducibility?

* 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.
* Process reproducibility 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: Process reproducibility (0 = failed process reproducibility check, 1 = passed process reproducibility 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 included data 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.

## 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 time frame has mostly followed this order, but not perfectly. 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.

In the original 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 data independent. As such, these results should be cautiously interpreted as the outcomes of an exploratory investigation.

### Outcome Reproducibility

The primary question was whether more stringent reproducibility policies were associated with greater success on outcome 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 cross-tabs are presented in Table S8 across papers and Table S9 across claims. Descriptively, papers were precisely reproduced more frequently with any transparency policy ({table\_s8\_1\_2\_p}, {table\_s8\_1\_3\_p}, and {table\_s8\_1\_4\_p}) than with no policy ({table\_s8\_1\_1\_p}); likewise, when comparing across claims ({table\_s9\_1\_2\_p}, {table\_s9\_1\_3\_p}, and {table\_s9\_1\_4\_p} with a policy; {table\_s9\_1\_1\_p} no policy). However, the chi-square test provided suggestive evidence across papers ({table\_s8\_chi2\_pval}) and significant evidence across claims ({table\_s9\_chi2\_pval}).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### **Table S8. Paper-level outcome reproducibility by journal policy (N =** **{table\_s8\_n})**

|  | **No policy** | **Data**  **required** | **Data and code**  **required** | **Data, code &**  **repro check required** | **Total** |
| --- | --- | --- | --- | --- | --- |
| Precisely  reproduced | {table\_s8\_1\_1} | {table\_s8\_1\_2} | {table\_s8\_1\_3} | {table\_s8\_1\_4} | {table\_s8\_1\_5} |
| Approximately  reproduced | {table\_s8\_2\_1} | {table\_s8\_2\_2} | {table\_s8\_2\_3} | {table\_s8\_2\_4} | {table\_s8\_2\_5} |
| Not reproduced | {table\_s8\_3\_1} | {table\_s8\_3\_2} | {table\_s8\_3\_3} | {table\_s8\_3\_4} | {table\_s8\_3\_5} |
| Total | {table\_s8\_4\_1} | {table\_s8\_4\_2} | {table\_s8\_4\_3} | {table\_s8\_4\_4} | {table\_s8\_4\_5} |

**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:** X²(6) = {table\_s8\_chi2}, p = {table\_s8\_chi2\_pval}

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

|  | **No policy** | **Data**  **required** | **Data and code**  **required** | **Data, code &**  **repro check required** | **Total** |
| --- | --- | --- | --- | --- | --- |
| Precisely  reproduced | {table\_s9\_1\_1} | {table\_s9\_1\_2} | {table\_s9\_1\_3} | {table\_s9\_1\_4} | {table\_s9\_1\_5} |
| Approximately  reproduced | {table\_s9\_2\_1} | {table\_s9\_2\_2} | {table\_s9\_2\_3} | {table\_s9\_2\_4} | {table\_s9\_2\_5} |
| Not reproduced | {table\_s9\_3\_1} | {table\_s9\_3\_2} | {table\_s9\_3\_3} | {table\_s9\_3\_4} | {table\_s9\_3\_5} |
| Total | {table\_s9\_4\_1} | {table\_s9\_4\_2} | {table\_s9\_4\_3} | {table\_s9\_4\_4} | {table\_s9\_4\_5} |

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

**Chi-squared test:** X²(6) = {table\_s9\_chi2}, p = {table\_s9\_chi2\_pval}

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 {table\_s10\_2\_3} times higher than an Approximate outcome if a journal has no policy (the baseline); however, if journals have any level of policy, the likelihood of being in a lower reproducibility outcome category decreases by around 70%. As the likelihood of a Precise or Approximate category is not significantly different (p = 0.120), this applies specifically to being less likely to have a Not Reproduced outcome than any other. This decrease in Not Reproduced outcome is 66% for a data requirement (p = 0.079), 70% for a data and code requirement (p = 0.003) and 67% for a data, code, and reproducibility check requirement although confidence is low for this category (p = 0.273) presumably due to very low case numbers (See Table S8).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

| **Term** | **Coeff.** | **Std. Error** | **Odds Ratio** | **p** |  |
| --- | --- | --- | --- | --- | --- |
| Precisely | Approximately | -0.33 | 0.21 | 0.72 | 0.120 |  |
| Approximately | Not Reproduced | 0.58 | 0.21 | 1.78 | 0.007 | \*\* |
| Data required | -1.09 | 0.62 | 0.34 | 0.079 | . |
| Data and code required | -1.19 | 0.40 | 0.30 | 0.003 | \*\* |
| Data, code, and reproducibility checks required | -1.12 | 1.02 | 0.33 | 0.273 |  |

Note: Weighted N=145. Significance levels: \*\*\* p < 0.001, \*\* p < 0.01, \* p < 0.05, . p < 0.1. Each claim outcome was weighted as a fraction of all outcomes for that paper.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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 is 3.57 times more likely than an Approximate outcome if a journal has no policy (the baseline); however, if journals have any level of policy, the likelihood of being in a lower reproducibility outcome category decreases by over 90%. As the likelihood of a Precise or Approximate category is not significantly different (p = 0.075), this applies specifically to being less likely to have a Not Reproduced outcome. This decrease in Not Reproduced outcome is 93% for a data requirement (p = 0.025), 94% for a data and code requirement (p < 0.001) and 91% for a data, code, and reproducibility check requirement although confidence is low for this category (p = 0.229) presumably due to very low case numbers (See Table S9).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

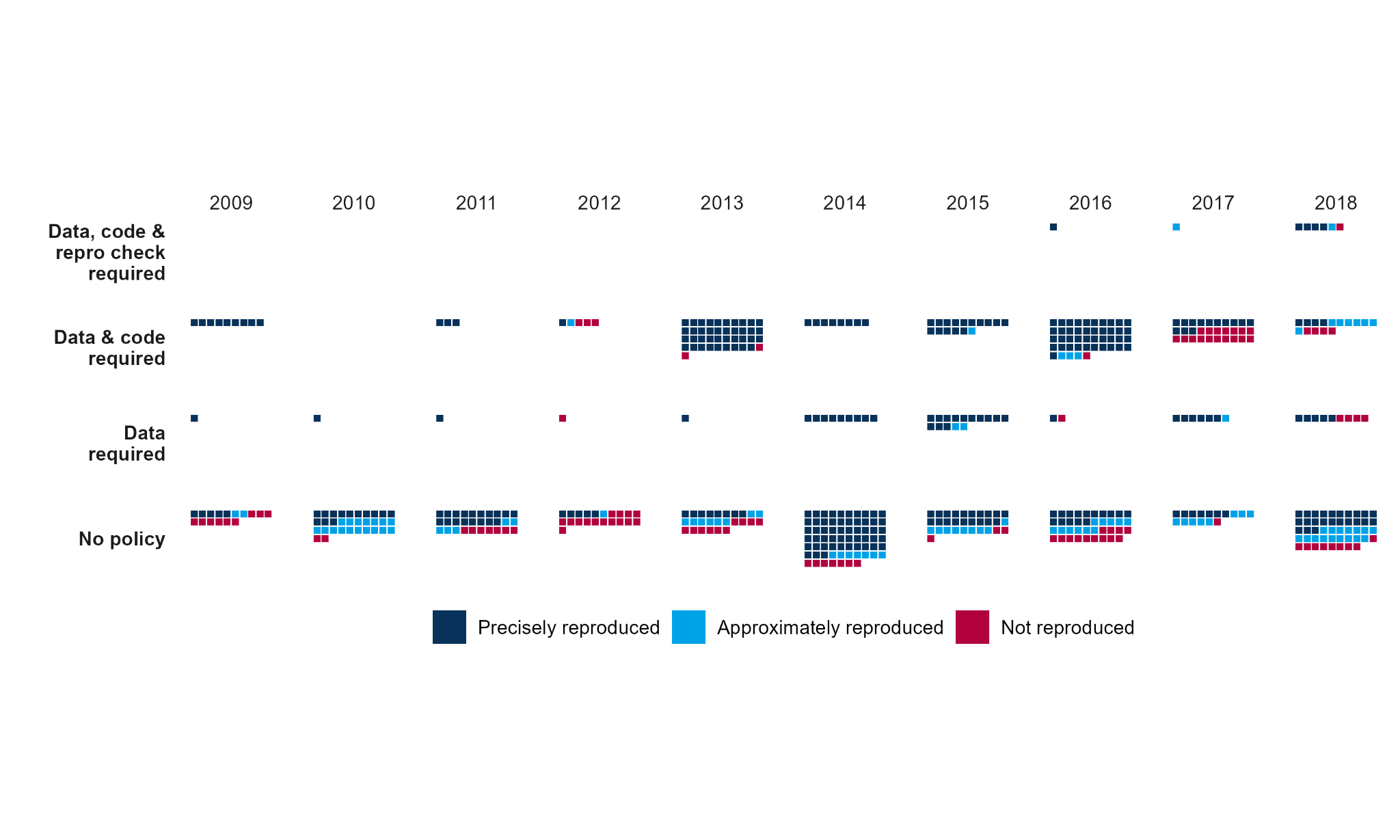
| **Term** | **Coeff.** | **Std. Error** | **Odds Ratio** | **p** |  |
| --- | --- | --- | --- | --- | --- |
| Precisely | Approximately | -0.71 | 0.40 | 0.49 | 0.075 | . |
| Approximately | Not Reproduced | 1.27 | 0.40 | 3.57 | 0.002 | \*\* |
| Data required | -2.67 | 1.19 | 0.07 | 0.025 | \* |
| Data and code required | -2.81 | 0.81 | 0.06 | 0.000 | \*\*\* |
| Data, code, and reproducibility checks required | -2.36 | 1.96 | 0.09 | 0.229 |  |

Note: N=553. Significance levels: \*\*\* p < 0.001, \*\* p < 0.01, \* p < 0.05, . p < 0.1. Includes random-intercepts at the paper-level (for N = 145 papers).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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% 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 outcome 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 outcome reproducibility, but further investigation is needed to provide clear evidence of a causal relationship.

## Process Reproducibility

A secondary research question concerned whether more stringent reproducibility policies were associated with greater success on process reproducibility. We conducted the same series of analyses as for process reproducibility on the sample of 600 papers that were subjected to process reproducibility 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 cross-tabs are presented in Table S12. Descriptively, papers achieved process reproducibility more frequently with any transparency policy (56.2%, 57.4%, and 100%) than with no policy (3.3%), with a significant chi-square test (p < 0.0001). This aligns with the definition of process reproducibility for this investigation: success if data were accessible for reproduction. The pattern would likely be different if we had adopted a more stringent definition of process reproducibility that included code.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### **Table S12. Process reproducibility outcomes by journal policy (N = 600)**

|  | **No policy** | **Data**  **required** | **Data and code**  **required** | **Data, code &**  **repro check required** | **Total** |
| --- | --- | --- | --- | --- | --- |
| Did not succeed | 494  (96.7%) | 7  (43.8%) | 29  (42.6%) | 0  (0.0%) | 530  (88.3%) |
| Succeeded | 17  (3.3%) | 9  (56.2%) | 39  (57.4%) | 5  (100.0%) | 70  (11.7%) |
| 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:** X²(3) = 240.93, p < 0.0001

For process reproducibility, a logistic regression reported in Table S13 predicts succeeded or did not succeed. For the baseline condition when a journal has no policy then reproduction of the procedure is 29 times (2,800%) more likely to not succeed (p < 0.001); however, when the journal has either a data sharing or data and code sharing requirement then the same unsuccessful outcome is 33.33 times (97%) less likely (p < 0.001 in both cases). Data, code and reproducibility check has no discernable effect because the sample size was very small and one outcome has 0 case numbers (see Table S12).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Table S13. Logistic Regression Model. Process Reproducibility.**

| **Term** | **Coeff.** | **Std. Error** | **Odds Ratio** | **p** |  |
| --- | --- | --- | --- | --- | --- |
| Succeeded | Did not succeed | 3.37 | 0.25 | 29.06 | 0.000 | \*\*\* |
| Data required | -3.62 | 0.56 | 0.03 | 0.000 | \*\*\* |
| Data and code required | -3.67 | 0.35 | 0.03 | 0.000 | \*\*\* |
| Data, code, and reproducibility checks required | -19.94 | 1073.11 | 0.00 | 0.985 |  |

Note: N=600. Significance levels: \*\*\* p < 0.001, \*\* p < 0.01, \* p < 0.05, . p < 0.1.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Journals with transparency policies had much higher rates of process reproducibility successes 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 outcome 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. … (2025). 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., … (2025). Investigating the replicability of the social and behavioral sciences.