ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

CC BY

Research

Cite this article: Krähmer D, Schächtele L, Auspurg K. 2026 Code sharing and reproducibility in survey-based social research: evidence from a large-scale audit. *R. Soc. Open Sci.* **13**: 251997.
https://doi.org/10.1098/rsos.251997

Received: 17 October 2025
Accepted: 6 January 2026

Subject Category:
Science, society and policy

Subject Areas:
behaviour, psychology, software

Keywords:
reproducibility, replication, code-sharing, meta-science, social sciences, research transparency, repeatability

Author for correspondence:
Daniel Krähmer
e-mail: daniel.kraehmer@soziologie.uni-muenchen.de

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8337305.

THE ROYAL SOCIETY PUBLISHING

Code sharing and reproducibility in survey-based social research: evidence from a large-scale audit

Daniel Krähmer, Laura Schächtele and Katrin Auspurg

Department of Sociology, LMU Munich, Munich, Germany

DK, 0000-0002-4100-5372; LS, 0009-0005-8627-9595; KA, 0000-0003-4504-0391

Reproducibility—the ability to obtain original results by reapplying the original analyses to the original data—is an essential component of empirical research. In this study, we assess the reproducibility of articles using the European Social Survey (ESS), a large-scale repeated cross-sectional dataset widely used across the social sciences. Drawing on more than 1000 ESS-based articles published between 2015 and 2020, we investigate whether authors share their code for reproduction purposes and whether published results are reproducible. We find that only about one in three authors (35%) share code. From the articles with code, we randomly selected 100 which reported 699 results. Of these 699 results, about half (51%) are numerically reproducible, while the others either fail (23%) or are different (26%). For those that are different, numerical deviations are usually minor and do not indicate systematic bias. Overall, about one in six published results (18%) is exactly reproducible. Reproducibility differs somewhat between disciplines, but reproducibility problems persist throughout. Reproducibility failure mostly stems from unavailable, poorly documented, or incomplete code. We propose low-cost measures for authors, editors, journals and data providers to improve code availability and reproducibility in large-*N* observational social research.

1. Introduction

When dropping an apple under the same conditions from the same height twice, it falls to the ground the same way both

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

times. Science should work the same way: applying identical methods to identical data under identical conditions (e.g. software environments) should produce identical outcomes. This principle, referred to as computational reproducibility[1] [5,6], is vital for empirical research. It is a prerequisite for any discussion about validity, generalizability and robustness since irreproducible findings cannot meaningfully be scrutinized or built upon [7,8]. Reproducibility thus serves as an entry-level criterion—a necessary but not sufficient condition—for research to contribute to the stock of scientific knowledge [9].

In this article, we present findings from a large-scale, systematic audit of social research based on survey data. Specifically, we audit more than 1000 articles published between 2015 and 2020 using data from the European Social Survey (ESS)—a large-scale, interdisciplinary survey conducted biennially in about 25 countries [10]. We investigate whether ESS analyses are open, i.e. whether authors share their code—a practical prerequisite for reproducibility with complex survey data [11]—and whether authors' original code reproduces their published results. In our audit, we answer four questions: (i) How open and reproducible is research using ESS data? (ii) Are discrepancies between original and reproduced results biased towards 'strong' and significant results? (iii) Do openness and reproducibility differ across disciplines? (iv) What prevents reproducibility?

Existing studies have investigated reproducibility in specific journals [12–19], disciplines [20–22] or research fields [23–25]. By shifting the focus to a specific *dataset*, we add a distinct perspective to the growing body of reproducibility literature. First, ESS studies exemplify the complexity inherent in secondary large-*N* data analysis. As a repeated cross-sectional dataset, the ESS confronts researchers with considerable flexibility in managing and analysing data, often requiring ambiguous decisions (e.g. in specifying variables and statistical models). Such flexibility makes survey analyses prone to error and hard to recreate based on an article alone, rendering code sharing and reproducibility checks particularly relevant [11]. Second, the ESS has a diverse user base. Between 2003 and 2023, it has been used in about 4000 international journal publications across a vast range of disciplines [26]. The dataset's broad use allows us to study reproducibility in disciplines that still lack systematic reproducibility evidence (e.g. sociology [27]), and enables us to describe disciplinary differences. Third, focusing on a single, publicly available dataset increases efficiency and standardization. Because all primary studies in our sample use a common dataset, we can follow a standardized reproduction protocol, maximizing internal consistency and enabling us to process a large number of studies. In fact, our audit is among the largest to date (see electronic supplementary material, table A1).

While ESS studies may not be representative of social science in general, they still offer an informative test case. The ESS has been used to study salient topics such as welfare sustainability, discrimination, integration, work–family dynamics, fertility decisions and public health, often directly informing public policy [26]. Given the ESS's academic and political impact, it is crucial that ESS studies meet basic metascientific quality requirements, including computational reproducibility. The ESS also provides an important methodological advantage for assessing reproducibility: because its data are publicly available, it arguably represents a best-case scenario for reproducibility as data availability is not an issue. If reproducibility rates are low even among studies using public data, this is particularly concerning.

Our audit proceeds in two steps: first, in our *code sharing audit*, we contact 1206 corresponding authors of ESS studies and request code for their article. This allows us to quantify code availability as a prerequisite for computational reproducibility. Second, we use authors' original code together with public ESS data to reproduce the main empirical findings of 100 randomly selected articles employing Stata-based multivariable regression analyses. These articles, which report 699 empirical main results, form the basis of our *reproducibility audit*.

At both stages, we uncover inefficiencies in the machinery of social science knowledge production. Even when authors share their materials, which is still the exception, there is only a roughly 50–50 chance that we can reproduce their original results. This pattern holds across social science disciplines working with ESS data, including those that have not yet attracted much metascientific attention. Irreproducibility stems about equally often from code execution problems (e.g. due to disorganized files, incomplete scripts or insufficient documentation) and different numerical estimates. Even when results are ultimately reproducible, they require substantial effort: instead of being 'push-button' reproducible, analyses often require sifting through authors' code, reorganizing files, deciphering opaque datasets and fixing coding errors. These hurdles turn straightforward verification into laborious reconstruction, hindering scientific progress, as 'time spent resolving non-replicability issues […] is time not spent expanding scientific understanding' [8]. Our audit thus corroborates the

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

impression that reproducibility is 'seldom straightforward, often utterly frustrating, and, for many articles, impossible' [22].

Despite the problems our audit uncovers, we remain cautiously optimistic. Most deviations between original and reproduced results in our study are minor and unlikely to alter substantive conclusions. Furthermore, reproducibility seems fixable. Many reproducibility barriers we encountered could have been avoided with marginally more effort in documenting and structuring original code. We propose several low-cost, practical interventions that could substantially improve reproducibility without unduly burdening authors, editors, journals or data providers.

2. Literature review

Researchers must be able to understand and inspect each others' work to ensure the cumulative advancement of science [8]. In observational research, this is not straightforward. Analysts make hundreds of decisions when preparing data and specifying models—a concept known as the 'garden of forking paths' [28]. Choices about case selection, controls, weighting, imputation and outlier treatment can substantially alter both the derived dataset and the model specification, and hence the reported findings [2]. Because it is infeasible to document every step of the analysis in a conventional methods section, much of the workflow usually remains opaque [29]. Consequently, even when two researchers interrogate the same dataset with the same research question, divergent choices can yield different datasets and results [30,31].

Open code provides a map through this garden of forking paths. When code is available, researchers no longer have to rely solely on verbally reported workflows but can scrutinize the work of others through reproductions and replications. Without code, however, reproductions are time-consuming [17,22] and uncertain: if reproduced findings differ from the originals, it is unclear whether discrepancies stem from different analytic choices or inaccurate reporting. Such uncertainty can lead to unproductive debate and academic stalemates [1,32]. In theory, open materials should enable data reuse, foster reproduction and provide clarity about the provenance of scientific evidence, thereby strengthening the credibility of science [8].

In practice, however, the extent to which reproduction materials are openly available remains unclear. Prior studies have shown that post-publication requests for data and code are often ignored [33,34], and that the availability of materials tends to deteriorate over time due to broken links, misplaced files or abandoned e-mail accounts [35,36]. In response to these challenges, some disciplines, most notably political science and economics, have begun adopting mandatory open science policies, with leading journals even employing dedicated data editors to check reproducibility prior to publication. Yet, whether such policies have shifted norms and practices towards open science in social research more broadly remains unknown.

What is more: even when data and code are accessible, reproductions may still fail. Reasons include missing or inconsistently labelled datasets [22], unclear hard- and software requirements [13,21], inadequate documentation [13,37], rounding errors [12,22] and misreporting [38].

In figure 1, we summarize the existing evidence on reproducibility, listing prior audits (y-axis) and their success rates (x-axis). Figure 1A visualizes *overall* reproducibility rates, i.e. reproducibility starting from the published articles, counting cases of missing code and data as instances of reproduction failure. Figure 1B displays *conditional* reproducibility rates, i.e. the share of reproducible articles when all relevant code and data are available.

As figure 1 illustrates, reproducibility is far from perfect. In figure 1A, almost all reproducibility estimates fall below 50%—indicating that more than half of the scrutinized research is irreproducible by auditors. The box plot ('Total') shows that three-quarters of studies report overall reproducibility rates below 33%, with a median reproducibility rate of only 15%. In figure 1B, which shows conditional reproducibility rates, estimates are naturally higher. Still, even when data and code are available, reproductions frequently fail (interquartile reproducibility range: 30% to 82%; median: 56%). Taken together, this evidence suggests that non-availability of materials is a major source of reproduction failure, but clearly not the only one. Even the most optimistic studies report reproducibility rates well below 100% [14,38,39].

Although these findings demonstrate that reproducibility is imperfect, they allow few other generalizations. Several audits have examined reproducibility in journals that introduced open science policies [13,14,29] or employ data editors who conduct pre-publication reproducibility checks [38]. As others have noted, these journals may attract submissions from authors whose workflow is already

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

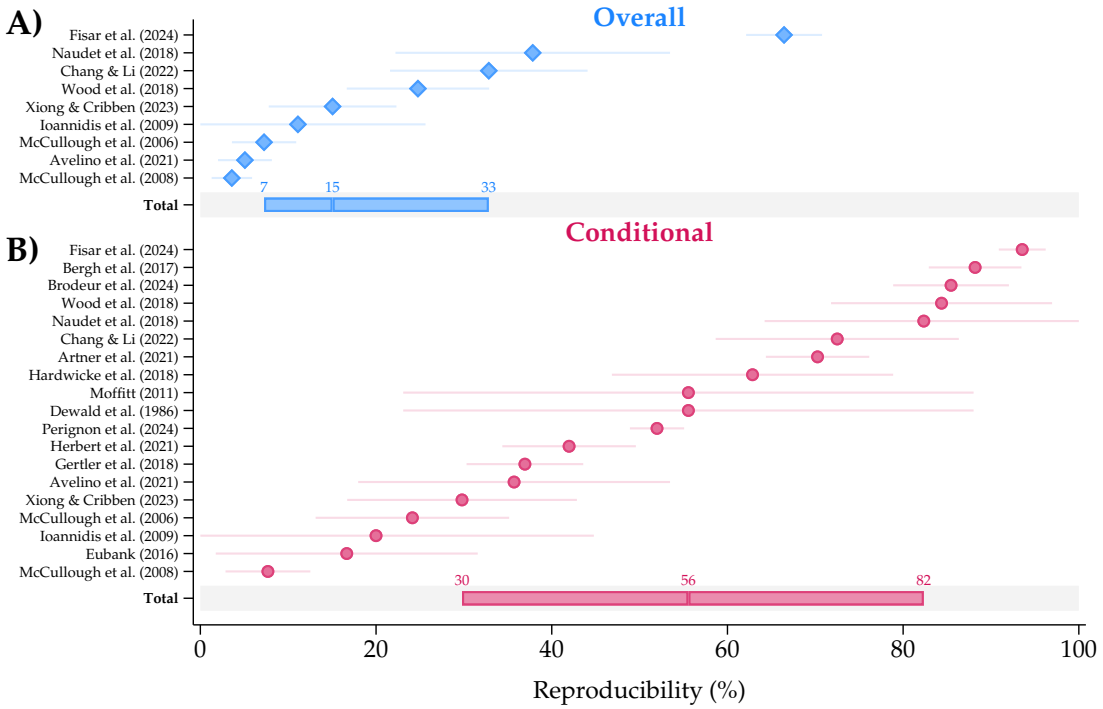


Figure 1. Reproducibility rates as reported by previous reproducibility audits. Reproducibility estimates with 95% confidence intervals. *Overall* rates in (A) include reproduction failure due to missing code or data. *Conditional* rates in (B) reflect reproducibility when all relevant code and data are available. 'Total' rows display central tendencies via box plots that visualize the median and interquartile range of reproducibility estimates. Most audits report reproducibility rates at the article level. Detailed information on individual studies [12–16,18–25,29,37–41] is provided in electronic supplementary material, table A1. An interactive version of this figure is shared as a separate supplement.

more open and reproducible, leading to overly optimistic conclusions [13,14,29]. By contrast, studies may underestimate reproducibility when the identification of replication targets is based on community nominations [42,43], delegated to replication crowds [44] or selected without any sampling frame at all, leading to isolated commentary pieces [45–48]. If researchers target results they expect to fail—anticipating that failed reproductions and replications are more publishable [49–51]—success rates will reflect a selective, overly pessimistic sample.

An aspect many reproducibility audits have in common is a narrow disciplinary focus. Most audits have investigated reproducibility in economics [15,16,18–21,41] and/or political science [12,38,40]. For other social sciences, such as sociology, no systematic reproducibility evidence has been published to date.[2]

Moreover, the reproducibility literature is rife with definitional and procedural differences [53]. In defining what constitutes a successful reproduction, some audits demand numerical identity between original and reproduced results [16,22]; others tolerate small deviations [29]; still others focus on direction or significance [21,40]; and some apply composite scores [18,37,38]. Regarding procedures, some audits request materials from the original authors [16,21,22,40]; others restrict themselves to publicly available files [13,29,38]; some involve original authors [17,38]; others do not [15,19,25]; some rely on small teams [12,21,22,43]; others on crowdsourcing [13,37,38]. This large diversity in research designs (for an overview, see electronic supplementary material, table A1) paired with the considerable variation in results underscores how difficult it is to compare and generalize the results of individual reproducibility audits.

3. The current study

The current study assesses reproducibility among articles using the ESS, a popular dataset for cross-sectional and trend analyses in the social sciences [26]. The ESS is a biennial survey conducted since 2002 in around 25 European countries using 1-h face-to-face interviews. Each wave includes hundreds

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

of variables from more than 40 000 respondents, totalling over 500 000 respondents across all waves (see [54] for methodological details). Like other cross-country trend studies, the ESS is publicly funded and designed to support both scientific research and evidence-based policymaking. Datasets like the ESS offer sample sizes that individual researchers could not feasibly collect on their own and therefore form the basis of hundreds of secondary analyses and publications every year [26]. Because the ESS serves diverse research fields, it comes as a multi-topic survey, requiring researchers to undertake extensive preprocessing steps to extract relevant samples and variables and to specify appropriate measurements and model specifications.

The ESS provides a useful sampling frame for investigating reproducibility in the social sciences for several reasons. First, ESS-based studies exemplify the inherent analytical flexibility of secondary large-*N* data analysis—a flexibility that makes reproductions without code largely infeasible and renders code sharing particularly important for reproducibility checks [11]. Researchers working with such data face complex data-management tasks (e.g. selecting, cleaning, and recoding variables or defining country subsamples) and analytical decisions (e.g. choosing model specifications, handling weights or dealing with missing data), all of which require substantial coding effort and are difficult to capture in verbal prose [29]. Because secondary analyses of large-*N* survey data, unlike experiments, are rarely preregistered [55], there is usually no formal record besides code that would enable others to follow and reproduce the original analytical decisions.

Second, articles using the ESS span diverse disciplines, including many that have received little or no attention in prior reproducibility audits. In line with scholars' calls for more diverse reproducibility samples [38,39], our audit provides first evidence for several disciplines for which no reproducibility audit has been conducted so far. Cross-discipline comparison is aided by the fact that we do not mix observational studies with experiments, which are typically easier to reproduce [11], and that we circumvent data availability issues that may be more prevalent in some disciplines than in others.

Third, starting from one common dataset has practical advantages. Most notably, it increases standardization and scalability. Because all studies in our sample use the same data, our audit follows a highly standardized reproduction protocol. This protocol proved essential, as preparatory analyses revealed that even in seemingly straightforward reproductions many complex decisions arise—such as whether to correct coding errors, how much runtime to allow or which results to select as reproduction targets—that potentially influence results[3] [3]. Using a standardized protocol and conducting our audit in a dedicated team of four (the three authors and one research assistant) allowed close control over the research process and maximized internal consistency [43]. At the same time, standardization allowed processing a large number of studies. While previous reproducibility audits have often covered fewer than 30 papers [12,16,18,23], our audit examines more than 1000 articles for code sharing and 699 results from 100 articles for reproducibility—placing it among the most comprehensive in the field [13,15,20,38].

To our knowledge, only one other metascientific audit has focused on a specific dataset [56]. That study investigated code availability among users of the German Socio-Economic Panel (GSOEP) but did not examine reproducibility based on the original code. We consider this second step crucial as large-*N* observational studies are notoriously complex and easily compromised by erroneous code.

4. Methods

4.1. Code sharing audit: target sample and procedure

Because openness is a practical prerequisite for reproducibility, we first attempted to obtain code for as many ESS articles as possible. In our *code sharing audit*, we targeted research articles published based on ESS data between 2015 and 2020. We excluded older articles to ensure that authors could still reasonably be expected to have access to their code files, and excluded more recent articles as the ESS Bibliography is continuously updated and usually incomplete for the latest period due to lags [26].

We drew a full sample of corresponding authors from the official ESS Bibliography (https://bibliography.europeansocialsurvey.org/). When authors were listed as corresponding authors for multiple publications in our time frame, we randomly selected one. This procedure was in line with recommendations by our institutional ethics committee and sought to minimize respondent burden by avoiding repeated requests.[4] After removing 22 duplicate records and 7 articles without full-text access, we identified 1206 unique journal articles suitable for our audit (see electronic supplementary material, figure A1, for details on our sampling procedure and case numbers).

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

Table 1. Sample descriptives.

	n	min.	max.	median	mean	s.d.
net code sharing sample (*n* = 1123)						
publication year	1123	2015	2020	2018	2018	1.68
no. of academic citations	1107	0	1499	19	40	78.32
discipline	1123					
political science	273					
economics	139					
sociology	131					
health	100					
psychology	70					
demography	55					
business and management	44					
social sciences (interdisciplinary)	87					
other	224					
reproducibility sample (*n* = 100)						
publication year	100	2015	2020	2018	2018	1.65
no. of academic citations	100	0	221	22	37	38.58
discipline	100					
political science	36					
economics	20					
sociology	15					
health	4					
psychology	3					
demography	7					
business and management	1					
social sciences (interdisciplinary)	6					
other	8					

Citation counts from Open Alex [58]; discipline classification based on the Web of Science (see electronic supplementary material, table A2).

We contacted the corresponding authors of these 1206 articles on 6 and 7 July 2022 and sent up to three reminders in case of non-response.[5] Our request asked for access to all preprocessing and analysis code (e.g. Stata, R or SPSS scripts) to reproduce the original findings from raw ESS data. We also asked authors to share relevant additional materials (e.g. supplementary external data), while explicitly noting that they did *not* need to share the raw ESS data themselves. This clarification served to prevent privacy-related concerns and allowed us to measure code sharing independently of data-sensitivity issues. Of the 1206 articles, 83 were found to be sampling-neutral overcoverage, most of them because they did not substantively use ESS data (n = 59).[6] Excluding these cases leaves a net sample of 1123 articles. Table 1 provides descriptive information on this sample.

To test whether low-cost interventions could increase code sharing, we embedded a field experiment in our code requests, randomly varying our message's framing (negative versus positive), the highlighted benefits of code sharing (e.g. citation gains) and the expected effort (e.g. whether code cleaning was required). Results from this experiment are reported elsewhere [57]. In brief, variations in wording had negligible effects, with code sharing differences of at most 6.9 percentage points across treatments, and only one out of the three stimuli showing a statistically significant effect (contrary to our pre-registered hypothesis). For our reproducibility audit, this implies good generalizability: articles with code are distributed across all treatment conditions, and do not come from one specifically

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

worded code request. We can therefore ignore the experimental component of our code request from here onward, focusing on the barriers to code sharing and, ultimately, reproducibility.

A central step in our code sharing audit was defining what counted as 'code sharing'. This proved non-trivial. Some authors shared fragmentary or unrelated code; others provided related files (e.g. datasets) but no relevant code. We conducted initial plausibility checks to assess whether the shared materials could in principle be suitable for reproduction (i.e. whether they included at least one code file). When packages were clearly incomplete, we recontacted authors, pointed out the problems with their materials, and asked them to complete them. When they stated that no further materials were available, we classified them as non-sharing. Our later reproducibility audit revealed that some authors provided code that ran but did not start from ESS raw data (e.g. analysis scripts that executed only a few regressions on a pre-processed dataset). Because data preprocessing (variable recoding, cleaning and sample definition) involves substantial researcher discretion [31], such incomplete code restricts transparency. Nevertheless, we adopted a conservative approach: all cases where *some* code was shared were counted as instances of 'code sharing'. We provide more details on code completeness in §5.5.

4.2. Reproducibility audit: target sample and procedure

Using the code researchers sent us, we attempted to reproduce the findings of 100 articles. These 100 articles were randomly selected from all articles that met three criteria. First, articles had to substantively rely on ESS data, meaning that we excluded articles that made only cursory use of the ESS (e.g. for robustness checks). Second, analyses had to be conducted in Stata. This ensured a common computational framework and allowed us to abstract from software versioning issues, as Stata is explicitly designed with reproducibility and backward compatibility in mind [59]. Moreover, all members of our research team were proficient in Stata, minimizing the risk that reproducibility failure might arise from our own technical limitations rather than authors' coding practices. Third, articles had to use multivariable regression analysis—the workhorse of observational survey research. We excluded purely descriptive studies, as these often reported hundreds of estimates of trends or group differences, making it difficult to define any clear reproduction target.

We assume that focusing on multivariable regression analyses in Stata does not introduce serious selectivity compared with the pool of ESS studies that shared code. To our knowledge, no prior research has systematically linked different software to reproducibility (for one exception based on a selective sample, see [60]), and attempting to reproduce studies in unfamiliar programming languages (e.g. Mplus, SAS) would likely have posed a far greater threat to the validity of our findings—particularly when authors combined several programming languages [61]. In addition, holding the software environment constant helped to ensure that reproducibility differences across disciplines were not due to software-related factors.

We also assessed empirically whether our sampling criteria introduced bias. Regarding key contextual variables (e.g. articles' publication year, citations, disciplines) they did not. When comparing the composition of our sample before and after applying the selection criteria, we find that differences along these contextual factors are mostly minor, unsystematic and statistically insignificant (see electronic supplementary material, figures A2–A4). Thus, we remain confident that our reproducibility sample offers a reasonable representation of ESS studies. For descriptive information about our reproducibility sample, see the bottom half of table 1.

4.3. Defining and conducting reproductions

Beyond sampling, we needed to decide *which* parts of an article to reproduce, *how* to reproduce them and *what* to consider a successful reproduction.

Regarding our reproduction targets (*which*), we focused on an article's 'main claims'. We defined a main claim as any statement in the abstract referring to the article's empirical findings. This approach ensured that claims were anchored in what original authors themselves highlighted (see also [21]). Articles could, and often did, contain multiple claims, all of which we recorded. In total, we identified 272 claims—an average of 2.7 per paper (see table 2).

Because claims were often broad, we identified more specific reproduction targets in the form of empirical results. We defined a result as any numerical quantity, typically a regression coefficient, that provided statistical evidence for a broader claim (for a similar distinction between claims and results, see [62]). When claims were based on multiple results, we recorded all. In total, we identified

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

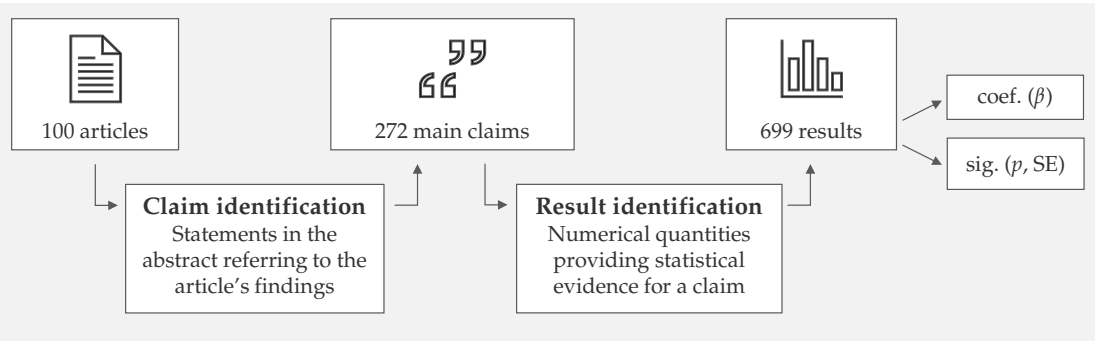


Figure 2. Identification of main claims and results. Visualization of the identification process for claims and results. An article could feature multiple claims, and a claim could be based on multiple results. Results usually consisted of two components: a coefficient (β or a comparable point estimate) and its significance (p-value, standard error). If the latter information was missing, we recorded only the coefficient.

Table 2. Claims and results from 100 articles in the reproducibility sample.

	min.	max.	median	mean	s.d.
claims ($n = 272$)					
— claims per paper	1	8	2	2.7	1.7
results ($n = 699$)					
— results per claim	1	16	2	2.6	2.3
— results per paper	1	29	5	7.0	5.9

699 results—an average of 2.6 per claim and of 7.0 per paper (see table 2). For each result, we recorded its coefficient (β or a comparable point estimate) and, when available, information on statistical significance (e.g. p-value, standard error, significance stars). Figure 2 provides a conceptual overview of the identification process for claims and results.

Identifying appropriate reproduction targets was surprisingly challenging. Articles in our sample frequently presented imprecise claims with no explicit link to specific results or without numerical results at all. Sometimes, authors concealed contradictory results behind seemingly unambiguous claims, or reported ad hoc claims without presenting the corresponding statistical tests. For instance, one article claimed a moderation effect but failed to estimate the significance of the relevant interaction term. Another failed to provide the appropriate reference category in their regression to support the main claim. Such practices make reproduction cumbersome. In our audit, we addressed this challenge through extensive crosschecking within our team: whenever the coding of a claim or result was uncertain, we applied a 'many-eyes' review principle that helped to maintain a high degree of process consistency [43], eliminating idiosyncratic researcher decisions on our side.

For the reproduction process (*how*), we followed a strict protocol. We relied exclusively on the materials provided by the original authors and executed their code with only minor bug fixes, such as installing missing user-written packages or adjusting file paths. We consider these to be reasonable fixes that any good-faith replicator would implement. To err on the side of caution, we even removed code passages that were problematic but clearly irrelevant for the main results (e.g. hard-coded comments that prevented execution). More substantial errors, however, were left untouched; if they blocked code execution, we classified the corresponding results as failed.

We deliberately refrained from contacting original authors, as this often leads to prolonged back-and-forth without meaningfully improving reproducibility [21,29]. Moreover, we share the view of many scholars [9,19,22,23] that reproducibility should not depend on personal communication with the original authors. Otherwise, the ability to reproduce results would be biased towards more accessible researchers—those still active in academia or from more recent cohorts.

Regarding the evaluation of reproduction success (*what*), we combine several criteria to balance the strengths and weaknesses of existing metrics. Specifically, we compare effect sizes, confidence intervals and statistical significance. This multifaceted approach avoids overreliance on any single indicator and follows recent recommendations [53,63]. For effect sizes, we apply both a strict comparison (numerical

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

identity) and a more lenient criterion (±10% deviation). For confidence intervals, we check whether the reproduced estimate falls within the original's 95% confidence interval [53,63]. For statistical significance, we compare reproduced and original *p*-values (often calculated from the reported standard errors) and assess whether both estimates fall on the same side of the conventional 5% threshold. Taken together, this multi-metric approach bolsters a comprehensive reproducibility assessment.

During our reproducibility audit, we encountered two unforeseen practical challenges. First, the ESS allows users to create customized data extracts via an online query system, generating unique datasets without explicit versioning.[7] While convenient for original researchers, this system hampers reproducibility as data extracts are hard to identify and not archived (see [64] for a similar critique). Second, the ESS hosts only the latest edition of each survey wave on its website publicly. For instance, if a study used edition 1.0 of survey wave 7, but edition 2.0 was later released, the original dataset would no longer be publicly accessible. While designed to prevent the spread of data errors, this release policy complicates reproduction. Although the ESS team kindly provided us with all historical data editions upon request, our experience highlights the importance of proper data versioning and citation practices—two issues we revisit in the discussion.

4.4. Exploring differences across disciplines

Finally, the broad use of ESS data allows us to investigate differences across disciplines. Are articles from disciplines that fall outside the scope of prior audits (e.g. sociology) less open and reproducible than articles from disciplines in the metascientific spotlight (e.g. economics, political science)? We believe that cross-discipline description can meaningfully advance reproducibility efforts. If certain disciplines exhibit substantially lower reproducibility rates, targeted interventions in these disciplines might be especially warranted. Given our sample of 100 articles, we focus on broad disciplinary categories based on Web of Science subject categories (see electronic supplementary material, table A2). We employ simple bivariate tests of proportions to quantify differences across disciplines. We discuss the limitations of this approach in §6.2.

5. Results

5.1. Code sharing audit

Of the 1123 articles in our net code sharing sample, 385 authors (34%) shared code upon request.[8] Including 11 authors who did not respond to our request but provided materials online,[9] we obtained code for 396 articles (35%), most of which had previously not been public (n = 359). This reveals an important first finding: code availability in the social sciences remains the exception rather than the rule.

When code was shared, the apparent quality of the materials varied greatly. Some packages included extensive README files and clearly numbered scripts, while others consisted of a disorganized assortment of files with little or no guidance. In extreme cases, authors sent dozens or even hundreds of files for a single article without providing any guidance on how to navigate them.

The 727 articles for which no code was shared fall into three categories (see figure 3A). First, 92 authors were unreachable, often due to changes of institutions, inactive e-mail accounts, long-term leave or exit from academia. Initially, this number was higher (n = 221), but we were able to track some authors through manual Web searches—an obviously inefficient process. Second, 439 authors remained silent. They, as far as we can tell, received our e-mails but never replied. Whether they overlooked or deliberately ignored our request,[10] the outcome was the same: no code. This high non-response rate is largely in line with previous research on post-publication requests for replication material [6,33,34]. Third, 196 authors replied but were either unwilling or unable to share. Some asked for additional information or promised to share code later but never delivered (n = 64), while others stated upfront that code sharing was impossible (n = 132).

The responses of authors who replied but did not share shed light on common barriers to code sharing (see figure 3B). Most had lost their code (n = 62), either because they misplaced it or switched institutions, with several noting that the code had been written 'a long time ago' (n = 16). Surprisingly many (n = 29) reported having never created permanent code in the first place, explaining that their 'sequential research practice' did not require it, that they used non-code-based software, that they had not saved their code, or that they did not bother to keep records 'when it is just a regression-based

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

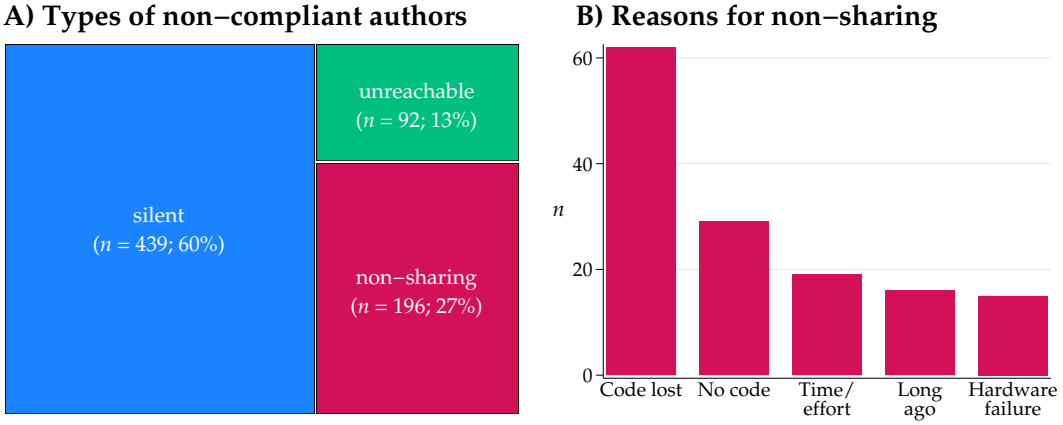


Figure 3. Results of our code sharing audit. Panel (A) distinguishes non-compliant authors by their (non-)response using a treemap [65]. Panel (B) zooms in on authors who were responsive but non-sharing, plotting their most common excuses (multiple mentions possible). Counts in (B) differ slightly from those reported elsewhere [57] because responses were recoded for this analysis.

methodology'. Others mentioned time constraints (n = 19) and hardware failure (n = 15). Collectively, these responses reveal poor code writing and archiving practices among social scientists.

Encouragingly, most authors were constructive. Several researchers invested hours searching through old files to satisfy our request—often without success—and some even offered to rewrite their analysis code.[11] While these efforts are commendable, they underscore the severe inefficiencies caused by missing or poorly archived research code.

Few researchers were openly opposed to code sharing. They questioned whether our request 'was justified', arguing that 'there is nothing magical' about their code and that 'anybody around the world must be able to achieve the same results, the decimal points after the comma included'. The view that reproduction should be possible based solely on the published article was held by 28 researchers, even if they were otherwise supportive of our request. This view contrasts sharply with metascientists' experience that conducting reproductions without code is like 'assembling flat pack furniture without an instruction booklet' [29].[12] One author's response from our sample illustrates this tension perfectly, claiming that 'the results can be reproduced with the information provided in the paper' while admitting that 'this is clearly not the easiest since the paper included hundreds of models'.

In sum, our code sharing audit shows that reproduction materials are available for only about one in three articles, even when data privacy concerns do not apply. Authors are difficult to reach, ignore requests, or are unable to comply because of poor code-writing, storing, and archiving practices. Moreover, the apparent quality of the shared materials varies greatly, raising an important follow-up question: is the shared code suitable for reproducing the published results?

5.2. Reproducibility audit

5.2.1. Are results reproducible?

Our reproducibility audit focused on 100 articles that contained a total of 699 results. Of these, 355 (51%) were exactly reproducible using authors' own code and publicly available ESS data when judged by the coefficient. One hundred and eighty-two (26%) differed from the originals, 83 of them within a ±10% bandwidth, and 162 (23%) completely failed. In other words, running the original code on the original data produced the published result only about half of the time. Using a more lenient criterion, about two-thirds of results (63%) fall within a ±10% bandwidth around the original. Figure 4A visualizes these findings.

When a reproduction *failed*, this meant that applying the original code to the data did not yield *any* output. The reasons for such failure were manifold (see figure 4B): 42 results were based on some preprocessed version of the ESS that was not provided and could not be recreated because authors did not provide the relevant cleaning code—despite our explicit request—and running their analyses on the raw data generated error messages (e.g. due to missing variables); 32 results failed because the relevant statistical tests were not implemented in the code, presumably because authors shared a preliminary version; 25 relied on additional context data that were inadequately documented and thus

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

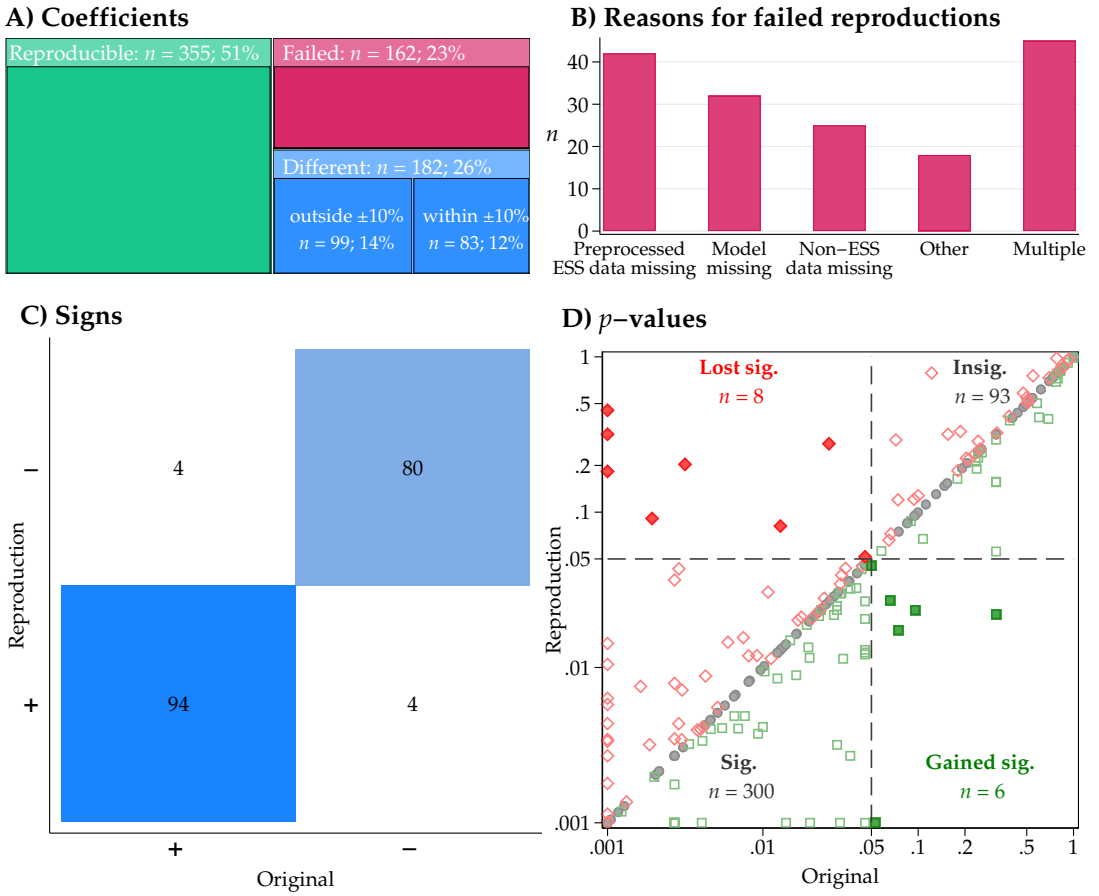


Figure 4. Results of our reproducibility audit. Panel (A) classifies reproduction success by comparing original and reproduced coefficient values ($n = 699$) using a treemap [65]. Panel (B) zooms in on failed reproductions, plotting the most common causes for failure ($n = 162$). Panel (C) focuses on reproductions that yielded different results, comparing original and reproduced coefficient signs ($n = 182$). Panel (D) plots the p-values of original and reproduced results when they were reported or could be calculated from the standard error ($n = 407$). Overall, reproducibility is clearly imperfect, although differences between original and reproduced estimates are generally minor.

impossible for us to obtain; 18 failed for other reasons, including packages too chaotic to be operated; 45 suffered from multiple problems.

Importantly, we did not judge irreproducibility prematurely. Had we required results to be *push-button reproducible*, i.e. reproducible without any modifications, our success rate would have been near zero as almost all packages required some intervention. These ranged from the mundane (e.g. changing a file path) to the considerable (e.g. installing missing user-written packages[13]) to the severe (e.g. rearranging entire code sections). When in doubt, we were deliberately generous rather than restrictive. For instance, when datasets were ambiguously referenced in the code, we tried several ESS waves and editions that seemed plausible based on information from the article. When code generated hundreds of pages of output, we tried to identify the model that corresponded to the result presented in the article. When code caused error messages, we even commented or reorganized problematic parts to facilitate reproduction. Hence, we believe that our estimates represent conservative estimates of reproducibility problems.

What if reproduction did not fail, but results were also not identical to the original (classified as 'different')? To judge the severity of such deviations, we consider signs and statistical significance. On both accounts, we see large overlap between original and reproduced results. For signs, 96% of results ($n = 174$) matched their original while only eight did not (see figure 4C). In most cases of sign flips, the original and/or reproduced result was not significant at the 5% level (only in one case did a significant negative result flip to a significant positive one). For significance, 97% of results ($n = 393$) matched their original under the 5% threshold while only 14 did not (see figure 4D). Among results that switched significance, slightly more lost than gained significance (8 versus 6). Although p-values did vary—not

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

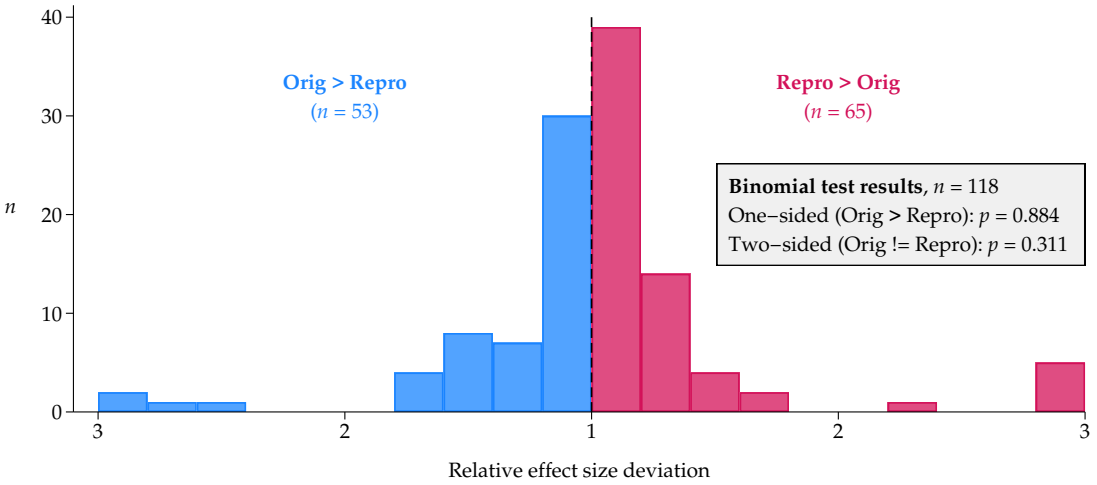


Figure 5. Relative effect size deviations. Distribution of relative effect size deviations when original and reproduction differ. Original null results ($n = 44$) and results without information on significance ($n = 20$) are excluded in line with recent literature [63] (see electronic supplementary material, figure A5, for a robustness check that includes these estimates). Numbers on the *x*-axis indicate the factor by which results are larger. For example, the number 2 on the left means that the original effect size was twice as large as the reproduced one. Colours indicate whether values fall below or above 1, i.e. whether the original or the reproduced coefficient is larger. Extreme values (>3) are winsorized for plotting ($n = 12$).

all points in figure 4D fall onto the diagonal—these deviations were mostly inconsequential when judged by a binary significance criterion. Furthermore, 96% of reproduced results ($n = 377$) fell within the 95% confidence interval of their original estimate.[14]

In sum, we find that around half of the coefficients in our sample are exactly reproducible using authors' original code, while the other half are either different (26%) or fail (23%). If an output could be produced, results usually matched the originals in sign and/or statistical significance. Because hypotheses in the social sciences are often evaluated based on these parameters, the deviations we find may not call the authors' main claims into question. Still, differences in effect sizes can be problematic as they matter for publication chances [66] and policy interventions [67]. We proceed by comparing original and reproduced effect sizes.

5.2.2. Are effect size deviations systematic?

Discrepancies between original and reproduced results may reflect random noise or systematic bias. While random deviations 'only' reduce the efficiency of cumulative research [8], systematic deviations—such as effect sizes skewed towards larger or 'stronger' results—inflate the rate of false positives. Systematic bias may arise from selectivity in error detection [62]. When authors, consciously or unconsciously, check 'strong' results confirming their expectations less carefully, errors that *inflate* effect sizes (e.g. from ad hoc, undocumented coding decisions) may persist into publication.

To test whether deviations are systematic, we visually compare original and reproduced findings. Figure 5 displays the distribution of relative effect size deviations, i.e. the ratio of original and reproduced coefficients. Values near 1 (i.e. the centre) indicate that the two estimates are consistent; values <1 (on the left) imply that original effect sizes are larger (as expected under a bias towards reporting 'stronger' results); and values >1 (on the right) indicate that reproduced effects are larger. Overall, figure 5 shows no signs of systematic bias. The shape of the distribution appears mostly symmetric, with deviations equally likely to fall below and above 1. A binomial test supports this visual impression, indicating that the deviations are statistically indistinguishable from random variation (see box in figure 5). Thus, we find no evidence that imperfectly reproducible effect sizes were systematically inflated in original publications.

5.3. From results to articles: reproducibility at the article level

So far, we have established that about one in three articles comes with code, and that about every second result is reproducible using the original code. But how is reproducibility distributed at the

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

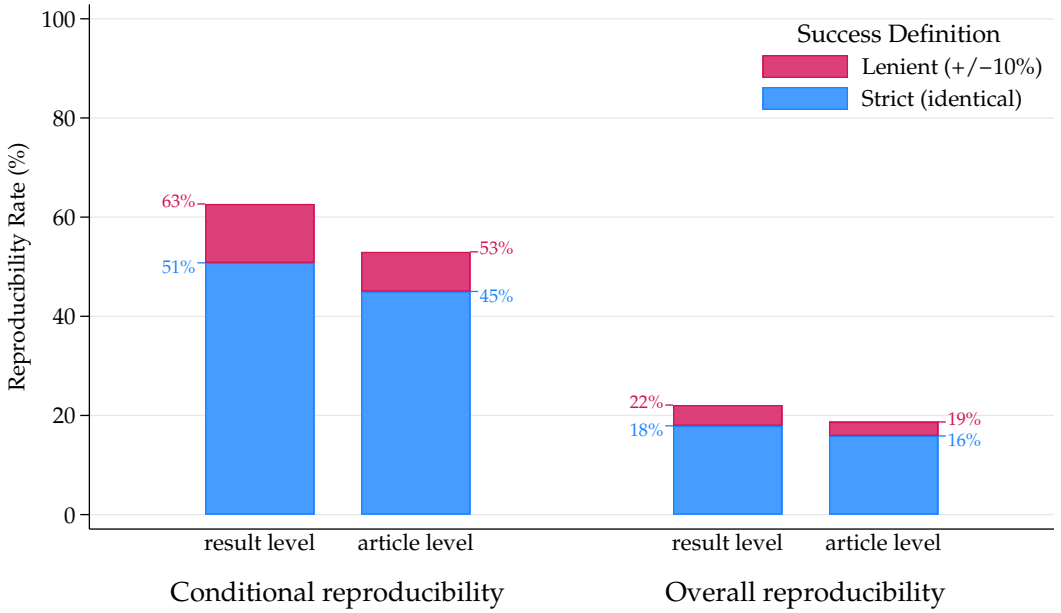


Figure 6. Reproducibility rates on result level and article level. Stacked bars summarizing conditional and overall reproducibility rates under lenient and strict success definitions at the result and article levels. Overall reproducibility rates, which combine both stages, are calculated by multiplying the conditional reproducibility rate by the overall code sharing rate of 35%.

Table 3. Reproducibility at the article level by stringency and scope.

		stringency	
		strict (numerically identical)	lenient (± 10%)
scope	all results	45	53
	>50% results	53	63

This table is based on the 100 articles from our reproducibility audit, so cell values can be interpreted as percentages.

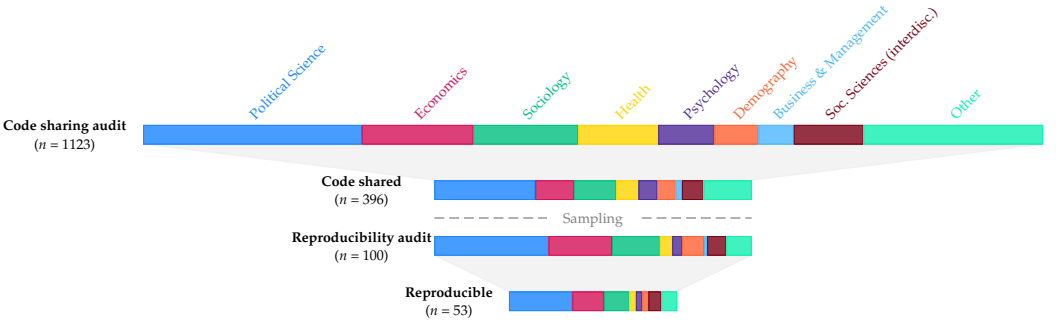
article level? Do irreproducible results cluster? Table 3 sheds light on this question, using four possible aggregation rules to calculate reproducibility at the article level.

Under a strict criterion (i.e. numerical identity), 45 of the 100 audited articles are reproducible *across all main results*, while 53 are reproducible regarding *more than half of their main results*. Tolerating minor numerical deviations (± 10%) increases these numbers to 53 and 63, respectively. Hence, depending on the scope and stringency of the aggregation rule, between 45% and 63% of articles are reproducible. For the following analyses, we define reproducibility at article level as *all* results being reproducible *within a* ± 10% *margin* (i.e. the value in the upper right corner of table 3). Our rationale is that minor differences (± 10%) are unlikely to alter articles' substantive conclusions, but one might expect *all* main claims of an article to be reproducible. Based on this definition, reproduction at the article level succeeds also only around half of the time (53%).

Figure 6 provides an alternative visual summary of our results, contrasting result level and article level reproducibility rates. The bar on the far left reiterates what we previously presented in §5.2.1: conditional on code being available, 63% of results are reproducible within a ± 10% bandwidth (51% exactly reproducible). At the article level, this rate is 53% (45% exactly reproducible; see also table 3). The slightly *lower* reproducibility rate at the article level indicates that irreproducibility is widespread rather than concentrated in a few specific articles.

The bars on the right extend the distinction between the result and article level to *overall reproducibility* rates. To estimate overall reproducibility, we multiply conditional reproducibility rates from our reproducibility audit with the code sharing rate (35%) from our code sharing audit. At the result level, this reveals that approximately one in six results (51% × 35% = 18%) are exactly reproducible and approximately one in five (63% × 35% = 22%) are reproducible under the more lenient definition. These success rates are lower at the article level: 16% (45% × 35%) and 19% (53% × 35%), respectively.

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026



	Political Science	Economics	Sociology	Health	Psychology	Demography	Business & Management	Soc. Sciences (interdisc.)	Other	All
$n_{\text{code sharing audit}}$	273	139	131	100	70	55	44	87	224	**1123**
$n_{\text{code shared}}$	126	48	53	28	23	23	8	27	60	**396**
Sharing rate	46%	35%	40%	28%	33%	42%	18%	31%	27%	**35%**
$n_{\text{reproducibilty audit}}$	36	20	15				29			**100**
$n_{\text{reproducible}}$	20	10	8				15			**53**
Reproducibility rate, conditional	56%	50%	53%				52%			**53%**
Reproducibility rate, overall	26%	17%	22%				15%			**19%**

Figure 7. The leaky reproducibility pipeline. Code sharing and reproducibility rates by discipline. The graph shows, from top to bottom, the full sample ($n = 1123$), the subset of articles with code ($n = 396$), as well as the subset of articles for which reproduction was attempted ($n = 100$) and successful ($n = 53$). Colours represent different disciplines. Reproducible articles are defined as having all main results reproducible within a ±10% margin. The sampling step between our code sharing and our reproducibility audit—a random draw of 100 Stata-based articles—slightly shifts the sample towards disciplines where Stata is more common (e.g. political science, economics, sociology). However, because we report code sharing and reproducibility *rates*, proportions are still comparable. The table below the graph reports case numbers and code sharing/reproducibility rates in numerical format. Disciplines with small case numbers in the reproducibility audit are grouped together. 'Other' disciplines include, for example, statistics and methods ($n = 39$), communication ($n = 23$), criminology and law ($n = 22$) and education ($n = 17$). Disciplines are based on subject categories from Web of Science data (see electronic supplementary material, table A2, for details).

Although we caution against extensive cross-study comparison, table 3 and figure 6 allow us to situate our results in prior literature, which usually reports reproducibility at the article level. Interestingly, our estimates closely match the medians of prior work (see figure 1): our conditional reproducibility (53%) is almost identical to the median of prior studies (56%). Similarly, our overall reproducibility rate (19%) resembles the median of the literature (15%).

5.4. Widening the scope: reproducibility of articles across disciplines

How do different disciplines fare? Figure 7 displays descriptive evidence for discipline-specific code sharing and reproducibility rates. The graph in the upper panel shows the disciplinary composition of our sample across all stages of our audit, from the full sample ($n = 1123$) to the subset of articles with code ($n = 396$) to the sample of articles for which reproduction was attempted ($n = 100$) and successful ($n = 53$). The graph's funnel shape indicates the leakiness of the reproducibility pipeline, i.e. articles dropping out at every stage of the process either because they are not open or not reproducible. Within stages, discipline-specific bars indicate how many articles from each discipline are retained at every stage. Comparing bars of a discipline *across* stages allows gauging discipline-specific code sharing and reproducibility rates (see the table below the figure for the same information in numerical format).

Overall, code sharing and reproducibility are clearly imperfect across all disciplines. Whether code sharing, conditional reproducibility, or overall reproducibility is considered, none of the disciplines in our sample come close to 100%. Still, there is relevant variation. For example, code sharing is much higher in political science (46%) than in business and management (18%). In fact, political science stands out as the discipline with the highest code sharing rate. It exceeds the rate of all other disciplines by a statistically significant margin, except for sociology and demography (see electronic supplementary material, figure A6, for bivariate tests of proportions). In economics, a discipline that has received considerable attention in previous reproducibility literature, the code sharing rate is about 11 percentage points lower than in political science.

At the reproducibility stage (see bottom panel of the table in figure 7), disciplinary differences are smaller. Conditional reproducibility rates fluctuate around 50%, with none of the minor cross-disciplinary differences being statistically significant (see electronic supplementary material, figure A7). This may indicate that conditional reproducibility does not meaningfully differ across disciplines using the

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

ESS, or that the statistical power in our sample is insufficient to detect them, given the relatively small number of cases at the reproducibility stage.

The overall reproducibility rate, which reflects both code sharing and reproducibility, ranges between 15% and 26% across disciplines.[15] Even in political science, which exhibits the highest overall reproducibility, we approximate that only about one in four articles is reproducible.

We treat these patterns as exploratory and discuss their limitations in §6.2. One firm conclusion emerging from our cross-discipline comparison, however, is that among studies using ESS data openness and reproducibility are imperfect across disciplines.

5.5. Further obstacles not reflected in our statistics

Our reproducibility estimates are conservative in two important regards. First, of the 53 articles in our sample that were reproducible, about one quarter ($n = 14$) start from pre-processed data. In these cases, we cannot reproduce variable recodings, sample selection or other data-cleaning steps—steps that are crucial in large-*N* observational research. It could reasonably be argued that these 14 articles should be classified as irreproducible. Doing so would reduce the conditional reproducibility rate to 39%, and the overall rate to approximately 14%.

Second, reproductions required substantial time and effort. Although we did not keep detailed time logs, reproduction attempts frequently required several hours, *excluding* computational runtime. Reproductions were particularly time-consuming when extensive detective work was required—for example, when hundreds of pages of undocumented output had to be reviewed to locate a single result. In some cases, we spent several hours on a reproduction attempt but were still unable to generate any result. By contrast, replication packages that included README files and clearly annotated code typically allowed us to execute analyses and identify key results within minutes. Considering that other studies applied much stricter time limits,[16] our results should be viewed as upper-bound estimates of reproducibility rates.

6. Conclusion

6.1. Summary of main findings

Openness and reproducibility are essential prerequisites for the accumulation of reliable scientific knowledge. Yet, as our study shows, both are suboptimal in studies using ESS data. Four findings from our audit stand out.

First, reproducibility problems are common in ESS studies. In the first stage—our *code sharing audit*—only about one-third of authors (35%) shared their analysis code, meaning that roughly two-thirds of articles (65%) could not be reproduced because code—an essential prerequisite for reproducibility of complex survey research—was missing. At the second stage—our *reproducibility audit*—we found that even when code was available, only about half of all main results (51%) were exactly reproducible (conditional reproducibility rate). Taking both findings together, our results imply that only about one in six published results (18%) is exactly reproducible by obtaining and rerunning the original code (overall reproducibility rate). Moreover, reproducibility problems are widespread: even when allowing for minor numerical deviations, only about half of the articles in our sample (53%) are reproducible on all main results using the original materials.

Second, deviations between originals and reproductions are unsystematic. Neither effect sizes nor statistical significance are systematically 'stronger' in original research. Furthermore, deviations are mostly inconsequential for signs and significance and unlikely to alter substantive conclusions.

Third, openness and reproducibility are low across all disciplines: sharing rates range from 18% to 46%, and overall reproducibility rates range from 15% to 26%. Disciplinary differences are most clearly visible at the code sharing stage.

Fourth, reproductions are resource-intensive and often require substantial detective work. Identifying suitable reproduction targets (i.e. main claims and results) is already time-consuming, and handling poorly written and documented code adds further burden. On the bright side, this means that reproducibility can easily be improved through better documentation and organization of research materials.

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

6.2. Implications, limitations and directions for future research

More important than any single number is that our systematic audit confirms a central conclusion of prior research: the current level of computational reproducibility in social research is low, and the lack of accessible replication materials is a major contributing factor. The fact that reproductions require substantial time and effort highlights another problem: if reproductions are not straightforward, researchers are disabled or deterred from effectively building on each other's work [7,8]. This is especially detrimental as social research not only advances academic debate, but also informs public policy [69]. When empirical findings inform decisions affecting millions of people, it is essential to ensure that results are based on sound empirical evidence. Yet, across all disciplines which we examined, results fell short of what is deemed desirable given the substantial policy relevance of research in these fields [69].

An encouraging finding is that our analyses did not reveal systematic bias among results that are different. Although further research is needed, the patterns we observe are, at least overall, more consistent with random variation than with systematic bias (i.e. inflated 'significance' or effect sizes) that may mislead practical conclusions. At the same time, any lack of reproducibility can undermine trust in science as a reliable foundation for policymaking [8]. Moreover, when reproducibility rates are low, current practices of communicating uncertainty solely through sampling error—without considering the reproducibility of findings—are insufficient [8].

Our study's limitations provide promising avenues for future research. First, we concentrated on ESS-based studies. Although we are confident that ESS studies provide an interesting test case for large-*N* survey-based social research, future audits could assess reproducibility rates in other samples. Such audits may also either focus on publicly available datasets to simplify reproduction—including panel data, which often require even more complex pre-processing and analytical decisions—or concentrate on administrative data which pose entirely different challenges to reproducibility (e.g. data sensitivity).

In addition, larger audits beyond our efforts are needed to strengthen the evidence base for cross-discipline comparisons. Although our audit is among the largest to date, resource constraints allowed us to analyse only a few dozen studies per discipline. Several disciplines, especially those in our 'other' category, feature case numbers too small for any firm conclusions. Additionally, evidence on discipline-specific differences may not generalize to research beyond the ESS. Without broader audits, it remains speculative whether results from ESS-based research reflect broader disciplinary patterns or are driven by selective use of ESS data across disciplines. Hence, our study can provide only tentative evidence regarding disciplinary differences, and larger and broader samples are essential to corroborate these patterns.

It is also important to bear in mind that our audit covers articles published between 2015 and 2020. Therefore, our sample cannot reflect recent open-science developments that have occurred since then. In some disciplines, such as political science and economics, initiatives like the Institute for Replication (I4R) [70] and increasingly stringent journal policies may have improved reproducibility rates. At the same time, journals in other disciplines, including top-tier ones such as *Psychological Science* and *Sociological Science*, have also implemented stronger transparency requirements in recent years [71,72]. Without systematic audits, the effects of these measures remain speculative. Beyond continuous coverage on the causal effects of such policies (see e.g. [13,29] for similar approaches), future studies should engage in descriptive monitoring exercises to identify disciplines in which measures may be particularly necessary.

The generalizability of our findings is further potentially limited by the fact that our reproducibility audit focused on studies using Stata. Because of its strong focus on version control [59], Stata may be a best-case scenario for reproducibility. Conversely, if researchers with open and reproducible workflows moved on to non-proprietary software (e.g. R, Python), Stata studies may be less reproducible due to self-selection. We encourage other researchers to investigate reproducibility in several programming languages.

We assessed computational reproducibility only for articles with code. This overestimates true reproducibility if reproducible authors are more inclined to share code. To mitigate this limitation—and unlike most prior audits that condition on material availability (e.g. [38,39])—we explicitly audited both stages (code sharing and reproducibility) within one sampling frame, allowing us to report overall as well as conditional reproducibility rates. Furthermore, anecdotal evidence from our code sharing audit challenges the notion that only the most reproducible researchers shared their code with us. Several authors explicitly acknowledged that their code may be unsuitable for reproduction,

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

sometimes openly wondering whether it even belonged to the article we requested. Future research could compare analytic reproducibility of articles with and without code (see e.g. [73] for such an approach), although such endeavours have been characterized as time-consuming [17,22].

Finally, like all computational reproducibility audits, our study does not speak to the correctness of the original results. Results may be perfectly reproducible but still wrong [19,37,74], and analyses must not always match those reported in the methods section. In fact, our audit already accidentally uncovered several such errors. One study on voting behaviour, for instance, mistakenly included respondents under 16 years of age, below the legal voting age in any European country. Similarly, several analyses implemented different strategies for weighting, imputing or truncating their data than they reported in the article. While it is beyond the scope of the present study to address such errors in depth, our audit provides fertile ground to investigate them further, possibly developing systematic ways to detect them. In an era of growing mistrust towards science, such scrutiny is arguably more important than ever [69,75].

6.3. Improving reproducibility: recommendations for data providers, authors, editors and journals

Our results have several implications for data providers, authors, editors and journals. The fact that reproducibility rates have remained low despite years of open science advocacy suggests that appeals alone are insufficient [11]: measures to improve reproducibility need to be institutionally mandated and enforced. This is especially true given the public good problem associated with open science [76,77]. While open, well-documented materials are desirable for the scientific community, preparing them requires individual effort which is costly in a 'publish or perish' environment. Meaningful interventions must therefore meet two criteria: they must be low-cost and prevent freeloading (i.e. profiting from others' shared code while not sharing one's own).

While our recommendations for authors, editors and journals partly echo recommendations from previous audits, we also highlight issues unique to large-*N* observational research with secondary data (e.g. data provenance and extensive pre-processing) that have received less attention. Moreover, data providers have thus far largely been overlooked as stakeholders in enhancing reproducibility, and our focus on a single dataset offers distinctive insights into their role. We structure our recommendations by key actors and highlight concrete, actionable steps to improve reproducibility without adding extensive additional cost or workload.

6.3.1. Data providers

A major takeaway from our audit is that identifying the exact data*set* a study used can be challenging even when the general data *source* is clear. While all studies in our sample used the ESS, it was often unclear which exact wave and/or edition they used. While authors are partly responsible for documenting their data sources transparently (see following section), data providers must ensure that raw data are clearly identifiable and accessible to the scientific community. This can be achieved by a few straightforward steps outlined below:

- *Unambiguous identifier*: Each data release must have a unique, citable identifier (DOI). This applies to entirely new datasets (e.g. new survey waves) and updates of existing datasets (e.g. new editions to correct data entry errors).
- *Citation guidelines*: Data providers must lay out citation guidelines for their datasets, including reference to a specific DOI. Users must be required to cite the data appropriately when using it.
- *User-generated data extracts*: If users are allowed to create custom data extracts via point-and-click interfaces, the resulting datasets should come with (open source) code to recreate the desired datasets from the raw data. Alternatively, users may obtain *only* such code and could be pushed towards implementing the subsetting themselves. When implemented right, this would add little burden for authors while ensuring the reproducibility of custom datasets.
- *Archiving*: All existing data versions must remain accessible in the data providers' archive to ensure reproducibility even after new releases.

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

6.3.2. Authors

We maintain that it is the authors' responsibility to clearly state their central claims and supporting results. It is also the authors' responsibility to ensure that their methods are correct and transparently set out in their code and materials (see e.g. [78] for similar calls). Our audit demonstrates that substantial improvements are required in these areas.

- *Clarity of claims and results*: Researchers should state their key claims in an article clearly, and indicate which figures or numerical estimates exactly support each claim [79,80]. In their replication materials, all models producing reported estimates must be clearly labelled (e.g. 'Table 1—Model 1'). Conversely, models *not* reported in an article should *not* appear in the replication materials (or be clearly flagged) as they clutter the code and cause confusion.
- *Data citation*: Authors must state both their data source and exact version, wave or edition (ideally using a DOI, see above). Currently, this standard is often violated (see §5.2.1 and [26]).
- *Open materials*: All files required to reproduce results must be deposited in a sustainable public repository. This includes code for data preparation and analysis, as well as user-written software (e.g. Stata ados). When non-public data are used, authors should provide as much detail as legally and ethically permissible. Unless journals suggest any specific repository, authors might find a suitable registry via re3data [81] or FAIRsharing [82], or use established repositories such as the Open Science Framework [83], the Dataverse [84] or Zenodo [85].
- *Documentation and structure*: Replication packages must include a comprehensive README file disclosing the data source, required software, instructions on how to execute the code, and a runtime estimate [21,24,37]. They should follow established templates [86].
- *Code quality*: Authors are encouraged to follow best-practice guidelines for structuring and documenting code (see e.g. [87]). More advanced researchers might opt for fully reproducible manuscripts, where code and text are interwoven [17,29].
- *Pre-submission reproducibility checks*: Before submission, authors should check their code for reporting errors (as these may lead to results being falsely identified as irreproducible) and ensure the correctness and reproducibility of their analyses. At a minimum, such reproducibility checks should be conducted on a different machine to rule out local dependencies [12]. Even better, pre-submission reproducibility checks should be conducted by a co-author or external researcher [22] to ensure that all materials are interoperable and that the replication package does not contain logical gaps unnoticed by the original author.

6.3.3. Journals and editors

Our audit also underscores the limited success of post-publication requests for replication materials: many authors did not respond or declined to share their code (see also [21,29]). Several authors, however, indicated that the materials could have been easily shared at the time of submission. This highlights the crucial role of journals in enforcing reproducibility standards.

- *Mandatory policies*: Journals must impose mandatory data and code sharing requirements [21,29]. These requirements must be clearly formulated, easy to follow and consistently enforced.
- *Guidelines*: Clear, standardized guidelines must specify the minimum methodological information to be provided in the article, the code and README files. Wherever possible, templates should be provided to ensure standardization, e.g. regarding README files.
- *Completeness checks*: Authors should be required to confirm compliance with the journal's guidelines upon submission [24]. This could be easily implemented through a mandatory checklist, for example asking authors to verify that they included a README file, checked the reproducibility of their analyses, and properly cited all relevant datasets in accordance with journal and/or data provider requirements. Technological tools including artificial intelligence may offer new ways to assist in these processes. AI-based systems could automatically verify completeness of this checklist, and may even assist in checking whether these statements are true (e.g. whether the replication package truly includes a README file) [88].
- *Verification*: The review process is crucial for ensuring reproducibility of a submitted manuscript. As a gold standard, journals would implement pre-publication reproducibility checks by in-house data editors or external organizations (for similar proposals, see [13,20,24]). While these efforts are effective, they are also very cost-intensive, even if reproducibility checks are only conducted for a random subset of submitted articles. A low-cost measure to increase reproducibility is to perform basic checks establishing minimum requirements (e.g. checking file endings

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

to ensure the availability of code). This is another promising avenue for automated tools and artificial intelligence, as these may be employed to substantially reduce the manual burden of reproducibility checks [60,89].

— *Archiving and discoverability*: Journals that maintain their own repositories must ensure that replication packages are automatically archived with citable DOIs [5,24], so that authors can be easily credited for these efforts. When in-house archiving is not possible, authors should be guided towards certified external repositories during submission (see recommendations for authors). In published articles, data availability statements must appear prominently and in a standardized place (e.g. on the title page) to make replication materials easy to locate [24].

6.3.4. Additional stakeholders

Improving reproducibility is a collective effort and involves actors beyond authors and journals. We therefore offer selected recommendations for other parties as well.

— *Reviewers* can promote higher reproducibility standards by declining to review papers that do not comply with minimal transparency requirements [90].
— *Funding agencies* can require that grantees publish in journals adhering to open science standards and make all materials—including code—publicly available after completion of the project.

As our audit demonstrates, despite public data, reproducibility in studies using ESS data is far from guaranteed. Replication materials are often unavailable, and even when shared, they frequently contain errors or insufficient documentation that hinder reproduction. However, these problems are preventable. Enforcing minimum reproducibility standards could help weed out 'bad apples', fostering more efficient and reliable empirical social science. We hope that our audit contributes to such practical improvements.

Ethics. Approval to conduct this study was granted by the Institutional Review Board of the Faculty of Social Sciences at LMU Munich (GZ 22-03). The participating authors were informed that their materials would be used to assess reproducibility. Participation was optional, and refusal did not incur any consequences. Informed consent regarding the experimental nature of our code request could not be obtained prior to conducting the study, as this would have jeopardized the validity of the results. In line with the recommendations by the ethics board, we debriefed all authors unless they had explicitly objected to receiving any further e-mails.

Data accessibility. All data, materials and analysis scripts related to this study are publicly available on the Open Science Framework (https://osf.io/ytu9d/) and our results can be reproduced following the instructions in the README file. Because original authors shared their code confidentially, we cannot share their materials in our replication package and only provide anonymized data from our reproducibility audit. As we recognize the irony of a study on openness being unable to share its raw data, we offer researchers access to our raw data upon request at our institution under appropriate privacy policies. We acknowledge that this arrangement introduces practical hurdles, but deem it necessary to protect participants' privacy.

Electronic supplementary material is available online [91].

Declaration of AI use. The following AI tools were used exclusively for final language editing, after the completion of the manuscript's conceptual and argumentative development: ChatGPT-5, DeepL and Claude Sonnet 4.5.

Authors' contributions. D.K.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, visualization, writing—original draft, writing—review and editing; L.S.: conceptualization, data curation, formal analysis, investigation, methodology, resources, validation, writing—review and editing; K.A.: conceptualization, funding acquisition, methodology, project administration, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the German Research Foundation (DFG) through the Priority Program META-REP (464507200; grant no. AU 394/5-2). The funder had no role in designing or conducting this study.

Acknowledgements. We are grateful to all authors who shared their code with us. We thank participants at the 'Seminar on Analytical Sociology: Theory and Empirical Applications' and members of the META-REP Priority Programme, especially Jörg Ankel-Peters, Julian Rose, Florian Neubauer and the working group on reproducibility in observational research. Andreas Schneck co-started this project. Josef Brüderl provided valuable advice throughout and contributed particularly thorough feedback on the manuscript and the presentation of results. We thank Anna Joraschek and Niklas Ippisch for excellent research assistance.

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

Endnotes

[1]Synonyms include 'verification' [1], 'verifiability' [2], 'pure replication' [3] and 'analytic reproducibility' [4].
[2]The SCORE programme [52] aims to provide evidence on a wide set of social and behavioural disciplines, including sociology. Their reports are, however, not yet available.
[3]We share our protocol as part of our replication package on OSF to enable reuse in other reproducibility studies.
[4]Correspondence about multiple articles with the same author occurred only when a corresponding author forwarded our request to a co-author who was already in our sample. In such cases, we randomly selected one article for which we requested code.
[5]Before contacting original authors, we pre-screened all published articles for pointers to replication materials. Specifically, we extracted all hyperlinks from the PDF articles and scanned them for references to journal websites, personal homepages or replication archives. We found links to materials in 37 articles (3.1%), most of them hosted on the journal website or Dataverse. Because some of these replication materials appeared incomplete, we still contacted their authors to give them the chance to provide updated and complete code. Our request explicitly mentioned our pre-screening procedure and apologized to authors if they had published replication materials which we missed, asking them to point us in the right direction. The exact wording of our code request can be found in the supporting information of [57].
[6]When authors claimed that their study did not substantively use ESS data, we manually verified these cases and excluded them if necessary. Because not all authors may have pointed out this issue, our net sample might still include some undetected overcoverage. We address this point in a sensitivity analysis in electronic supplementary material, text A1.
[7]Such a point-and-click system for generating custom ESS datasets is still in place, but each downloaded dataset is now accompanied by an HTML document that details the underlying data versions and selected variables. This was not the case for articles from our target period.
[8]As mentioned above, the experimental results of our code sharing audit have been presented in a previous paper [57]. Our findings here mirror those reported there, though some of the numbers differ slightly due to different sampling frames. For our experimental study, we excluded authors with invalid e-mail addresses since they did not receive treatment. However, for this study we included them, since failure to share code due to authors being unreachable is itself meaningful data.
[9]Recall that we found links to replication materials for 37 articles (i.e. 9% of all articles for which we ultimately obtained code). Of these, 11 authors did not reply to our request (four due to inactive e-mail accounts); 26 shared code upon request or confirmed that their publicly available material was complete.
[10]We minimized the risk of getting caught in spam filters by sending requests from an institutional e-mail account.
[11]We did not accept such offers for two reasons. First, it probably would have delayed our audit significantly. Second, it seemed unethical given that we were mainly interested in the technical reproducibility of the original materials.
[12]In theory, reproductions are possible without original code if replicators can reconstruct the original analysis based on the article's methods section—so-called 'recreate reproductions' [63]. In practice, scholars have had great difficulty with this task [22], even when methods sections are meticulously written. Furthermore, rewriting code adds uncertainty as to whether reproduction failure stems from original reporting errors or mistakes in reconstructing the appropriate analysis pipeline [29]. If recreate reproductions are already challenging for experimental research, they are practically impossible for observational studies where analysis code is usually longer and more complex [11].
[13]While installing user-written packages in Stata is generally straightforward, it becomes burdensome when a specific ado version is required, when ados require manual adaptions, or when ados are hosted on personal homepages.
[14]Confidence intervals could be calculated only for $n = 391$ original estimates due to insufficient reporting in some studies (e.g. p-values reported as 0).
[15]Because the overall reproducibility rate is the product of two rates derived from different sample sizes, we cannot provide a statistical significance test for it.
[16]For instance, the Guide for Accelerating Computational Reproducibility in the Social Sciences [68] classifies any reproduction that requires more than an hour (excluding computational runtime) as irreproducible. We did not set such a strict time limit because the time needed to reoperate code depends not only on the clarity of the code but also on the complexity of the analyses. Furthermore, keeping exact time logs is difficult when smaller, easy-to-fix error messages occur after parts of the analyses already ran for hours or even days. In such cases, it is particularly helpful when authors provide information on the time needed to rerun their code in README files [21].

References

1. Clemens MA. 2017 The meaning of failed replications: a review and proposal. *J. Econ. Surv.* **31**, 326–342. (doi:10.1111/joes.12139)
2. Freese J, Peterson D. 2017 Replication in social science. *Annu. Rev. Sociol.* **43**, 147–165. (doi:10.1146/annurev-soc-060116-053450)
3. Hamermesh DS. 2007 Viewpoint: replication in economics. Can. J. Econ. **40**, 715–733. (doi:10.1111/j.1365-2966.2007.00428.x)
4. LeBel EP, McCarthy RJ, Earp BD, Elson M, Vanpaemel W. 2018 A unified framework to quantify the credibility of scientific findings. Adv. Methods Pract. Psychol. Sci. **1**, 389–402. (doi:10.1177/2515245918787489)
5. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JPA, Taufer M. 2016 Enhancing reproducibility for computational methods. *Science* **354**, 1240–1241. (doi:10.1126/science.aah6168)
6. Stodden V, Seiler J, Ma Z. 2018 An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl Acad. Sci. USA* **115**, 2584–2589. (doi:10.1073/pnas.1708290115)
7. Balafoutas L, Celse J, Karakostas A, Umashev N. 2025 Incentives and the replication crisis in social sciences: a critical review of open science practices. J. Behav. Exp. Econ. **114**, 102327. (doi:10.1016/j.socec.2024.102327)

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

8. National Academies of Sciences, Engineering, and Medicine. 2020 *Reproducibility and replicability in science*. Washington, DC: National Academies Press.
9. King G. 1995 Replication, replication. PS: Polit. Sci. Polit. **28**, 444 (doi:10.2307/420301)
10. Jowell R, Roberts C, Fitzgerald R, Eva G. 2007 *Measuring attitudes cross-nationally: lessons from the European Social Survey*. London, UK: SAGE. (doi:10.4135/9781849209458)
11. Auspurg K, Brüderl J. 2022 How to increase reproducibility and credibility of sociological research. In *Handbook of sociological science* (eds K Gërxhani, N De Graaf, W Raub), pp. 512–527. Cheltenham, UK: Edward Elgar Publishing. (doi:10.4337/9781789909432.00037)
12. Eubank N. 2016 Lessons from a decade of replications at the *Quarterly Journal of Political Science*. *PS: Polit. Sci. Polit.* **49**, 273–276. (doi:10.1017/S1049096516000196)
13. Fišar M, Greiner B, Huber C, Katok E, Ozkes A, Management Science Reproducibility Collaboration. 2024 Reproducibility in management science. *Manage. Sci.* **70**, 1343–1356. (doi:10.1287/mnsc.2023.03556)
14. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, Ioannidis JPA. 2018 Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in the *BMJ* and *PLoS Medicine*. *BMJ* **360**, k400. (doi:10.1136/bmj.k400)
15. Herbert S, Kingi H, Stanchi F, Vilhuber L. 2023 The reproducibility of economics research: a case study. *SSRN J.* (doi:10.2139/ssrn.4325149)
16. Dewald WG, Thursby JG, Anderson RG. 1986 Replication in empirical economics: the journal of money, credit and banking project. *Am. Econ. Rev.* **76**, 587–603.
17. Hardwicke TE *et al*. 2021 Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: an observational study. *R. Soc. Open Sci.* **8**, 201494. (doi:10.1098/rsos.201494)
18. Moffitt RA. 2011 Report of the editor: *American Economic Review*. *Am. Econ. Rev.* **101**, 684–699. (doi:10.1257/aer.101.3.684)
19. McCullough BD, McGeary KA, Harrison TD. 2006 Lessons from the JMCB Archive. *J. Money Credit Bank.* **38**, 1093–1107. (doi:10.1353/mcb.2006.0061)
20. Gertler P, Galiani S, Romero M. 2018 How to make replication the norm. *Nature* **554**, 417–419. (doi:10.1038/d41586-018-02108-9)
21. Chang AC, Li P. 2022 Is economics research replicable? Sixty published papers from thirteen journals say 'often not'. *Crit. Financ. Rev.* **11**, 185–206. (doi:10.1561/104.00000053)
22. Artner R, Verliefde T, Steegen S, Gomes S, Traets F, Tuerlinckx F, Vanpaemel W. 2021 The reproducibility of statistical results in psychological research: an investigation using unpublished raw data. *Psychol. Methods* **26**, 527–546. (doi:10.1037/met0000365)
23. Ioannidis JPA *et al*. 2009 Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155. (doi:10.1038/ng.295)
24. Xiong X, Cribben I. 2023 The state of play of reproducibility in statistics: an empirical analysis. *Am. Stat.* **77**, 115–126. (doi:10.1080/00031305.2022.2131625)
25. Wood BDK, Müller R, Brown AN. 2018 Push button replication: is impact evaluation evidence for international development verifiable? *PLoS One* **13**, e0209416. (doi:10.1371/journal.pone.0209416)
26. Malnar B. 2025 *European Social Survey academic impact monitoring: annual report 2024*. Ljubljana, Slovenia: University of Ljubljana.
27. Moody JW, Keister LA, Ramos MC. 2022 Reproducibility in the social sciences. *Annu. Rev. Sociol.* **48**, 65–85. (doi:10.1146/annurev-soc-090221-035954)
28. Gelman A, Loken E. 2014 The statistical crisis in science. *Am. Sci.* **102**, 460. (doi:10.1511/2014.111.460)
29. Hardwicke TE *et al*. 2018 Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *Cognition*. *R. Soc. Open Sci.* **5**, 180448. (doi:10.1098/rsos.180448)
30. Silberzahn R *et al*. 2018 Many analysts, one data set: making transparent how variations in analytic choices affect results. Adv. Methods Pract. Psychol. Sci. **1**, 337–356. (doi:10.1177/2515245917747646)
31. Huntington-Klein N *et al*. 2021 The influence of hidden researcher decisions in applied microeconomics. *Econ. Inq.* **59**, 944–960. (doi:10.1111/ecin.12992)
32. Nosek BA, Errington TM. 2020 The best time to argue about what a replication means? Before you do it. *Nature* **583**, 518–520. (doi:10.1038/d41586-020-02142-6)
33. Krawczyk M, Reuben E. 2012 (Un)available upon request: field experiment on researchers' willingness to share supplementary materials. *Account. Res.* **19**, 175–186. (doi:10.1080/08989621.2012.678688)
34. Wicherts JM, Borsboom D, Kats J, Molenaar D. 2006 The poor availability of psychological research data for reanalysis. *Am. Psychol.* **61**, 726–728. (doi:10.1037/0003-066x.61.7.726)
35. Tedersoo L *et al*. 2021 Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192. (doi:10.1038/s41597-021-00981-0)
36. Vines TH *et al*. 2014 The availability of research data declines rapidly with article age. *Curr. Biol.* **24**, 94–97. (doi:10.1016/j.cub.2013.11.014)
37. Pérignon C *et al*. 2024 Computational reproducibility in finance: evidence from 1,000 tests. *Rev. Financ. Stud.* **37**, 3558–3593. (doi:10.1093/rfs/hhae029)
38. Brodeur A, Mikola D, Cook N. 2024 Mass reproducibility and replicability: a new hope. *IZA Discussion Paper Series*. (doi:10.2139/ssrn.4790780)
39. Bergh DD, Sharp BM, Aguinis H, Li M. 2017 Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strateg. Organ.* **15**, 423–436. (doi:10.1177/1476127017701076)
40. Avelino G, Desposato S, Mardegan I. 2021 Transparency and replication in Brazilian political science: a first look. *Dados* **64**, e20190304. (doi:10.1590/dados.2021.64.3.242)

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

41. McCullough BD, McGeary KA, Harrison TD. 2008 Do economics journal archives promote replicable research? *Can. J. Econ. Can. D'économique* **41**, 1406–1420. (doi:10.1111/j.1540-5982.2008.00509.x)
42. Klein RA *et al*. 2018 Many Labs 2: investigating variation in replicability across samples and settings. Adv. Methods Pract. Psychol. Sci. **1**, 443–490. (doi:10.1177/2515245918810225)
43. Davis AM, Flicker B, Hyndman K, Katok E, Keppler S, Leider S, Long X, Tong JD. 2023 A replication study of operations management experiments in *Management Science*. *Manage. Sci.* **69**, 4977–4991. (doi:10.1287/mnsc.2023.4866)
44. Open Science Collaboration. 2015 Estimating the reproducibility of psychological science. *Science* **349**, aac4716. (doi:10.1126/science.aac4716)
45. Herndon T, Ash M, Pollin R. 2014 Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Camb. J. Econ.* **38**, 257–279. (doi:10.1093/cje/bet075)
46. Malter D. 2014 Female hurricanes are not deadlier than male hurricanes. *Proc. Natl Acad. Sci. USA* **111**, E3496. (doi:10.1073/pnas.1411428111)
47. Sanchez C, Sundermeier B, Gray K, Calin-Jageman RJ. 2017 Direct replication of Gervais & Norenzayan (2012): no evidence that analytic thinking decreases religious belief. *PLoS One* **12**, e0172636. (doi:10.1371/journal.pone.0172636)
48. Ritchie SJ, Wiseman R, French CC. 2012 Failing the future: three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS One* **7**, e33423. (doi:10.1371/journal.pone.0033423)
49. Galiani S, Gertler PJ, Romero M. 2017 Incentives for replication in economics. *NBER Working Papers*. (doi:10.3386/w23576)
50. KuehbergerA,Schulte-MecklenbeckM.2018Selectingtargetpapersforreplication.*Behav.BrainSci.***41**,e139.(doi:10.1017/S0140525X18000742)
51. Romero F. 2018 Who should do replication labor? *Adv. Methods Pract. Psychol. Sci.* **1**, 516–537. (doi:10.1177/2515245918803619)
52. Alipourfard N *et al*. 2021 Systematizing confidence in open research and evidence (SCORE). *SocArXiv*. (doi:10.31235/osf.io/46mnb)
53. Heyard R *et al*. 2025 A scoping review on metrics to quantify reproducibility: a multitude of questions leads to a multitude of metrics. *R. Soc. Open Sci.* **12**, 242076. (doi:10.1098/rsos.242076)
54. Schnaudt C, Weinhardt M, Fitzgerald R, Liebig S. 2014 The European Social Survey: contents, design, and research potential. *SCHM* **134**, 487–506. (doi:10.3790/schm.134.4.487)
55. Thibault RT, Kovacs M, Hardwicke TE, Sarafoglou A, Ioannidis JPA, Munafò MR. 2023 Reducing bias in secondary data analysis via an Explore and Confirm Analysis Workflow (ECAW): a proposal and survey of observational researchers. *R. Soc. Open Sci.* **10**, 230568. (doi:10.1098/rsos.230568)
56. Fink L, Marcus J. 2025 Replication code availability over time and across fields: evidence from the German Socio-Economic Panel. *Econ. Inq.* **63**, 357–386. (doi:10.1111/ecin.13267)
57. Krähmer D, Schächtele L, Schneck A. 2023 Care to share? Experimental evidence on code sharing behavior in the social sciences. *PLoS One* **18**, e0289380. (doi:10.1371/journal.pone.0289380)
58. Priem J, Piwowar H, Orr R. 2022 OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. (https://arxiv.org/abs/2205.01833)
59. StataCorp. 2025 *Reproducibility and backward compatibility*. See https://www.stata.com/flyers/reproducibility19.pdf (accessed 8 September 2025).
60. Brodeur A, Valenta D, Marcoci A, Aparicio JP, Mikola D, Barbarioli B, Alexander R, Deer L, Stafford T. 2025 Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science. *SSRN J.* (doi:10.2139/ssrn.5118632)
61. Trisovic A, Lau MK, Pasquier T, Crosas M. 2022 A large-scale study on research code quality and execution. *Sci. Data* **9**, 60. (doi:10.1038/s41597-022-01143-6)
62. Kohrt F, Melinscak F, McElreath R, Schönbrodt FD. 2024 A conceptual framework for computational reproductions: formal definitions and epistemic functions. Zenodo. (doi:10.5281/ZENODO.14282874)
63. Dreber A, Johannesson M. 2025 A framework for evaluating reproducibility and replicability in economics. *Econ. Inq.* **63**, 338–356. (doi:10.1111/ecin.13244)
64. Vilhuber L. 2020 Reproducibility and replicability in economics. *Harv. Data Sci. Rev.* **2**. (doi:10.1162/99608f92.4f6b9e67)
65. Naqvi A. 2024 TREEMAP: Stata module for treemaps (v1.6). Statistical Software Components. S459123.
66. Franco A, Malhotra N, Simonovits G. 2014 Publication bias in the social sciences: unlocking the file drawer. *Science* **345**, 1502–1505. (doi:10.1126/science.1255484)
67. Imbens GW. 2021 Statistical significance, *p*-values, and the reporting of uncertainty. *J. Econ. Perspect.* **35**, 157–174. (doi:10.1257/jep.35.3.157)
68. Berkeley Initiative for Transparency in the Social Sciences. 2020 *Guide for advancing computational reproducibility in the social sciences*. See https://bitss.github.io/ACRE/ (accessed 9 October 2025).
69. Christensen GS, Freese J, Miguel E. 2019 *Transparent and reproducible social science research*. Oakland, CA: University of California Press.
70. Brodeur A, Dreber A, Hoces de la Guardia F, Miguel E. 2024 Reproduction and replication at scale. *Nat. Hum. Behav.* **8**, 2–3. (doi:10.1038/s41562-023-01807-2)
71. Hardwicke TE, Vazire S. 2024 Transparency is now the default at *Psychological Science*. *Psychol. Sci.* **35**, 708–711. (doi:10.1177/09567976231221573)
72. Sociological Science. 2023 *Reproducibility policy*. See https://sociologicalscience.com/reproducibility-policy/ (accessed 11 December 2025).
73. Breznau N *et al*. 2025 The reliability of replications: a study in computational reproductions. *R. Soc. Open Sci.* **12**, 241038. (doi:10.1098/rsos.241038)
74. Leek JT, Peng RD. 2015 Opinion: reproducible research can still be wrong: adopting a prevention approach. *Proc. Natl Acad. Sci. USA* **112**, 1645–1646. (doi:10.1073/pnas.1421412111)

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026

75. National Academies of Sciences, Engineering, and Medicine. 2025 *Understanding and addressing misinformation about science*. Washington, DC: National Academies Press. (doi:10.17226/27894)
76. Fecher B, Friesike S, Hebing M. 2015 What drives academic data sharing? *PLoS One* **10**, e0118053. (doi:10.1371/journal.pone.0118053)
77. Kraft-Todd GT, Rand DG. 2021 Practice what you preach: credibility-enhancing displays and the growth of open science. *Organ. Behav. Hum. Decis. Process.* **164**, 1–10. (doi:10.1016/j.obhdp.2020.10.009)
78. Academy of Sociology. 2020 *Checklist for quantitative social science articles*. OSF. See https://osf.io/vkjg8/files/mw59u.
79. Lundberg I, Johnson R, Stewart BM. 2021 What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.* **86**, 532–565. (doi:10.1177/00031224211004187)
80. Kohler U, Class F, Sawert T. 2024 Control variable selection in applied quantitative sociology: a critical review. *Eur. Sociol. Rev.* **40**, 173–186. (doi:10.1093/esr/jcac078)
81. Pampel H *et al*. 2013 Making research data repositories visible: the re3data.org registry. *PLoS One* **8**, e78080. (doi:10.1371/journal.pone.0078080)
82. Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M. 2019 FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **37**, 358–367. (doi:10.1038/s41587-019-0080-8)
83. Foster ED, Deardorff A. 2017 Open Science Framework (OSF). *J. Med. Libr. Assoc.* **105**. (doi:10.5195/jmla.2017.88)
84. King G. 2007 An introduction to the Dataverse network as an infrastructure for data sharing. *Sociol. Methods Res.* **36**, 173–199. (doi:10.1177/0049124107306660)
85. European Organization For Nuclear Research, OpenAIRE. 2013 Zenodo. See https://www.zenodo.org/ (accessed 15 October 2025).
86. Vilhuber L, Connolly M, Koren M, Llull J, Morrow P. 2020 A template README for social science replication packages. Zenodo. (doi:10.5281/ZENODO.4319999)
87. The Turing Way Community. 2025 The Turing way: a handbook for reproducible, ethical and collaborative research. Zenodo. (doi:10.5281/ZENODO.15213042)
88. Goldberg A, Ullah I, Khuong TGH, Rachmat BK, Xu Z, Guyon I, Shah NB. 2024 Usefulness of LLMs as an author checklist assistant for scientific papers: NeurIPS'24 experiment. (https://arxiv.org/abs/2411.03417)
89. Bibal A, Minton SN, Khider D, Gil Y. 2025 AI copilots for reproducibility in science: a case study. (https://arxiv.org/abs/2506.20130)
90. Morey RD *et al*. 2016 The peer reviewers' openness initiative: incentivizing open research practices through peer review. *R. Soc. Open Sci.* **3**, 150547. (doi:10.1098/rsos.150547)
91. Krähmer D, Schächtele L, Auspurg K. 2026 Supplementary material from: Code sharing and reproducibility in survey-based social research: evidence from a large-scale audit. FigShare (doi:10.6084/m9.figshare.c.8337305)

Downloaded from http://royalsocietypublishing.org/rsos/article-pdf/doi/10.1098/rsos.251997/6123953/rsos.251997.pdf
by guest
on 17 March 2026