Sylvérie Herbert
Banque de France

Hautahi Kingi
Google

Flavio Stanchi
Airbnb

Lars Vilhuber*
Labor Dynamics Institute, Cornell University

Abstract. Journals have pushed for transparency of research through data availability policies. Such data policies improve availability of data and code, but what is the impact on reproducibility? We present results from a large reproduction exercise for articles published in the American Economic Journal: Applied Economics, which has had a data availability policy since its inception in 2009. Out of 363 published articles, we assessed 274 articles. All articles provided some materials. We excluded 122 articles that required confidential or proprietary data, or that required the replicator to otherwise obtain the data (44.5% of assessed articles). We attempted to reproduce 152 articles, and were able to fully reproduce the results of 68 (44.7% of attempted reproductions). A further 66 (43.4% of attempted reproductions) were partially reproduced. Many articles required complex code changes even when at least partially reproduced. We collect bibliometric characteristics of authors, but find no evidence for author characteristics as determinants of reproducibility. There does not appear to be a citation bonus for reproducibility. The data availability policy of this journal was effective to ensure availability of materials, but is insufficient to ensure reproduction without additional work by replicators.

Résumé.

JEL classification: B41; C80; C81; C87; C88

*Corresponding author: lars.vilhuber@cornell.edu. This work has benefited from insightful comments by the editor, Abel Brodeur and three anonymous referees. We thank Christophe Pérignon, Ben Greiner, Jan Höffler, and participants of the Banque de France DGSEI seminar and the 2018 BITSS annual meeting at UC Berkeley. The paper has been previously circulated as Kingi et al. (2018) and Herbert et al. (2021). The opinions expressed in this article are solely the authors', and do not represent the views of the American Economic Association, the US Census Bureau, Banque de France, Google, or Airbnb.

Canadian Journal of Economics / *Revue canadienne d'économique* 20XX 00(0)
January 20XX. Printed in Canada / *Janvier 20XX. Imprimé au Canada*

ISSN: 0000-0000 / 20XX / pp. 1–38 /

1. Introduction

Replication, reproduction, and falsification of published articles are important steps in the scientific process. These activities help to make science "robust and reliable" (Bollen et al. 2015) and are a *sine qua non* condition for the credibility of economic research. Robust and replicable research is especially important in policy institutions such as central banks or governments since it provides input needed for its core activities and informs their decisions. Given its importance for both research and policy purposes, the reproducibility and replicability[1] of articles has been discussed in economics for at least thirty years[2]. Since the early 2000s, many economics journals have a data availability policy requiring authors to deposit materials sufficient to reproduce the analysis in their article. The goal is twofold: the exercise of compiling the replication package itself provides an opportunity for authors to fix errors on their own. And the availability of these materials is meant to reduce the cost of replication and as a "starting point for other researchers" (Bernanke 2004). However, compliance is typically only lightly monitored (some authors have referred to it as the "honors system"), and no pre-publication verification is conducted. In this paper, we carry out a large reproducibility exercise of published articles in a journal which implements such data availability policy.

Actual published reproductions or replications are rare (Bell and Miller 2013, Duvendack et al. 2017), notwithstanding some recent examples (Höffler 2017, Chang and Li 2017; 2015, Camerer et al. 2016). For example, Mueller-Langer et al. (2018) found that just 0.1% of the 126,505 articles published between 1974 and 2014 in the top 50 economics journals were replications. Sukhtankar (2017b) found that, of the 1,138 empirical development economics articles published between 2000 and 2015 in the "top 10" economics journals,[3] just 6.2% were replicated in a published or working paper. The paucity of replications in economics is, in part, because it is often difficult to find the materials required to conduct reproducibility or replication exercises (Dewald et al. 1986, McCullough et al. 2006, McCullough and Vinod 2003). Despite a long standing explicit recognition of the importance of replication in economics (Frisch 1933), it has been suggested that "there is no tradition of replication in economics" (McCullough et al. 2006, p. 1093).[4] An exception has been the

1 For precise definitions of these terms, see (Bollen et al. (2015), National Academies of Sciences, Engineering, and Medicine (2019)) and definitions in the next section.

2 (Dewald et al. 1986, Vinod 2005, Anderson et al. 2005, King 1995, Burman et al. 2010, Duvendack et al. 2017, Hamermesh 2017, Sukhtankar 2017a, Hoeffler 2017, Coffman et al. 2017, Chang and Li 2017, Berry et al. 2017, Anderson and Kichkha 2017)

3 Sukhtankar (2017b) combines the traditional "top 5" with the American Economic Journal: Applied Economics (AEJ:AE), the American Economic Journal: Economic Policy (AEJ:EP), the Economic Journal (EJ), the Journal of the European Economic Association (JEEA) and the Review of Economics and Statistics (ReStat).

4 Though Hamermesh (2007), Hamermesh (2017) disagrees.

Journal of Applied Econometrics (since 2003) and more recently the newly created Journal of Political Economy (JPE) Micro, announced in February 2022.

To promote transparency, journals have adopted "data (and code) availability" policies (DAP). The American Economic Association (AEA) announced one in 2004 (Bernanke 2004), the JPE in 2004, the Quarterly Journal of Economics (QJE) was one of the last major economics journals to do so in 2016. Data availability policies require authors to make all data (and generally code) necessary to reproduce their study available, or when legal restriction prohibit the sharing of the data, authors should explain how to get access to the data. As of 2016, the top five economics journals all have a DAP. However, it is still not the norm among economics journals. Studies find that between 57% and 66% of economics journals (with some empirical content) have a DAP that requires or suggests deposit of materials, and between 12 and 14% let authors provide materials on request (Hoeffler 2017, Vlaeminck 2021). There are also differences in the coerciveness of the DAP across journals, with only a small minority making deposit of materials mandatory[5]. Such DAP are not necessarily a panacea, as authors who have tested them in some way note it does not guarantee obtaining all the papers' codes and data, nor does it guarantee reproducibility (Glandon 2011, Hoeffler 2017, Stodden et al. 2018).

More recently, journals and societies have appointed "data editors" responsible for monitoring compliance. The Canadian Economic Association appointed a data editor in 2017. The AEA appointed one of this article's authors as data editor in 2018 (Duflo and Hoynes 2018) and subsequently updated its data and code availability policy. Similar positions have been created at the Review of Economic Studies, the Economic Journal, Management Science, and most recently, Econometrica.

In this paper, we set out to assess how well a particular journal's "data availability" policy, lightly monitored as described in the next section, yields *reproducible* articles. We call the monitoring "light", because no verification of completeness or functionality was conducted prior to publication, in contrast to the more recent monitoring conducted by dedicated data editors. In other words, we ask the question: given that authors are required to provide code and data, without pre-publication verification, is that enough to generate reproducible results?

Our protocol sets a relatively high bar: Can *undergraduates*, armed only with the information provided by authors on the journal website, successfully reproduce the tables and figures presented by the author in the article? Unlike

5 Hoeffler (2017): in 2015, of 343 journals, 26 had mandatory DAP, 110 had voluntary DAP, and 49 allowed authors to provide data and code on request, possibly in combination with the offer of a deposit. Vlaeminck (2021): of 327 journals, 50 had a mandatory DAP, 135 had a voluntary DAP, and 38 had some sort of provision on request.

Dewald et al. (1986) and McCullough and Vinod (2003), who requested data and programs from the original authors, we did not attempt any contact with authors to clarify issues that arose. While our replicators were instructed to do their best to fix any bugs or inconsistencies that they encountered, they were limited both by time and training. On the other hand, it would seem that pure computational reproducibility should be the lowest standard met by an article. We conducted this experiment over several summers starting in 2014 and during the 2018 fall semester, using the AEJ:AE as our source of articles.

We find a moderate replication success, with a reproduction rate of 37.8% overall. 44.7% of articles were successfully reproduced, conditional on data being available, with an additional 43.4% partially reproduced. Compared to a reproduction rate of 13% found by Dewald et al. (1986) in the context of a journal (Journal of Money, Credit and Banking (JMCB)) with no data or code availability policy, our results seem to suggest that journal policies that enhance transparency are helpful, yet not sufficient to reach full reproducibility. We further show that fully reproducible papers do not seem to benefit from a citation bonus — authors' reputation seems to matter the most when it comes to citations. We speculate that we may be in a relatively low reproducibility equilibrium because the costs of producing reproducible research (for instance in terms of time) outweigh the advantages (given it does not lead to more citations). Our contribution to the literature is thus twofold: (1) we provide an estimate of reproducibility standards of a journal that imposed, since its founding, a data availability policy; (2) we provide a rationale for authors' lack of incentives to produce reproducible research, absent journals' verification of reproducibility.

We start by describing, in Section 3, our methods of selection, analysis, and reproduction. We present our results in Section 4 before concluding in Section 5.

2. Background

The focus of the exercise is to assess whether availability of replication packages, in compliance with a posted data availability policy, leads to reproducibility. It is thus worthwhile to briefly discuss the policy as it was applied at AEA journals.

The policy was first announced by (Bernanke 2004). Authors would provide "in electronic form, data and code sufficient to permit replication [sic]." It is unknown when the policy was separately posted on the AEA's website, but the policy as of 2018 (American Economic Association 2008) is largely consistent with the original (very brief) statement. Authors would email (or upload to an FTP server) "data set(s) and programs used to run the final models", including a Readme PDF with instructions. Exemptions for "proprietary" data needed to be approved by the journal's editor. The American Economic Journals

launched in 2009, and applied this policy from the outset. Exemptions are (at least in later years, and through 2020) tabulated in the annual editors' reports (see f.i. Mas 2019).[6]

We chose the AEJ:AE for two reasons. First, because of the empirical nature of its articles and its policy of publishing papers "only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication" (American Economic Association 2008). As we outline in Section 3, we wanted replication materials to be present for all articles in the journal, without needing to contact authors. The AEJ:AE applied the AER's data availability policy from its creation in 2009. From past experience of this team (Vilhuber 2020) and others (e.g. Dewald et al. 1986, McCullough and Vinod 2003, Stodden et al. 2018), requesting materials from authors is fraught with non-response.

The second reason for the choice of the AEJ:AE was based on feasibility. A cursory review of the articles on the AEJ:AE website suggested that articles would be less complex than articles in the American Economic Review (AER) or Econometrica, including by using mostly high-level software commonly used in economics, such as Stata and MATLAB.[7] We wanted articles to be reproducible by undergraduate, students who are mostly armed with knowledge of Stata and MATLAB.[8] However, nothing in the methodology used in this paper is specific to the AEJ:AE, and could be applied to other journals.

An additional, though not crucial, motivation, was to test the feasibility of "pre-publication verification" at AEA journals, similar to what was, at the time, done at the American Journal of Politial Science (AJPS) (Jacoby et al. 2017, Christian et al. 2018). The results obtained from the exercise reported in this article ultimately lead to the implementation of pre-publication verification at the AEA journals (Vilhuber 2019; 2022).

The exercise reported here differs somewhat from many other articles in the space of reproducibility and replicability. Some author teams have attempted to replicate, rather than computationally reproduce, published articles, in particular in the field of behavioral economics and psychology (Open Science Collaboration 2015, Camerer et al. 2016; 2018). Others have taken a less structured approach to computational reproducibility, by simply attempting to run any code within a replication package and check for failure (Trisovic et al. 2022, Wang et al. 2020). Finally, others have conducted replication exercises that were designed to identify cross-researcher degrees of freedom (Menkveld et al. 2023, Huntington-Klein et al. 2021).

6 We briefly touch on assessed reasons for exemptions in Section 4.

7 See Vilhuber et al. (2020b) for a distribution of software used in replication packages deposited by authors in AEA journals from 1999 to 2018.

8 In practice, all articles in AEJ:AE use Stata for most of the analysis, but some use additional softwares. We address this in Section 4 in more details.

How does our exercise compare with and differ from similar exercises conducted in the past? Dewald et al. (1986) attempted to reproduce 54 articles from the JMCB, a few years after JMCB had introduced a pioneering DAP. The policy was so pioneering that many authors did not comply: Only 49 out of 65 (75%) provided replication materials to the journal upon request, despite the policy requiring it. They report a reproduction rate of 7 out of 54 (13%) articles, though they only attempted to actually compute the results for nine articles, absent complete data and code for others (7 out of 9, 78%). Only two were perfectly reproduced (either 2 out of 54, or 2 out of 9). In later work using the same journal, McCullough et al. (2006) found only 14 of 196 (7.1%) articles selected from the JMCB that should have had archives were actually reproducible. Only 69 articles actually had replication archives, some of which were incomplete. Conditional on having code and data, they therefore succeeded in fully reproducing 14 out of 62 (22.6%), though they did not have sufficient computational resources to test the reproducibility of seven others.

This paper is not the first assessment of the AEA's DAP. In 2008, six PhD students attempted to reproduce 39 empirical articles out of 135 papers published in the AER between 2006 and 2008 (Glandon 2011). Out of 39 articles, 11 were based on proprietary data and only 9 papers were actually assessed for reproducibility. However, they also assessed completeness of the provided packages, and estimated that 95% would be reproducible without help from the authors if data were accessible. By the replicators' assessment, none were perfectly reproducible, but more than half (five out of nine) had enough information that substantial effort to fill in the blanks would yield a perfect replication. The other four had immaterial discrepancies between the results and the published paper.

Chang and Li (2015), using slightly different methodology in selection, selected all articles from 13 well-regarded economics journals, including the AER, satisfying certain criteria (empirical paper using data on U.S. gross domestic product), and successfully reproduced the results of only 22 of 67 papers (32.8%). This seems to suggest that journal policies to enhance publications are helpful, but insufficient to foster reproducibility. Most closely related may be Fišar et al. (2023), who enrolled 700 (unpaid) replicators for nearly 500 articles previously published in *Management Science*, both before and after the implementation of a data and code availability policy. Conditional on having data, code, and IT infrastructure in place, they were able to reproduce 95% of articles after implementation of the policy. They find that introduction of the data and code availability policy both increased availability and quality of replication packages.

In a related exercise, Stodden et al. (2018) used articles, including in economics, published in *Science* as the basis of their analysis. Given Science's policy, they had to request data for 180 (88%) of the articles in their sample by contacting authors. They only managed to obtain data and code for 89 (43.6%)

of their sample, and assessed 56 of these. In the end, they only attempted reproduction of 22 (10.8%), for which 95% were reproducible.

A similar exercise, albeit not focused on published articles, is Pérignon et al. (2022). In their case, they were able to re-purpose the 163 replication packages generated by the Menkveld et al. (2023) project, and assess reproducibility of the code used to produce the 6 research questions of that project. It is a much narrower exercise, as all researchers who participated came from a small subfield of finance, and accomplished the programming task within a much more limited timeframe than the typical economics paper. At the same time, it is also more complex, as the choice of software varied more widely within this narrowly focused field than in the entire AEJ:AE corpus, which had to be mastered by a single replicator. The lack of field diversity, topics, and replicator skill diversity allow a clean focus on the researcher skills leading to lack of reproducibility, something which in our context will be somewhat obscured by the intersection of researcher programming skills and replicator execution skills.

3. Methodology

We start by defining terms and concepts, describing the setup for the reproducibility exercise, followed by the bibliometric analysis. As in any modern study on "reproducibility" and "replicability," we need to define those terms, since historically, multiple, often conflicting or confusing definitions have been used (Pesaran 2003, Hamermesh 2007, Bollen et al. 2015, Clemens 2015, National Academies of Sciences, Engineering, and Medicine 2019) Throughout the text, we use (computational) *reproducibility* to refer to "the ability [...] to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" (Bollen et al. (2015), see also National Academies of Sciences, Engineering, and Medicine (2019)). (Christensen and Miguel 2018, p. 942) argue that is the "basic standard [that] should be expected of all published economics research, and hope this expectation is universal among researchers." In contrast, *replicability* refers to "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."[9]

We refer to the collection of instructions, code, and data as "replication packages," because they enable not just the exact computational reproduction

9 We note that we have ourselves not always been consistent in the use of these terms. For instance, our instructions to our "replicators" ask whether the article has been "replicated", not "reproduced," see Appendix A6 and A7. While Bollen et al. (2015) certainly first defined the concepts as we use them here, they only became (relatively well) accepted with the publication of National Academies of Sciences, Engineering, and Medicine (2019), after the conclusion of our experiment. We reproduce our replicator instructions material verbatim in the appendix, and do not modify the terminology.

of an article, but also more generally replication.[10] In our context, we will refer to the (junior) researchers re-executing the code in a replication package as "replicators," primarily for linguistic convenience (see also footnote 9).

Our reproducibility exercise was conducted over the course of five summers, from 2014 to 2018 as well as the fall semester of 2018. The exercise was split into three steps. First, an "assessor" evaluated an article and its replication package as to the availability of the required components and its difficulty, and recorded a selection of article characteristics. A "replicator" then conducted the actual reproducibility check. Often but not always, the same person would be assessor and replicator. Finally, upon completion of the attempt at reproduction, replicators filled out a guided report with questions about the reproduction, such as whether or not the main results of the article could be reproduced and, if not, the main barriers impeding a successful reproduction.

To complement the information collected during the reproducibility exercise, we also obtained descriptive article and author information, such as citations and h-indexes, from the OpenAlex (OA) application programming interface (API) (Priem et al. 2022, OurResearch 2023).

3.1. Article selection

The AEJ:AE published 363f articles between the first issue in 2009 (volume 1 issue 1) and the April issue in 2018 (volume 10 issue 2). Replicators worked on this project from the summer of 2014 until Fall of 2018. Articles were added to the assignment set in groups by year, though somewhat haphazardly. The pilot project in the summer of 2014 worked on articles from 2013, and completed many of them. Subsequently, the collection of information was systematized, but the articles already completed from 2013 were not transferred to the new collection instrument, and are thus missing from our analysis in this paper. In 2015, using the systematized collection instrument, years 2010 and 2011 were added. Years 2009 and 2012 were added in the summer of 2016, and publication year 2014 was added in 2017. Finally, in 2018, the years 2015 through 2017 as well as the first two issues in 2018 were added (see Appendix Table A1). Articles were added in batches to separate online spreadsheets, based on a pull from the CrossRef API, which yielded articles in somewhat random order, augmented by a manual culling of unprocessed articles from the previous summer's list.[11] Once a particular cohort of replicators had completed a batch, a new batch was pulled. Replicators were assigned an initial article by the supervising team, but could subsequently self-select (first come, first serve) from the list of unprocessed articles. They were instructed to finish the initial assessment regardless of the outcome. We do not believe

10 Journals sometimes do refer to these as "supplemental data," but in economics, they almost always contains code, if not data.

11 Ex-post inspection of our assignment lists shows no particular pattern, though articles with lower issue and page numbers seem to appear higher in the original lists.

TABLE 1
Articles Published and Selected by Year

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
Published	34	40	37	40	38	40	35	37	42	20	363
Selected	21	23	32	39	8	40	21	33	39	18	274
Percent	61.76	57.5	86.49	97.5	21.05	100	60	89.19	92.86	90	75.48

Notes: Assessments made using the entry questionnaire by replicators, prior to attempting to reproduce any tables or figures. The sample of assessed papers comprises those for which we had complete assessment questionnaires.

that there was intentional systematic assignment based on data availability, software usage, or topic.

Table 1 shows the number of eligible articles (i.e., published articles) per year, and the number of articles each year that were assessed and scored, as described in the next few sections.

3.2. Initial Assessment

Before attempting a reproduction, each article was assessed. An assessor filled out the entry questionnaire, gathering descriptive information, and providing an initial assessment of the expected level of reproducibility of an article.[12] Each assessor was provided the Digital Object Identifier (DOI) of the article,[13] and then verified the following elements:

- The presence of one or more downloadable datasets (including DOI or URL, if any), an online appendix, the programs used by the authors and documentation on how to run them (the "Readme");[14]
- The clarity and completeness of the documentation and of the program metadata;

12 See Appendix A6. We will refer to questions on the entry questionnaire by "Entry Q1", "Entry Q2", etc.

13 DOIs are a managed identifier space built on top of the Handle System (Sun et al. 2010), a technology for distributed, persistent, and unique naming for digital objects. Virtually all academic publishers assign DOIs at the article level in all of their publications. In addition, DOIs are increasingly used to identify data (Pollard and Wilkinson 2010). In particular, each DOI provides a persistent identifier (International DOI Foundation (IDF) 2012) for a digital object: an article or data artifact.

14 In theory, the author might have provided a DOI to a third-party data archive for some or all of the content. In the case of the AEJ:AE, at the time this exercise was conducted, only the article itself was assigned a DOI. Supplemental data, programs, and online appendices are linked from the landing page associated with the article's DOI. This has changed since completion of the exercise, see Vilhuber et al. (2020b).

- The presence of clear references to the original data provenance and a description of how to construct the initial datasets;
- Data availability, and some basic categories if not available as part of the replication package (e.g., restricted access data, private data, public use data, etc.)

While many supplemental data packages contained some content, they often did not contain all the data or programs. Sometimes, the data provenance might be described in an online appendix, while the instructions for the programs might be enclosed in the supplemental "data" package. Thus, a "clerical" review of each article's webpage, and some careful reading of the actual article and online appendix were the only way to collect all the information requested.

Based on the initial objective enumeration of the characteristics of the article and a subjective evaluation by the assessor of the complexity of the task described in the "Readme" document, the assessor was asked to provide a subjective rating of the replication difficulty, from 1 (easiest) to 5 (most difficult), based on a set of heuristics (see Table 2, Entry Q60).[15] Assessors also recorded the programming languages (Entry Q8) and separately the data storage formats (Entry Q21, Q28, Q36, Q44) contained within each archive. Archives can and do contain programs and data in multiple formats. While all articles had some supplementary data, not all articles were accompanied by the datasets necessary for replication. Assessors recorded whether the articles were accompanied by any data (Entry Q14), and if not, the (apparent) reason why data were not provided (Entry Q15). This included an assessment of whether the data were confidential or proprietary, for which we provided some guidance.

15 A score of 1 assigned to an article does not imply that its replication package cannot be improved. For example, the programs provided by authors might lack documentation, or the provenance of the included datasets might be obfuscated, despite the article being easily reproducible.

TABLE 2
Criteria for Assessment of Difficulty of Replication

Rating	Description
1	The article possesses all desired features that ensure replicability. Datasets are provided and their use is public. The documentation and program metadata are clear and complete. Negligible changes might be required to run the programs e.g. path redirection).
2	The article possesses most desired features that ensure replicability. Datasets are provided and their use is public. The documentation and program metadata are present but the programs might need some changes to run cleanly.
3	The article replication may present some difficulties. Datasets are provided and their use is public, but the documentation is incomplete or unclear. Substantial changes might be needed to run the programs.
4	The article replication may present substantial difficulties and/or additional steps are required to recover the datasets used by the authors. Datasets may not be provided but their use is public or available on request.
5	The article is not replicable. The dataset are not provided and their access is private or restricted. The programs are not provided. Documentation is absent or incomprehensible.

NOTE: The description is reproduced verbatim. The language used in the instructions used "replication" instead of "reproduction", as we generally use it in this article.

3.3. Reproducing the Empirical Analyses

Once the article was assessed and determined to be amenable for replication, a replicator was assigned based on characteristics of the article, and in particular the type of software required and the assessed difficulty of the replication task. Most often, the initial assessor self-assigned themselves as the replicator although some replications were conducted by the supervising team or distributed to a replicator based on their familiarity with a particular programming language. The materials for each article were downloaded, and used to populate an article-specific repository. Unlike Dewald et al. (1986), McCullough and Vinod (2003), Stodden et al. (2018), who requested data and programs from the original authors when no materials were published by the journal, but similar to Glandon (2011), we only tried to obtain data that was available on the journal's archive, and replicators were explicitly instructed not to contact the authors. This is very much in the spirit of King (1995), McCullough et al. (2006), Glandon (2011) and others, who emphasize the importance of being able to reproduce the results without assistance from

the authors. Thus, we did not follow-up on missing datasets, private data, protected data, or data available upon request. All attempts at reproducing the analysis were done based exclusively on the materials provided by the authors at the time of publication of the articles, or references therein leading to external data archives.

The replicator was instructed to document any changes made to the author-provided programs using a version control system (VCS). They also recorded information in two additional files: Any URLs that were used to download materials were recorded in *'SRC.txt'*, and a high-level summary of steps undertaken for the replication and results obtained in a short report (called *'REPLICATION.txt'*).

Once the VCS was populated and all data files downloaded, each replicator was instructed to read the author-provided "Readme", and attempt to run the programs in the author-provided archive. Replicators were told to keep modifications to the absolute minimum, starting with the adjustment of path structures to the replication system (for an example, see Appendix Figure A1). Whenever a program required more extensive changes to run, the replicator would do so to the best of their ability. The entire team discussed problems encountered in regular meetings, and sought to find solutions. However, we also limited the time to find a solution for problems encountered to about a week. If unsuccessful in finding a solution, the reproduction attempt was marked as "not reproducible." There was no time limit for running programs.

Replicators were free to use any computer they had access to for the replication, unless the author materials specifically mandated a particular operating system. This was quite rare, but did happen a few times. We note that the journal does not require authors to specify the required software, operating system (OS), and versions thereof, and most articles were silent on the topic. In addition to replicators' laptops, our team had access to university-provided Windows remote desktops and a Linux cluster, and were unlikely to be constrained by computing resources. Both Windows and Linux systems had access to the latest version of the computing software, typically updated yearly. In general, we assume that versions of OS and software are different from the version originally used by the authors, given the considerable time lag between submission, publication, and the time of the reproducibility exercise. In some cases, programs needed small modifications to run cleanly due to software version discrepancies. If successful, these modifications were treated as 'negligible' and did not affect the score of the article. If unsuccessful, however, the article was classified as non- or partially reproducible. We discuss the necessary code changes required for successful reproduction along with our results in Section 4.

3.4. Report on Reproducibility

Once the reproduction attempt was completed, replicators described outcomes via the exit questionnaire, recording information about the success or failure

of the reproduction attempt and other descriptive information.[16] We asked replicators to classify the attempt as fully, partially, or not reproducible (Exit Q3), and elicited perceived causes (Exit Q4-Q12). For partial or full reproducibility, we asked them to further describe some of the possible issues: whether modifications to the programs were needed, and if yes, what those changes were, and whether all output was accurately reproduced (Exit Q7-Q12). We also collected additional information on the packages: Software and data formats they encountered, including versions, and whether any extra packages or libraries were required (Exit Q13-Q17). Finally, we asked them to describe the completeness (ex-post) of the description of the reproduction in the authors' README (Exit Q22), and how difficult they found the reproduction attempt (Exit Q26).[17]

3.5. Bibliometric data

In addition to the data collected through the replication process, we also obtained complementary descriptive data for each article and the (academic) characteristics of the authors and their institutions. We queried OpenAlex (OA, Priem et al. 2022, OurResearch 2023) for data on each article in the sample. We then queried OA for all works published by authors in our sample, using the per-year citation data to recompute authors' h-index (Hirsch 2005). We also computed author experience (years since the first publication). OA also identifies authors' institutions at the time of publication, and provides information on each institution (as of 2023), such as the total number of works published by all authors at the institution.[18] We use this number as a proxy for the research intensity of the authors' institutions, though we cannot be sure that the institution at the time of publication is the affiliation at the time of preparation of the manuscript. We also identify the location (country or region) of the institution. More details can be found in the Online Appendix.

3.6. Recruitment of Replication Lab Members

Over the course of five summers, we recruited undergraduate students (typically but not always rising seniors) for the reproducibility exercise as part of summer research, to serve as assessors and replicators. The team members needed to meet some minimal technical qualifications, such as experience working with the relevant programming languages, and acceptable

16 The print version of the online questionnaire is provided in Appendix A7. We refer to questions on the exit questionnaire as "Exit Q1", "Exit Q2", etc.

17 We note that information on completeness and difficulty are also queried in the entry questionnaire, but the analysis in this article will rely on the ex-post description recorded in the exit questionnaire.

18 Note that this number is an attribute of the institution in 2023, not at the time of publication of the article. For reference, OA reported 334943 works published cumulatively in their corpus by authors affiliated with Cornell University, as of October 2023.

performance in economics or technically equivalent courses (it turned out that many of our students had never taken an economics class). Team members attended a one-day training course covering the background and purpose of the reproducibility exercise and our approach. They were also given guidance about the subjective aspects of the exercise such as difficulty rating and classifying documentation clarity, and were instructed on some technical matters such as version control with Subversion and the use of remote computing clusters using materials from a Cornell High Performance Computing course designed for social science researchers. The team members were supervised by economics Ph.D. candidates from Cornell University (Kingi, Stanchi, Herbert) and a faculty-level researcher (Vilhuber). As this was not part of any coursework, the replicators were not graded. They were paid an hourly wage commensurate with Cornell University pay scales for student employees.[19]

The information gathered by the students in the entry and exit questionnaires allow us to grasp how documented the data and code is, and how easy the reproduction can potentially be.

4. Results

4.1. Analysis of replication packages

Out of 363 eligible articles in AEJ:AE, 274 were assigned and assessed. Table 3 shows the classification of the 274 assessed articles into three categories (Appendix Table A2 breaks these numbers down by year of publication). The assessors identified 86 articles that relied on confidential data, proprietary data, or data that would require payment, application, or registration. Another 8 provided neither data nor explanation for the absence of data. These were likely articles that obtained exemptions from the data availability policy, the standard policy at the time for articles with confidential data. Finally, 180 articles (65.7%) appeared to contain sufficient information, data, and code to attempt a reproduction.[20]

19 The selection, training, and supervision are precursors to the training and supervision described in Vilhuber et al. (2022b), but are not identical.

20 To be precise, we have complete entry and exit questionnaires for 180 articles. We exclude articles which were assigned for reproduction, but for which the reproduction attempt or the exit questionnaire are incomplete. These are mostly articles which were assigned towards the end of the summer, and which were aborted as the student left the summer job.For some articles, a second reproduction attempt was made, for a number of reasons, including explicit double-coding, "onboarding" of new replicators, and operator error. Duplicates are removed in the analysis to keep the most successful outcomes when several outcomes were available for a given article. This creates an intentional upward bias in terms of reproducibility, given the skill distribution of our replicators. Results do not materially change when switching to the first recorded outcome.

TABLE 3
Assessment of Data Availability

Assessment	Articles
Confidential Data	86
No Data or Reason	8
Some Data Provided	180
Total	274

Notes: Assessments made using the entry questionnaire by replicators, prior to attempting reproduction. Entry Q14, Q15.

We note that when authors requested an exemption due to data confidentiality, they may have understood that they were also exempted from the policy to "precisely document[]" the data. Thus, provision of no information, as in the case of the 8 articles in Table **??**, may be consistent with the implementation of the policy at the time. In what follows, we therefore concentrate on articles that did, in fact, provide some information that could be assessed. We first document the types of software required and data formats, a subjective measure of difficulty of reproduction, as well as ex-post documentation clarity. We then turn to the reproduction results themselves.

4.1.1. Software and data formats

Most replication packages relied on proprietary, non-open source software for statistical programs, and stored data in proprietary formats. The vast majority of code in replication packages used the Stata programming language for at least some portion of analysis (Table 4). This dominance of a single language is reflective of broader usage in economics, though the particular dominance of Stata might be specific to the AEJ:AE. From a reproducibility perspective, Stata has both advantages and disadvantages. While it is proprietary software, it is relatively cheap and accessible. Many packages to extend its usability are available, many of which are accessible from within the software from both peer-reviewed (Stata Journal) and crowd-sourced (RePEc/SSC) repositories. Unfortunately, in contrast to CRAN, the SSC does not currently support versioning of packages, making it sometimes difficult to find the original version of a package used by authors.

Table 4 also indicates that economists tend to provide data in the native format of the programming language used, instead of open formats (CSV and others). The Stata file format has proven to be quite robust, as newly released versions of Stata maintain backward compatibility to all previous

TABLE 4
Programming Languages and Data Formats

Software	Programming Language	Data Format
Stata	252	174
Not Reported	14	76
Matlab	11	3
SAS	8	0
R	6	2
SPSS	2	1
Excel	1	7
CSV	0	10
txt	0	2

Notes: Combination of answers from Entry Q8, Q21, Q28, Q36, Q44. Column sums are greater than the number of articles because articles can use more than one programming language or data format. Sample of assessments made using the entry questionnaire by replicators prior to attempting to reproduce any tables or figures.

versions of the data format. Furthermore, the data format is well understood (albeit not open-source), and can be read by many open-source software packages (R, python). The Stata file format allows the embedding of richer metadata, which is not feasible for (basic) CSV formats. We did not verify that metadata (variable labels, same variable names, etc.) complied with modern data curation standards.

4.1.2. Subjective difficulty

Replicators were asked to provide a subjective measure of the difficulty of "reproducibility," based on the criteria described in Table 2. Outcomes both for the full sample of assessed articles, as well as for the subset of articles for which reproducibility was attempted, are presented in Table 5. Note that the assessment was made before a decision was made whether or not to reproduce the article, though it was known that articles with confidential data would not be reproduced. Because data availability entered into the assessment, it is not surprising that replicators found articles that we ultimately attempted to reproduce to be easier (median rating of 2 for attempted articles vs. a median rating of 3 for all assessed articles). These ratings should be interpreted with respect to the skill level of the replicators, as well as the time and effort that could be devoted to these attempts.

TABLE 5
Difficulty of Reproduction

Difficulty Rating	Assessed articles	Percent	Attempted articles	Percent
1	57	20.8	49	27.22
2	56	20.44	56	31.11
3	56	20.44	43	23.89
4	35	12.77	20	11.11
5	70	25.55	12	6.67
Total	274	100	180	100

Notes: Entry Q60. Assessments made prior to attempting to reproduce any tables or figures.

4.1.3. Documentation of replication packages

Previous authors have argued for improved documentation of submitted data and programs (McCullough et al. 2006, Chang and Li 2015). Reproduction and in particular replication attempts are made significantly easier when replicators are not required to resolve ambiguity, and the source of each table and figure is well documented. In contrast to previous assessments, Table 6 presents a summary of the ex-post documentation quality, after a reproduction was attempted, by year of publication. As evaluated by the replicators, the quality of documentation seems fairly good, with a majority of articles being perceived as well documented. 133 articles out of 180 (73.9%) provided complete documentation, defined as a README file with step by step instructions on how to execute every provided program. However, 45 articles (25%) provided incomplete ReadMe files that either skipped some of the important steps required to run the programs or contained some ambiguous instructions. No documentation was provided in 2 articles.

4.2. Analysis of reproducibility

Once an article had been assessed and if data appeared to be available, replicators attempted to re-run the code, using the provided data.[21] The outcome for the 180 attempts was then recorded. Tables and figures obtained as a result of the attempted reproduction may differ in precision (small discrepancies, rounding errors) or coverage (all, some or few results being

21 We note that we recruited undergraduates students on the basis of knowledge of the two dominant software packages, Stata and MATLAB. We did not restrict our selection of papers to said software. Replicators did have knowledge of other software, and were assisted, if necessary, by the authors of this paper in running the software.

TABLE 6
Ex-Post Assessment of Documentation Clarity

	n	Percent
Complete	133	73.89
Incomplete	45	25
No Info	2	1.11

Notes: Articles with attempted reproduction. Exit Q22.
Complete = Provided all information required to run the programs.
Incomplete = Was ambiguous or left out crucial steps.

reproduced). Depending on whether the differences were with respect to precision or coverage, we instructed replicators to categorize articles between partial or non-reproducibility. Assessors were instructed to categorize articles as fully reproducible if all numbers and figures matched, up to some decimals (allowing for some minor rounding errors, hence mainly differences in precision). The second category, partial reproduction, concerned articles for which differences in tables and figures went beyond small precision differences. Articles were categorized as partially reproducible if replicators were able to execute the computer programs that produced the numerical values reported in the articles but that there were differences in the numerical values, beyond rounding errors, for some tables.[22] However, for articles to be partially reproducible, the results should still be qualitatively similar, or the main results had to hold, but other secondary results and robustness checks could differ. Papers that failed either of these criteria were deemed not reproduced.

Table 7 presents the main results of the reproduction exercise. 68 of 180 reproduction attempts fully reproduced the article's analysis (37.8%). A further 66 (36.7%) were at least partially successful. The dominant reason for unsuccessful reproductions is the absence of confidential or proprietary data. For these 28 cases, either assessors had missed the information that no data was available, or the information was not clearly available from the replication package's documentation. Combining with the 86 articles identified as relying on confidential data at the initial assessment stage, a total of 114 out of 274 (41.6%) relied on confidential or proprietary data, and were thus not reproducible by this project. If we exclude articles that critically rely on

22 Possible causes may lie in software version discrepancies, uninitialized random number generators, different operating systems, or even different machines. We did not identify the causes of the discrepancy.

confidential data, the headline percentages would be 44.7% fully reproduced, and v% partially reproduced. In most of our subsequent analysis, we will exclude articles relying on confidential data.

Table 7 identifies 18 articles that were not able to be reproduced for reasons other than absence of confidential data. These reasons are listed in Table 8. 6 articles did not provide data, with no indication that the data might be confidential. We did not investigate the reason for this apparent non-conformance to the AEJ:AE data availability policies, although we point out that some non-confidential data are still subject to terms of use that prevent redistribution (earlier years of IPUMS data and any version of PSID data are just two examples). Errors in the provided computer programs prevented the reproduction of 1 article[23], while the data provided in 3 articles was corrupted in some way so that the software available to us was not able to read the datasets. Our replicators did not have access to the software required to run 1 article.[24] For 5 articles, the computer programs successfully ran, but the numerical values were inconsistent with those reported in the articles, and the replicators were unable to find a convincing reason.

Many (partially) successful reproductions required complex code modifications. We tabulate in Table 9 the extent to which modifications to the provided computer programs were required to successfully reproduce the articles. The majority of successful reproductions required minimal work from the replicators. 38 of the 68 successful reproductions required, at most, a simple rerouting of directory references. The remaining 24 successful articles required a deeper understanding of the software, and a more in-depth analysis

23 For instance, one example of assessor's comment was "Could not replicate [sic] due to incorrect use of indicator variable function. Did not understand what the author was trying to achieve due to lack of comments in the code, and therefore could not come up with alternate way to generate the dta file."

24 The software was "PostGIS in PostgreSQL", which was not easily available.

TABLE 7
Reproduction Results

Outcome	No. of Articles	Percent		No. of Articles	Cond. on Data	
Successful	68	37.78	%	68	44.74	%
Partial	66	36.67	%	66	43.42	%
Confidential Data	28	15.56	%			
Other failure	18	10	%	18	11.84	%
Total	180	100	%	152	100	%

TABLE 8
Reason for Unsuccessful Reproduction

Cause	No. of Articles
Most Numbers Differ	1
Missing Data	6
Corrupted Data	3
Missing Code	1
Code Error	1
Software Unavailable	1
Other	5

Note: Articles for which a reproduction was attempted, the reproduction was unsuccessful, with valid Exit Q4 (see Appendix). Question was not asked for articles with partial reproduction success.

of the code and/or command of the subject matter. These "Complex Changes" to the code required more than simple directory adjustments such as, for example, the debugging of classical code errors or the adjustment of outdated commands to reflect newer versions of software or operating systems. The fact that about 35% required complex changed calls for at least better documentation in implementing these changes were they unavoidable, along with more robust coding practices.

TABLE 9
Code Changes for Successful Reproductions

	Fully	Partial
Complex Change	24	32
Directory Change	19	23
Missing	6	3
No Change	19	8

Note: Partially or fully reproduced articles. Answers to questions Exit Q8 and Q9. These questions were skipped when reproduction was not successful.

Good documentation is key to better reproducibility, as emphasized by many authors (McCullough et al. 2006, Chang and Li 2015, Stark 2018). In our sample, better documentation is positively correlated with reproduction success (Table 10), providing some support for this assertion.

TABLE 10
Correlation of Reproduction Success vs Documentation Quality

	Estimate	p-value
Pearson	0.25500	0.00057
Kendall	0.25500	0.00066

Notes: Sample of reproduced and scored papers, i.e, assessed papers for which we attempted reproduction.

Author characteristics may well affect the reproducibility of replication packages. The code changes needed to make a replication package reproducible, for which Table 9 provides some indication, are a function of the authors' skills and experience. Institutional support may also affect reproducibility. To explore this dimension, we used the OA data to obtain various characteristics of the authors, their institutions, and the articles. While the next section will investigate the impact on citations, we consider here the potential determinants of reproducibility. Ideally, we would like to measure elements of the training of the authors, for instance during their Ph.D. years. Limitations of the OA prevent that, though, as OA data was only available for the previous 10 years, preventing us from measuring some author characteristics as of time of publication of the paper. We computed the h-index (Hirsch 2005) for all authors of a paper for each year that they are present in OA, and then computed the average, the maximum, and the minimum per-paper h-index across all authors, for each year of the data. We also consider the highest experience (years since publication of the first article) amongst an article's authors, as well as location and productivity of their institution in the relevant year.

Table 11 presents a summary of these measures, as of four years after publication, categorized by reproduction success and data availability.[25] On average within the entire sample, articles have 2.32 authors and were cited 9.33 times per year, four years after publication. The author with the longest publication record within each author team has 24.4 years of experience. Amongst the authors' home institutions, the highest has a faculty that has produced about 210000 citable works. The vast majority of papers have at least one author in the United States.

The data reported in Table 11 do reveal some differences across the various outcome categories, albeit without clear patterns. For instance, articles with authors at highly productive institutions do not appear to produce more reproducible replication packages. Full reproducibility is associated with a higher percentage of authors in the US, but also a lower h-index. Articles that are fully reproduced do seem to be associated with a higher number of citations in the fourth year after publication, but are also associated with less experienced author teams. Table 12 therefore disentagles these correlates by means of the likelihood of observing a fully reproduced article, conditional on not relying on confidential data.[26] We consider average, minimum, and maximum h-index among co-authors, the productivity of affiliated institutions - both linearly and whether the highest institution falls into the top or bottom third of the productivity distribution of institutions, a linear component

25 Appendix Table A5 shows the same measures for the subset of articles for which data are available for the year of publication itself (Year 0).

26 Because of the aforementioned limitation of OA data availability, Table 12 and Appendix Table A4 only include data for articles first published in 2012 or later.

TABLE 11
Publication and Author Metrics

	Confidential data	Unsuccessful	Partial	Successful
Avg h-index	16.41	21.57	17.22	18.5
Lowest h-index	9.43	14.17	10.18	10.41
Number of Authors	2.21	2.17	2.32	2.56
Citations	9.24	6.61	8.23	11.26
Highest experience	24.31	27.11	23.45	24.84
Institutional productivity	20.07	26.25	18.5	23.1
Percent of authors in US	85.12	72.22	74.24	83.82
N	121	18	66	68

Notes: All assessed articles, except for 1 author dropped for inconsistent OA data.
Author and institutional characteristics are measured 4 years after publication, due to limitations in the OA data availability.
Institutional (cumulative) productivity measured in 10,000 publications.

in the number of authors, and in column (4), an indicator for whether the article is solo-authored. Only the kitchen-sink specification in column (4) shows some significant determinants of reproducibility. The higher the maximum h-index is, the less likely a replication package is reproducible. The productivity of the home institutions appears to be non-linear, with the middle third of the distribution (omitted from the regression) having a negative impact on reproducibility. Appendix Table A4 shows the same regressions when the outcome is defined as at least partial reproducibility, and shows no significant correlates. From this exercise, we cannot determine what may cause differential reproduction outcomes.

4.2.1. Computing the Reproduction Rate

The literature generally provides a summary or headline "reproducibility rate" to describe the outcome of reproducibility exercises like ours. For instance, McCullough et al. (2006)'s headline number in the abstract is "Of more than 150 empirical articles, fewer than 15 could be replicated" (14 out of 196, to be exact), for a reproducibility rate of 7.1%. On the other hand, the article notes that only 62 reproductions were attempted (and 7 were skipped due to lack of resources), which yields instead a reproducibility rate of 22.6%.

Choice of the denominator (how many articles to include as the basis of evaluation) and the numerator (what to consider a successful reproduction) matters. Part of the choice of the numerator depends on how to include the skill level of the replicator. Would a partially successful reproduction attempt have been classified as fully reproducible by a replicator with

TABLE 12
Probit: Determinants of Reproducibility, Year 0

	Outcome: Full Reproduction			
	(1)	(2)	(3)	(4)
'Avg. H-index'	0.010			0.102
	(0.015)			(0.066)
'Max H-index'		−0.002		−0.052*
		(0.008)		(0.029)
'Min H-index'			0.025	−0.018
			(0.023)	(0.043)
'Highest Institution Publications'	−0.001	−0.001	−0.001	
	(0.007)	(0.007)	(0.007)	
'Institution Publications (top)'				0.706*
				(0.398)
'Institution Publications (bottom)'				1.360**
				(0.638)
'Highest Co-author Experience'	−0.004	0.002	−0.005	−0.016
	(0.014)	(0.013)	(0.013)	(0.015)
'Number of authors'	0.084	0.098	0.121	0.161
	(0.112)	(0.116)	(0.116)	(0.158)
'Author at US university'	0.345	0.368	0.347	0.039
	(0.318)	(0.318)	(0.318)	(0.339)
'Solo-authored'				−0.467
				(0.439)
Constant	−0.712**	−0.707**	−0.838**	−0.955
	(0.345)	(0.347)	(0.368)	(0.596)
N	113	113	113	113

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$
Sample restricted to articles published in 2012 and later, with attempted reproduction.

more advanced skills, better access to support, or simply more time? How to classify reproductions that required changes to the code that would require an interaction with the author, but would ultimately lead to full reproducibility? In the literature, some of the attempts to reproduce analyses allowed knowledgeable replicators to "fix" the code. Choice of the numerator depends critically on how to define the basis. Should one include articles that should be reproducible (for instance, all articles published in a certain time period), but were never assigned and thus not tested? Should one include articles for which the current team of replicators did not have the software, hardware, data, or skill resources, but which replicators with higher endowments in any of these dimensions would succeed, because they have

access to licensed software, high-performance computers, restricted or massive data, or know how to compile code in arcane compilers?

In this exercise, aware of the limits of both the scope of the exercise and the skills of the replicators, we carefully define what we call the "reproduction rate" — the percentage of articles that were tested and found to be either reproduced or not. We define the reproduction rate as

$$R_{t,s} = \frac{n_t}{d_s}, \tag{1}$$

where n_t is the number of articles that were either fully or at least partially reproduced (t=$\{full, partial\}$), as defined above. We define the denominator d_s as the number of articles either s=$\{assessed, attempted, nonconfidential\}$, where the first (*assessed*) is the total number of articles we assessed and could have attempted a reproduction for (274). We did not randomly choose which of the assessed articles to attempt a reproduction for - we removed those where we knew that they relied on confidential data that would take extra time to obtain (86), or provided neither data nor a reason for the absence of data (8). We attempted to reproduce the remaining 180 (*attempted*). Of these, it turns out that 28 also relied on confidential data. Had we been able to properly assess these when making the initial assessment, they would also have been excluded, leaving 152 articles that legitimately should have been reproducible, given the parameters of the exercise (*nonconfidential*).

We report all three reproduction measures in Table 13, to allow the readers to come to their own conclusions. However, we emphasize the difference between "reproduced" (something our results show) and "reproducible." In particular when data are difficult, but not impossible to access, articles may be able to be reproduced by those with access — thus being plausibly "reproducible" until proven otherwise. Many articles that rely on confidential data have been shown to be reproducible in the same way we demonstrate reproducibility here (Pérignon et al. 2019, Vilhuber 2021; 2022).

4.3. Reproducibility and Impact on Citations

Journal editors have noted that any credible research paper must be reproducible, in order to serve as a starting point for future research.[27] Given citation practices in the social sciences, this should lead to a positive correlation between reproducibility and future citations. We would expect *a priori* that reproducible papers provide research which can be easily built upon and that other researchers are thus more likely to use. Hamermesh (2007) observes that heavily cited articles are also replicated (not reproduced) in the literature, while poorly cited articles are not. While this is also a positive

27 "[Reproducibility] is essential if empirical findings are to be credible and usable as a starting point for other researchers." (Bernanke 2004) "[Reproducible papers] should be the starting point for related future work. [...] Reproductions are the most essential type of reanalysis. They are the foundation of science." (Welch 2019)

TABLE 13
Reproduction Rates

Denominator	Number of Articles	n_{full} (%)	$n_{partial}$ (%)
$d_{assessed}$	273	24.90%	50.20%
$d_{attempted}$	180	37.80%	76.10%
$d_{nonconfidential}$	152	44.70%	88.20%

Notes:
Full or partial reproduction are defined in the text.

correlation ex-post, it does not answer the question of whether (unobserved) underlying reproducibility was possibly a condition for future replication, in the sense of Bernanke (2004).[28]

We collected the annual citations of each article from OA, and computed cumulative citations and per-year h-indexes up to year 5 post-publication.[29] In the following Tables 14 to 17, we investigate the relationship between cumulative citations measured four years after publication and reproducibility measures, as well as some of the same bibliometric measures and author characteristics identified earlier. Ideally, we would measure the bibliometric measures in the year of publication. We do so for the sample of articles published in 2012 and later, in the appendix. Here, we measure the bibliometric measures with a one-year lag, but for the whole sample.

We start by exploring the role of the type of data used in explaining citation outcomes, using the full sample. We regress cumulative total citations for an article on an indicator of whether the article used confidential data (Table 14). Results indicate a positive but non-significant citation bonus for papers with confidential data. The prime determinants of future citations are the number of authors, and the h-indexes of the various authors of an article, though that pattern appears to differ somewhat for articles using confidential data.

Conditional on not using confidential data, how does the reproducibility of an article affect its future citation count? We regress the inverse hyperbolic sine (arcsinh) transform of year-to-date citations on an indicator of whether

28 Hamermesh actually uses the positive correlation between presence of replications and citations to argue that (enforced) reproducibility is unnecessary: "Sparsely-cited articles in major journals are not killed by replications that cast doubt on their results; rather, they "die" from neglect." (Hamermesh 2017)

29 OA does provide a measure of total citations for each paper at the data availability boundary in 2012, making the use of cumulative citations feasible. Tables A8 and A9 in Appendix A4 provide summary statistics.

TABLE 14
OLS: Citations and Confidential Data

	Total Citations			
	(1)	(2)	(3)	(4)
'Avg. H-index'	1.160***			5.380***
	(0.294)			(1.360)
'Max H-index'		0.490***		−1.770***
		(0.163)		(0.582)
'Min H-index'			0.795**	−2.500***
			(0.399)	(0.840)
'Confidential data'	10.200	6.320	−4.300	9.010
	(8.770)	(7.820)	(8.170)	(8.920)
'Highest Institution Publications'	0.114	0.094	0.130	0.095
	(0.133)	(0.135)	(0.136)	(0.132)
'Highest Co-author Experience'	−0.301	−0.160	−0.108	−0.351
	(0.262)	(0.260)	(0.257)	(0.258)
'Number of authors'	6.390**	4.790	7.990**	9.390***
	(3.090)	(3.190)	(3.180)	(3.350)
'Author at US university'	4.050	5.570	5.180	2.310
	(6.360)	(6.410)	(6.470)	(6.280)
'Solo-authored'	2.740	0.669	−0.213	7.400
	(8.000)	(8.070)	(8.080)	(8.080)
'Avg. H-index':'Confidential data'	−0.714			−7.110**
	(0.449)			(2.870)
'Max H-index':'Confidential data'		−0.345		2.560*
		(0.249)		(1.320)
'Min H-index':'Confidential data'			0.327	4.670***
			(0.713)	(1.690)
Constant	7.840	15.100	10.800	1.040
	(11.600)	(11.500)	(11.900)	(12.100)
N	273	273	273	273
Adjusted R^2	0.085	0.063	0.055	0.118

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Notes:
YTD citations are cumulative citations to the article in Year 4.
H-index measures are computed across all authors of an article, in the previous year.
An author without citations has an h-index of 0.
"Confidential data" identifies if the article was assessed to have confidential data.
Results for all articles with complete assessment.

our team was able to partially or fully reproduce the article, along with the other potential determinants of citations, in addition to controls for author characteristics. Table 15 shows that there appears to be only a weak (non-significant) positive effect of reproducibility on citations, with author

TABLE 15
OLS: Arcsin Citations on Reproduction Outcomes

	(1)	(2)	(3)	(4)
'Avg. H-index'	0.025***			0.109***
	(0.008)			(0.041)
'Max H-index'		0.009**		−0.035**
		(0.004)		(0.016)
'Min H-index'			0.025**	−0.043
			(0.011)	(0.026)
'Fully reproduced'	0.154	0.113	0.156	0.193
	(0.238)	(0.215)	(0.208)	(0.244)
'Highest Institution Publications'	0.003	0.002	0.004	0.002
	(0.003)	(0.003)	(0.003)	(0.003)
'Highest Co-author Experience'	−0.008	−0.004	−0.003	−0.009
	(0.007)	(0.007)	(0.007)	(0.007)
'Number of authors'	0.191**	0.147*	0.230***	0.259***
	(0.074)	(0.077)	(0.077)	(0.083)
'Author at US university'	0.046	0.073	0.068	0.033
	(0.157)	(0.160)	(0.160)	(0.157)
'Solo-authored'	0.063	0.012	−0.035	0.172
	(0.207)	(0.209)	(0.209)	(0.216)
'Avg. H-index':'Fully reproduced'	0.003			−0.098
	(0.012)			(0.068)
'Max H-index':'Fully reproduced'		0.004		0.042
		(0.007)		(0.032)
'Min H-index':'Fully reproduced'			0.005	0.059
			(0.017)	(0.040)
Constant	3.190***	3.370***	3.150***	2.980***
	(0.290)	(0.288)	(0.299)	(0.312)
Observations	180	180	180	180
Adjusted R^2	0.158	0.126	0.131	0.164

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Notes:
YTD citations are cumulative citations to the article in Year 4.
H-index measures are computed across all authors of an article, in the previous year.
An author without citations has an h-index of 0.
Full or partial reproduction are defined in the text.
Results for all articles with attempted reproduction.

characteristics (h-index and the number of authors) playing a significant role.[30]

30 Tables A10, A11, and A12 in Appendix A5 show the same pattern using alternate specifications in levels and logs as well as a Poisson regression, respectively.

The bibliometric measures in Table 15 are measured with only a one-year lag to the measured citations, and may be inappropriately capturing some of the already-occurred citations. Table 16 uses the same specification on the sample of articles published in 2012 and later, but with bibliometric measures now collected in the year of the article's publication. The effect of the various h-index variables is no longer significant, but institutional productivity now has a slightly positive effect instead. However, generally, there is no impact of reproducibility on future citations. Using both partial and full reproduced results as the outcome is qualitatively similar (Appendix Table A17).

Finally, we consider the dynamic effect over time. Instead of restricting the sample to the fourth year after publication, we look at the effect on yearly citations over the five years following publication. Since 2018/2019 was a period in which multiple data editors were appointed (AEA, this journal, ReStud, EJ), we add a dummy variable for years 2019 and later. Indeed, the introduction of data editors could have changed economists' perceptions about papers in the AEJ:AE, even though all articles in this study were published prior to the data editor's appointment, and not subject to his scrutiny. Table 17 shows the usual (concave) evolution of citations over time (with a maximum in yearly citations around 7 years after publication). There is no significant effect of reproducibility, other than a small second-order effect through the h-index interaction. The hypothesized data editor effect is not significant.

5. Conclusion

In this paper, we carried out a large scale reproduction exercise of replication packages from a journal with a data availability policy that requires deposit of data and code. Out of 363 articles published during the period under consideration, we assessed 274 articles. All articles provided some materials. We excluded articles that required confidential or proprietary data, or that required the replicator to otherwise obtain the data, either at initial assessment or if discovered during the reproduction attempt. We attempted to reproduce 152 articles, and were able to fully reproduce the results of 68. A further 66 were partially reproduced. The overall reproduction rate, depending on how one counts the partially reproduced, is either $R_{full,nonconf}$ =44.7% or $R_{partial,nonconf}$ =88.2% of articles we attempted to reproduce. We have no information on the reproducibility of articles when data acquisition was too onerous from this exercise. When articles were not reproduced, the unexplained absence of data, without indication that it was due to confidentiality was the most frequent explanation.

How do these results compare with and differ from similar exercises conducted by other authors? Dewald et al. (1986) found that only 7 out of 54 (13%) articles from the JMCB were able to be reproduced $R_{full,assessed}$. In later work using the same journal, McCullough et al. (2006) found only 14 of

TABLE 16
OLS: Arcsin Citations on Reproduction Outcomes, post-2012

	(1)	(2)	(3)	(4)
'Avg. H-index'	0.019*			0.051
	(0.011)			(0.049)
'Max H-index'		0.008		−0.013
		(0.005)		(0.019)
'Min H-index'			0.021	−0.015
			(0.018)	(0.035)
'Fully reproduced'	−0.084	−0.051	0.021	−0.098
	(0.251)	(0.231)	(0.225)	(0.259)
'Highest Institution Publications'	0.007*	0.007*	0.008**	0.007*
	(0.004)	(0.004)	(0.004)	(0.004)
'Highest Co-author Experience'	−0.003	0.000	−0.001	−0.005
	(0.008)	(0.008)	(0.008)	(0.008)
'Number of authors'	0.238***	0.198**	0.278***	0.260***
	(0.080)	(0.083)	(0.084)	(0.091)
'Author at US university'	−0.078	−0.057	−0.034	−0.073
	(0.175)	(0.177)	(0.176)	(0.177)
'Solo-authored'	0.028	−0.005	−0.064	0.025
	(0.238)	(0.238)	(0.242)	(0.249)
'Avg. H-index':'Fully reproduced'	0.023			−0.065
	(0.015)			(0.088)
'Max H-index':'Fully reproduced'		0.015*		0.038
		(0.009)		(0.040)
'Min H-index':'Fully reproduced'			0.028	0.057
			(0.024)	(0.053)
Constant	3.010***	3.120***	2.950***	2.980***
	(0.311)	(0.307)	(0.329)	(0.340)
Observations	129	129	129	129
Adjusted R^2	0.259	0.239	0.237	0.245

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
Notes:
YTD citations are cumulative citations to the article in Year 4.
H-index measures are computed across all authors of an article, in the year the article was published. An author without citations has an h-index of 0.
Full or partial reproduction are defined in the text.
Sample restricted to articles published in 2012 and later, with attempted reproduction.

196 ($R_{full,assessed}$ =7.1%) articles selected from the JMCB were reproducible. However, McCullough et al. (2006) note that only 62 reproductions were attempted, which yields instead a reproduction rate of $R_{full,nonconf} = 22.6\%$. Chang and Li (2015), using slightly different methodology in selection, selected all articles from 13 well-regarded economics journals satisfying certain criteria (empirical paper using data on U.S. gross domestic product), and successfully

TABLE 17
OLS: ArcSinH Citations - Dynamic Effect

	(1)	(2)	(3)
'Avg. H-index'	0.035***	0.034***	0.042***
	(0.004)	(0.009)	(0.010)
'Fully reproduced'	0.125	−0.221	−0.210
	(0.125)	(0.273)	(0.302)
'Years since publication'		0.585***	0.657***
		(0.116)	(0.119)
'Years squared'		−0.043**	−0.045**
		(0.018)	(0.019)
'Year >= 2019'			0.831
			(0.514)
'Avg. H-index':'Fully reproduced'	0.007	0.026*	0.032*
	(0.007)	(0.016)	(0.017)
'Fully reproduced':'Years since publication'		0.110	0.158
		(0.081)	(0.101)
'Fully reproduced':'Year >= 2019'			−0.539
			(0.771)
'Avg. H-index':'Years since publication'		−0.002	−0.004
		(0.002)	(0.003)
'Avg. H-index':'Year >= 2019'			−0.050*
			(0.027)
'Avg. H-index':'Fully reproduced':'Years since publication'		−0.006	−0.009
		(0.004)	(0.005)
'Avg. H-index':'Fully reproduced':'Year >= 2019'			−0.030
			(0.044)
Constant	3.090***	1.850***	1.710***
	(0.070)	(0.192)	(0.194)
Articles	180	180	180
Observations	808	808	808
Adjusted R^2	0.169	0.332	0.356

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$
New citations are the new citations to the article in each year 1 to 5.
H-index measures are computed across all authors of an article, in the previous year.
An author without citations has an h-index of 0.
Full or partial reproduction are defined in the text.
Results for all articles with attempted reproduction.

reproduced the results of 22 of 67 papers ($R_{full,nonconf} = 32.8\%$). Fišar et al. (2023), using faculty and graduate students as replicators, achieved a reproduction rate of ($R_{full,nonconf} = 95.3\%$), although it is not clear how different their classification of "largely reproduced" is from our classification of "partially reproduced." They also found that for 29% of articles, replicators could not obtain data, which compares to 44.5% in our case, though our

replicators were asked to obtain no data at all. Broadly, after introduction of the data and code availability policy, the reasons for non-reproducibility are similar to the ones we find: the dominant reason is absence of access to data, with various code, information, or software issues relevant in a few cases.

We note that authors in general must have been aware of the policy at the AEJ:AE before submitting: The main journal of the AEA (the American Economic Review) had been requiring data and code for four years by the time the AEJ:AE was first published, and the journal from the start required such materials. We assume that the incentive effect must be higher at this journal than at journals that make provision of such materials voluntary (evidenced *inter alia* by Fišar et al. (2023)), or simply encourage it, and that the reproduction rate of replication packages that we observed may constitute an upper limit when compared to journals with a looser data and code availability policy.

On the other hand, we intentionally used replicators from the lower part of the skill distribution. More skilled replicators, armed with a Ph.D. and years of experience, might be able to reproduce more articles, by fixing minor bugs or by being more skilled in filling in gaps in the code with textual information from the paper (see Fišar et al. 2023). Yet we emphasize that there is no reason why such gaps should exist, or such bugs would need to be fixed.

The articles considered were relatively recent, and trying to reproduce papers even just a few years older might present more difficulties related to differences in software version or unavailability of data previously accessed over the internet. To assess the authenticity of the results, we would ideally use the same software version used by the authors of an article, but such software is often difficult - and, in some cases, impossible - to find or run. Most authors did not provide software version information and, to the best of our knowledge, the journals did not, at the time, attempt to capture this information from authors.[31] Our analysis highlighted that complex changes to the code were sometimes required to reproduce the papers, and documentation was lacking or inadequate. This is in line with the assessment of reproducibility conducted by Trisovic et al. (2022), who analyzed more than 2,000 replication packages in the Harvard Dataverse that used R, and found that 60% of R files crashed even after some small correction such as directory changes or packages installation (58% when restricting to data from journals).

We also show that reproducibility of papers does not appear to provide a citation bonus. This may appear to be disappointing, given that researchers should be able to more easily build on state-of-the-art research when such research is transparent and easily reproducible. Other authors have found a positive citation impact of policies (Gleditsch et al. 2003, Hoeffler 2017).

31 McCullough et al. (2006) suggested that this be provided. The recommended README created by multiple data editors (Vilhuber et al. 2020a; 2022a) and required by this journal asks for such information.

However, their findings are generally obtained when comparing journals with and without easily observable (and well-known) data and code availability policies, with potential differences in journal quality also affecting citation levels. We only compared articles from the same journal for which we attempted reproduction, something that is not easily observable and is costly to obtain. It may thus not be particularly surprising that there is more of a reputation effect (through prior publications and the h-index) than of the reproducibility of a specific article.

We interpret the relatively high reproduction rate as indicative of the limited success of data availability policies. While such policies reduce the likelihood of materials (data and/or code) not being available, they do not ensure that such materials are functional. The large number of complex code changes necessary to obtain the results reported here are more suggestive of a need to improve technical skills of authors. Testing code prior to publication will further reduce such issues. Testing can occur internally within research groups or by using outside services (Pérignon et al. 2019), before submission. It can also be conducted by active data editors at journals, such as the aforementioned journals in economics, but also in political science. Finally, as long as the materials can be corrected where readers expect to find them, for instance by updating them on journal repositories, the use of post-publication critique, as part of general academic discourse, or through structured activities (Brodeur et al. 2023), can further enhance the reproducibility of articles in economics.

References

American Economic Association (2008) "Data availability policy,"

Anderson, R. G., W. H. Greene, B. D. McCullough, and H. Vinod (2005) "The role of data and program code archives in the future of economic research," Working Paper 2005-014C

Anderson, R. G., and A. Kichkha (2017) "Replication, meta-analysis, and research synthesis in economics," *American Economic Review* 107(5), 56–59

Bell, M., and N. Miller (2013) "How to persuade journals to accept your replication paper,"

Bernanke, B. S. (2004) "Editorial Statement," *The American Economic Review* 94(1), 404–404, ISSN 0002-8282

Berry, J., L. C. Coffman, D. Hanley, R. Gihleb, and A. J. Wilson (2017) "Assessing the rate of replication in economics," *American Economic Review* 107(5), 27–31

Bollen, K., J. T. Cacioppo, R. M. Kaplan, J. A. Korsnick, and J. L. Olds (2015) "Social, behavioral, and economic sciences perspectives on robust and reliable science," Technical report, Subcommittee on Replicability in Science, National Science Foundation Directorate for Social, Behavioral, and Economic Sciences

Brodeur, A., A. Dreber, F. Hoces de la Guardia, and E. Miguel (2023) "Replication games: How to make reproducibility research more systematic," *Nature* 621(7980), 684–686

Burman, L. E., W. R. Reed, and J. Alm (2010) "A call for replication studies," *Public Finance Review* 38(6), 787–793

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016) "Evaluating replicability of laboratory experiments in economics," *Science* 351(6280), 1433–1436

Camerer, C. F., A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, and H. Wu (2018) "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour* 2(9), 637–644, ISSN 2397-3374

Chang, A. C., and P. Li (2015) "Is economics research replicable? sixty published papers from thirteen journals say "usually not"," Finance and Economics Discussion Series 2015-83, Board of Governors of the Federal Reserve System (U.S.)

——— (2017) "A preanalysis plan to replicate sixty economics research papers that worked half of the time," *American Economic Review* 107(5), 60–64

Christensen, G., and E. Miguel (2018) "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature* 56(3), 920–980

Christian, T.-M. L., S. Lafferty-Hess, W. G. Jacoby, and T. M. Carsey (2018) "Operationalizing the replication standard: A case study of the data curation and verification workflow for scholarly journals,"

Clemens, M. A. (2015) "The meaning of failed replications: A review and proposal," *Journal of Economic Surveys* 31(1), 326–342

Coffman, L. C., M. Niederle, and A. J. Wilson (2017) "A proposal to organize and promote replications," *American Economic Review* 107(5), 41–45

Dewald, W. G., J. G. Thursby, and R. G. Anderson (1986) "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project," *American Economic Review* 76(4), 587–603

Duflo, E., and H. Hoynes (2018) "Report of the search committee to appoint a data editor for the aea," *AEA Papers and Proceedings* 108, 745

Duvendack, M., R. Palmer-Jones, and W. R. Reed (2017) "What is meant by "replication" and why does it encounter resistance in economics?," *American Economic Review* 107(5), 46–51

Fišar, M., B. Greiner, C. Huber, E. Katok, A. Ozkes, and the Management Science Reproducibility Collaboration (2023) "Reproducibility in management science," Working paper, Wirtschaftsuniversität Wien

Frisch, R. (1933) "Editor's note," *Econometrica* 1(1), 1–4

Glandon, P. (2011) "Report on the american economic review data availability compliance project," *Appendix to American Economic Review Editors Report*

Gleditsch, N. P., C. Metelits, and H. Strand (2003) "Posting your data: Will you be scooped or will you be famous," *International Studies Perspectives* 4(1), 89–97

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

——— (2017) "Replication in labor economics: Evidence from data and what it suggests," *American Economic Review* 107(5), 37–40

Herbert, S., H. Kingi, F. Stanchi, and L. Vilhuber (2021) "The reproducibility of economics research: A case study," Working paper, Banque de France

Hirsch, J. E. (2005) "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America* 102(46), 16569–16572

Hoeffler, J. H. (2017) "Replication and economics journal policies," *American Economic Review* 107(5), 52–55

Huntington-Klein, N., A. Arenas, E. Beam, M. Bertoni, J. R. Bloem, P. Burli, N. Chen, P. Grieco, G. Ekpe, T. Pugatch, M. Saavedra, and Y. Stopnitzky (2021) "The influence of hidden researcher decisions in applied microeconomics," *Economic Inquiry* 59(3), 944–960, ISSN 1465-7295

Höffler, J. H. (2017) "ReplicationWiki: Improving transparency in social sciences research," *D-Lib Magazine* 23(3/4)

International DOI Foundation (IDF) (2012) "The Digital Object Identifier system home page,"

Jacoby, W. G., S. Lafferty-Hess, and T.-M. Christian (2017) "Should Journals Be Responsible for Reproducibility?,"

King, G. (1995) "Replication, replication," *PS: Political Science and Politics* 28(3), 443–499

Kingi, H., L. Vilhuber, S. Herbert, and F. Stanchi (2018) "The reproducibility of economics research: A case study," mimeo, BITSS

Mas, A. (2019) "Report of the Editor: American Economic Journal: Applied Economics," *AEA Papers and Proceedings* 109, 639–645, ISSN 2574-0768

McCullough, B. D., K. A. McGeary, and T. D. Harrison (2006) "Lessons from the JMCB Archive," *Journal of Money, Credit and Banking* 38(4), 1093–1107

McCullough, B. D., and H. D. Vinod (2003) "Econometrics and software: Comments," *Journal of Economic Perspectives* 17(1), 223–224

Menkveld, A. J., A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, M. Razen, U. Weitzel, D. Abad, M. M. Abudy, T. Adrian, Y. Ait-Sahalia, O. Akmansoy, J. Alcock, V. Alexeev, A. Aloosh, L. Amato, D. Amaya, J. Angel, A. Bach, E. Baidoo, G. Bakalli, A. Barbon, O. Bashchenko, P. C. Bindra, G. H. Bjonnes, J. Black, B. S. Black, S. Bohorquez, O. Bondarenko, C. S. Bos, C. Bosch-Rosa, E. Bouri, C. T. Brownlees, A. Calamia, V. N. Cao, G. Capelle-Blancard, L. Capera, M. Caporin, A. Carrion, T. Caskurlu, B. Chakrabarty, M. Chernov, W. M. Cheung, L. B. Chincarini, T. Chordia, S. C. Chow, B. Clapham, J.-E. Colliard, C. Comerton-Forde, E. Curran, T. Dao, W. Dare, R. J. Davies, R. De Blasis, G. De Nard, F. Declerck, O. Deev, H. Degryse, S. Deku, C. Desagre, M. A. van Dijk, C. Dim, T. Dimpfl, Y. Dong, P. Drummond, T. L. Dudda, A. Dumitrescu, T. Dyakov, A. H. Dyhrberg, M. Dzieliński, A. Eksi, I. El Kalak, S. ter Ellen, N. Eugster, M. D. D. Evans, M. Farrell, E. Félez-Viñas, G. Ferrara, E. M. Ferrouhi, A. Flori, J. Fluharty-Jaidee, S. Foley, K. Y. L. Fong, T. Foucault, T. Franus, F. A. Franzoni, B. Frijns, M. Frömmel, S. Fu, S. Füllbrunn, B. Gan, T. Gehrig, D. Gerritsen, J. Gil-Bazo, L. R. Glosten, T. Gomez, A. Gorbenko, U. Güçbilmez, J. Grammig, V. Gregoire, B. Hagströmer, J. Hambuckers, E. Hapnes, J. H. Harris, L. Harris, S. Hartmann, J.-B. Hasse, N. Hautsch, X. He, D. Heath, S. Hediger, T. Hendershott, A. M. Hibbert, E. Hjalmarsson, S. A. Hoelscher, P. Hoffmann, C. W. Holden, A. R. Horenstein, W. Huang, D. Huang, C. Hurlin, A. Ivashchenko, S. R. Iyer, H. Jahanshahloo, N. Jalkh, C. M. Jones, S. Jurkatis, P. Jylha, A. Kaeck, G. Kaiser, A. Karam, E. Karmaziene, B. Kassner, M. Kaustia, E. Kazak, F. Kearney, V. van Kervel, S. Khan, M. Khomyn, T. Klein, O. Klein, A. Klos, M. Koetter, J. P. Krahnen, A. Kolokolov, R. A. Korajczyk, R. Kozhan, A. Kwan, Q. Lajaunie, F. E. Lam, M. Lambert, H. Langlois, J. Lausen, T. Lauter, M. Leippold, V. Levin, Y. Li, M. H. Li, C. Y. Liew, T. Lindner, O. B. Linton, J. Liu, A. Liu, G. Llorente, M. Lof, A. Lohr, F. A. Longstaff, A. Lopez-Lira, S. Mankad, N. Mano, A. Marchal, C. Martineau, F. Mazzola, D. Meloso, R. Mihet, V. Mohan,

S. Moinas, D. Moore, L. Mu, D. Muravyev, D. Murphy, G. Neszveda, C. Neumeier, U. Nielsson, M. Nimalendran, S. Nolte, L. L. Norden, P. O'Neill, K. Obaid, B. A. Ødegaard, P. Östberg, M. Painter, S. Palan, I. Palit, A. Park, R. Pascual, P. Pasquariello, L. Pastor, V. Patel, A. J. Patton, N. D. Pearson, L. Pelizzon, M. Pelster, C. Pérignon, C. Pfiffer, R. Philip, T. Plíhal, P. Prakash, O.-A. Press, T. Prodromou, T. J. Putniņš, G. Raizada, D. A. Rakowski, A. Ranaldo, L. Regis, S. Reitz, T. Renault, R. W. Renjie, R. Renò, S. Riddiough, K. Rinne, P. Rintamäki, R. Riordan, T. Rittmannsberger, I. Rodríguez-Longarela, D. Rösch, L. Rognone, B. Roseman, I. Rosu, S. Roy, N. Rudolf, S. Rush, K. Rzayev, A. Rzeźnik, A. Sanford, H. Sankaran, A. Sarkar, L. Sarno, O. Scaillet, S. Scharnowski, K. R. Schenk-Hoppé, A. Schertler, M. Schneider, F. Schroeder, N. Schuerhoff, P. Schuster, M. A. Schwarz, M. S. Seasholes, N. Seeger, O. Shachar, A. Shkilko, J. Shui, M. Sikic, G. Simion, L. A. Smales, P. Söderlind, E. Sojli, K. Sokolov, L. Spokeviciute, D. Stefanova, M. G. Subrahmanyam, S. Neusüss, B. Szaszi, O. Talavera, Y. Tang, N. Taylor, W. W. Tham, E. Theissen, J. Thimme, I. Tonks, H. Tran, L. Trapin, A. B. Trolle, G. Valente, R. A. Van Ness, A. Vasquez, T. Verousis, P. Verwijmeren, A. Vilhelmsson, G. Vilkov, V. Vladimirov, S. Vogel, S. Voigt, W. Wagner, T. Walther, P. Weiss, M. van der Wel, I. M. Werner, P. J. Westerholm, C. Westheide, E. Wipplinger, M. Wolf, C. C. P. Wolff, L. Wolk, W.-K. Wong, J. Wrampelmeyer, S. Xia, D. Xiu, K. Xu, C. Xu, P. K. Yadav, J. Yagüe, C. Yan, A. Yang, W. Yoo, W. Yu, S. Yu, B. Z. Yueshen, D. Yuferova, M. Zamojski, A. Zareei, S. Zeisberger, S. S. Zhang, X. Zhang, Z. Zhong, Z. I. Zhou, C. Zhou, X. S. Zhu, M. Zoican, R. C. J. Zwinkels, J. Chen, T. Duevski, G. Gao, R. Gemayel, D. Gilder, P. Kuhle, E. Pagnotta, M. Pelli, J. Sönksen, L. Zhang, K. Ilczuk, D. Bogoev, Y. Qian, H. C. Wika, Y. Yu, L. Zhao, M. Mi, L. Bao, A. Vaduva, M. Prokopczuk, A. Avetikian, and Z.-X. Wu (2023) "Non-Standard Errors,"

Mueller-Langer, F., B. Fecher, D. Harhoff, and G. G. Wagner (2018) "Replication Studies in Economics: How Many and Which Papers Are Chosen for Replication, and Why?," JRC Working Papers on Digital Economy 2018-01, Joint Research Centre (Seville site)

National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and Replicability in Science*, Washington, D.C.: National Academies Press, ISBN 978-0-309-48616-3

Open Science Collaboration (2015) "Estimating the reproducibility of psychological science," *Science* 349(6251), aac4716, ISSN 0036-8075, 1095-9203

OurResearch (2023) "OpenAlex,"

Pérignon, C., K. Gadouche, C. Hurlin, R. Silberman, and E. Debonnel (2019) "Certify reproducibility with confidential data," *Science* 365(6449), 127–128, ISSN 0036-8075, 1095-9203

Pesaran, H. (2003) "Introducing a replication section," *Journal of Applied Econometrics* 18(1), 111–111

Pollard, T. J., and J. Wilkinson (2010) "Making datasets visible and accessible: Datacite's first summer meeting," *Ariadne* 64

Priem, J., H. Piwowar, and R. Orr (2022) "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts,"

Pérignon, C., O. Akmansoy, C. Hurlin, A. Dreber, F. Holzmeister, J. Huber, M. Johannesson, M. Kirchler, A. Menkveld, M. Razen, and U. Weitzel (2022) "Reproducibility of empirical results: Evidence from 1,000 tests in finance," Research Paper FIN-2022-1467, HEC Paris

Stark, P. B. (2018) "Before reproducibility must come preproducibility," *Nature* 557(7707), 613–613

Stodden, V., J. Seiler, and Z. Ma (2018) "An empirical analysis of journal policy effectiveness for computational reproducibility," *Proceedings of the National Academy of Sciences* 115(11), 2584–2589

Sukhtankar, S. (2017a) "Replications in development economics," *American Economic Review* 107(5), 32–36

——— (2017b) "Replications in development economics," *American Economic Review* 107(5), 32–36

Sun, S. X., L. Lannom, and B. Boesch (2010) "Handle system overview,"

Trisovic, A., M. K. Lau, T. Pasquier, and M. Crosas (2022) "A large-scale study on research code quality and execution," *Nature: Scientific Data* 9(60)

Vilhuber, L. (2019) "Report by the AEA Data Editor," *AEA Papers and Proceedings* 109, 718–29, ISSN 2574-0768, 2574-0776

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

——— (2021) "Report by the AEA Data Editor," *AEA Papers and Proceedings* 111, 808–817, ISSN 2574-0768, 2574-0776

——— (2022) "Report by the AEA Data Editor," *AEA Papers and Proceedings* 112, 813–23, ISSN 2574-0768, 2574-0776

Vilhuber, L., M. Connolly, M. Koren, J. Llull, and P. Morrow (2020a) "A template README for social science replication packages,"

——— (2022a) "A template README for social science replication packages," Technical Report v1.1.0, Zenodo

Vilhuber, L., H. H. Son, M. Welch, D. N. Wasser, and M. Darisse (2022b) "Teaching for large-scale Reproducibility Verification," *Journal of Statistics and Data Science Education* 30(3), 274–281

Vilhuber, L., J. Turitto, and K. Welch (2020b) "Report by the AEA Data Editor," *AEA Papers and Proceedings* 110, 764–75, ISSN 2574-0768, 2574-0776

Vinod, H. D. (2005) "Evaluation of archived code with perturbation checks and alternatives," in *Meetings of the American Economic Association*

Vlaeminck, S. (2021) "Dawning of a new age? Economics journals' data policies on the test bench," *LIBER Quarterly: The Journal of the Association of European Research Libraries* 31(1), 1–29, ISSN 2213-056X

Wang, J., T.-y. Kuo, L. Li, and A. Zeller (2020) "Assessing and restoring reproducibility of Jupyter notebooks," in *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pp. 138–149, Virtual Event Australia: ACM, ISBN 978-1-4503-6768-4

Welch, I. (2019) "Reproducing, Extending, Updating,Replicating, Reexamining, and Reconciling," *Critical Finance Review* 8(1-2), 301–304, ISSN 21645744, 21645760

Appendix: Acronyms Used

AEA American Economic Association
AEJ:AE American Economic Journal: Applied Economics
AEJ:EP American Economic Journal: Economic Policy
AER American Economic Review
AJPS American Journal of Politial Science
API application programming interface
DOI Digital Object Identifier
EJ Economic Journal
JMCB Journal of Money, Credit and Banking
JPE Journal of Political Economy
JEEA Journal of the European Economic Association
OS operating system
QJE Quarterly Journal of Economics
ReStat Review of Economics and Statistics
VCS version control system