WORKING paper

BANQUE DE FRANCE
EUROSYSTÈME

The Reproducibility of Economics Research: A Case Study

Sylvérie Herbert[1], Hautahi Kingi[2], Flavio Stanchi[3] & Lars Vilhuber[4]

December 2021, WP #853

ABSTRACT

Given the importance of reproducibility for the scientific ethos, more and more journals have pushed for transparency of research through data availability policies. If the introduction and implementation of such data policies improve the availability of researchers' code and data, what is the impact on reproducibility? We describe and present the results of a large reproduction exercise in which we assess the reproducibility of research articles published in the *American Economic Journal: Applied Economics*, which has implemented a data availability policy since 2005. Our replication success rate is relatively moderate, with 37.78% of replication attempts successful. 68 of 162 eligible replication attempts successfully replicated the article's analysis (41.98%) conditional on non-confidential data. A further 69 (42.59%) were at least partially successful. A total of 98 out of 303 (32.34%) relied on confidential or proprietary data, and were thus not reproducible by this project. We also conduct several bibliometric analyses of reproducible vs. non-reproducible articles and show that replicable papers do not provide citation bonuses for authors.

Keywords: Replication, Reproducibility, Transparency, Replicability, Journal Policies.

JEL classification: B41, C80, C81, C87, C88.

[1] Banque de France, sylverie.herbert@banque-france.fr
[2] Facebook, hautahikingi@gmail.com
[3] Airbnb, fs379@cornell.edu
[4] Labor Dynamics Institute, Cornell University, lars.vilhuber@cornell.edu. This work has benefited from insightful comments by participants of the Banque de France DGSEI seminar as well as the Berkeley BITSS conference. We thank Christophe Pérignon for his discussion.

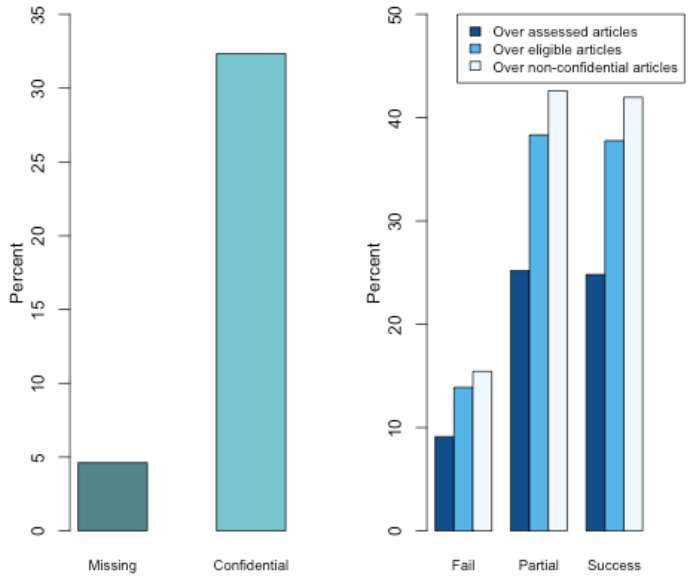
NON-TECHNICAL SUMMARY

Like for many scientific disciplines, transparency and openness are essential to the credibility of economics research. This is especially true given that economics research informs economic policy decisions. Indeed, in recent years, governments and policy institutions have pushed for "evidence-based" policy-making. They strive to base their policies on academic research, hence the need for economics research to be trustworthy for the policies to be credible. To be trustworthy, research needs to be *reproducible*.

The peer review process in economics journals ensures that original research is published and provide a stamp of "high quality". However, editors and referees do not have the obligation to check that "same data and same code give the same results". Hence, the referee process does not provide any feedback on the reproducibility. In an effort to foster reproducibility, journals have put in place data availability policies (DAP henceforth) as early as 1933 with *Econometrica*. Other journals followed quite early as well, such as the *Journal of Money, Credit and Banking* in the late 1990's. The first "top 5" journal, the A*merican Economic Review,* introduced a DAP in 2005, with the *Quarterly Journal of Economics* following in 2016. As of 2017, only 54% of 343 economic journals of the Thomson Reuters Social Science citation index had a DAP (Höffler 2017).

In this study, we aim to test whether data availability policies with light enforcement, such as the one enforced by the A*EA* in 2005, yields reproducible results. While there is no systematic check of the codes with such policies, making codes and data publicly available should in theory enhance transparency. Providing undergraduate students only with the code, data and information provided by the authors, we verify if they successfully replicate the results in the paper.

Proportion of articles with confidential and missing data (left) and reproducibility ratio (right)



Note: ratio of articles with missing and confidential data to the left, over the sample of assessed articles. Failure, partial and full reproducibility ratio according to different definitions of the sample (to the right).

First, we find a moderate replication success, in spite of a data availability policy. When considering the sample of assessed articles, only 25% were successfully replicated, rising to 38% when considering eligible papers only (those for which authors did not specify using confidential data). Conditional on non-confidential data, we find a higher reproducibility rate of 43%. Such low reproducibility is driven by confidential or missing data. Even restricting on non-confidential data, our analysis shows that a substantial numbers of articles had different results than those obtained from their replication packages. Secondly, we show that even reproducible articles required complex code changes to reach similar results as in the paper. Our analysis suggests that data availability policies are necessary but not sufficient to ensure full replicability.

Most importantly, we further show that replicated papers do not generate a citation bonus. Increased citations may thus not be enough of an incentive device to reach transparent documentation and coding practices. Our analysis therefore calls for a systematic check of articles during the referee process, as a way to reach a "good reproducibility" equilibrium.

La reproductibilité de la recherche en économie : un cas d'étude

RÉSUMÉ

Étant donné l'importance de la reproductibilité pour l'éthique scientifique, de plus en plus de revues ont poussé à la transparence de la recherche par le biais de politiques de disponibilité des données. Si l'introduction et la mise en œuvre de telles politiques de données améliorent la disponibilité du code et des données des chercheurs, quel est l'impact sur la reproductibilité ? Nous décrivons et présentons les résultats d'un vaste exercice de reproduction dans lequel nous évaluons la reproductibilité des articles de recherche publiés dans l'*American Economic Journal: Applied Economics*, qui a mis en place une politique de disponibilité des données depuis 2005. Notre taux de succès de réplication est relativement modéré, avec 37,78 % de tentatives de réplication réussies. 68 des 162 tentatives de réplication admissibles ont réussi à reproduire l'analyse de l'article (41,98 %), sous réserve de données non confidentielles. 69 autres tentatives (42,59 %) ont abouti à une réplication partielle. Un total de 98 sur 303 (32,34 %) reposaient sur des données confidentielles ou propriétaires, et n'étaient donc pas reproductibles par ce projet. Nous effectuons également plusieurs analyses bibliométriques des articles reproductibles par rapport aux articles non reproductibles et montrons que les articles reproductibles n'apportent pas de bonus de citation aux auteurs.

Mots-clés : réplication, reproductibilité, transparence, réplicabilité, politiques de journal.

1 Introduction

Replication, reproduction, and falsification of published articles is an important part of the scientific endeavor. It helps to make science "robust and reliable" (Bollen et al., 2015) and is 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 purpose, the reproducibilty of articles has been discussed in economics for at least thirty years[1]. While there are no systematic checks of replicabilty in most economic journals[2], some however have a data availability policy aiming at better transparency. In this paper, we carry out a large replication exercise of a journal which implements such data availability policy and test its efficiency.

As aforementioned, though not unheard of (Camerer et al., 2016; Chang and Li, 2015; Chang and Li, 2017; Höffler, 2017b), actual published reproductions or replications are rare (Bell and Miller, 2013; Duvendack et al., 2017). 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"[3] economics journals, 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 and Vinod, 2003; McCullough et al., 2006). 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]

To promote transparency, more and more journals are adopting "data and code availability" policies (the AEA announced one in 2003, the Journal of Political Economy (JPE) in 2004), though some doubt their effectiveness (Höffler, 2017a; Stodden et al., 2018). According to Duvendack et al. (2015) only 27 of the 333 economics journals listed in the Thomson Reuters *Web of Science* as of September 2013 regularly publish data and code for empirical articles, and 10 of those journals explicitly state that they publish replication studies. While that number seems low, it is higher than it was a decade earlier. More recently, the Journal of the American Statistical Association (JASA) has moved towards much more stringent replication requirements (Fuentes, 2016), and the AEA in 2017/2018 appointed as Data Editor the last author in lexicographic order of this article (Duflo and Hoynes, 2018).[5]

Making data and code available should enhance transparency hence replicability. In this paper, we set out to assess how well a particular journal's "data availability" policy, combined with light enforcement, yields *reproducible* articles. By *reproducibility* we 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,"(definition

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

[2] An exception is the American Economic Association (AEA)

[3] These were the traditional "top 5" and 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) and Hamermesh (2017) disagree.

[5] Much of the research reported in this article was started before the appointment.

articulated by Bollen et al. (2015), among others[6], which is related to the "narrow" sense of replication of Pesaran, 2003 [7]). Use of the "same procedures" may imply using the same computer code or re-implementing the statistical procedures in a different software package[8]. In simple words, the same code and data should produce similar results. Our protocol is set with 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 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.

We conducted this experiment over several summers and during the 2018 Fall semester, using the AEJ:AE as our source of articles, which we chose primarily 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," we expect that nearly all articles have some empirical component. Second, while other journals may also have theoretical or more complex empirical papers, using a variety of software, we wanted articles to be reproducible by the undergraduate student armed with knowledge of Stata and Matlab only. However, nothing in the methodology used in this paper is specific to the AEJ:AE, and it can be expanded to other journals. Part of the motivation is also to test the feasibility of "pre-publication verification" similar to what is done at the American Journal of Politial Science (AJPS) (Christian et al., 2018; Jacoby et al., 2017) and now at the AEA.

We find a moderate replication success, with a replication rate of 38 % overall, in spite of the data availability policy. 42 % of articles were successfully reproduced, conditional on available data, with an additional 43 % partially replicated. Compared for instance to the replication rate of 13% of Dewald et al. (1986), who conducted a large replication exercise on the Journal of Money, Credit and Banking (JMCB) which at the time had no similar policy, our results seem to suggest that journal policies that enhance transparency are helpful, yet not sufficient to reach full replicability. We further show that fully replicable papers do not benefit from a citation bonus, but authors' reputation seem to matter the most when it comes to citations. Our novel findings on what determines citations underline 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, from its creation, a data availability policy; (2) we provide a rationale for authors' lack of incentives to produce replicable research, absent journals' verification of reproducibility.

We start by describing, in Section 2, our methods of selection, analysis, and reproduction. We present our results in Section 3 before concluding in Section 4.

[6] A variety of replication concepts are used (Bollen et al., 2015; Clemens, 2015; Hamermesh, 2017)

[7] Hamermesh (2007) calls this "pure replication", which Christensen and Miguel (2018, p. 942) argue is the "basic standard [that] should be expected of all published economics research, and hope this expectation is universal among researchers."

[8] 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" (Pesaran, 2003, "wider" sense of replication), while *generalizability* refers to the extension of the scientific findings to other populations, contexts, and time frames, perhaps using different methods. Because there is a grey zone between these last two definitions, we will generally refer to either context as "replicability", which Hamermesh (2017) calls "scientific replication." In this text, we will use the terms as defined above when the distinction is material. However, we may refer to the overall concept of redoing the analysis as "replicability".

2 Description of Reproduction Procedure

Our replication exercise was conducted over the course of four summers from 2014 to 2018 as well as the fall semester of 2018. This section describes the procedure used to conduct the replication exercise, which was split into 3 parts. First, an "assessor" evaluated the task of replication as to the availability of the required components and its difficulty, and recorded a selection of article characteristics. A "replicator" then conducted the actual replication. Often but not always, the same person would be assessor and replicator. Finally, upon completion of the replication attempt, replicators filled out a guided report with questions about the exercise, such as whether or not the main results of the article could be replicated and, if not, the main barriers impeding a successful replication. To complement the information collected during the replication exercise, we also obtained descriptive article and author information such as citations and h-indices. Each of these steps are discussed in turn in the following subsections.

2.1 Initial Assessment

We first assessed each article. An assessor filled out a questionnaire (see Appendix C), gathering descriptive information, and providing an initial assessment of the expected level of 'replicability' of an article. Each assessor was provided the Digital Object Identifier (DOI) of the article,[9] 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");[10]
- The clarity and completeness of the documentation and of the program metadata;
- The presence of clear references to the original data provenance and a description of how to construct the initial datasets;
- Data availability (e.g., restricted access data, private data, public use data, etc.)

Although some of the responses to the questionnaire could have been captured via web-scraping tools, it is not possible to assess the completeness of the supplemental data without inspection by the assessor. 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. We will return to the aspect of machine-readability (machine-reproducibility) in the concluding discussion.

[9] 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.

[10] In theory, the author might have provided a DOI to a third-party data archive for some or all of the content. Ideally, each component – the article, the online appendix, the data and the programs – would have a separate DOI. In the case of the AEJ:AE, only the article itself is assigned a DOI. Supplemental data, programs, and online appendices are linked from the landing page associated with the article's DOI.

Based on the initial objective enumeration of the characteristics of the article and on 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 1).[11] Assessors also recorded the programming languages and separately the data storage formats 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 a dataset and, if not, the (apparent) reason why data was not provided. This included an assessment of whether the data were confidential or proprietary, for which we provided some guidance.

Table 1: 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.

2.2 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 replicator was instructed to document any changes made to the author-provided programs using a version control system (VCS).[12] The materials for each article were downloaded, and used to populate a article-specific repository. If multiple replicators worked on the same article, they would work in separate subdirectories of the same repository. In addition to recording changes in the VCS, replicators were asked to record information in two additional files. They recorded one or more Uniform Record Locators (URLs) of materials obtained in *'SRC.txt'*. And, in addition to any VCS commit messages, they were asked to provide a high-level summary of steps undertaken for the replication and results obtained in *'REPLICATION.txt'*.

[11] A score of 1 assigned to an article does not imply that its 'replicability' cannot be improved. For example, the programs provided by authors might lack a complete header or DOI for each database, but overall the article appears to be easily replicable.

[12] Until August 2018, the team used an restricted-access Subversion repository. Since September 2018, the team uses a restricted-access Git repository.

Once the VCS area was populated and all data files were 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 (Figure 1). Whenever a program required more extensive changes to run, the replicator would do so to the best of their ability. 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 available at the time of replication (Stata 13 and 14, SAS 9.4, and SPSS 23 at the time of the replication). 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 replication 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 replication along with our results in Section 3.[13]

It is important to note that the articles considered were relatively recent, and trying to replicate papers even just a few years older might present more difficulties related to differences in software version. 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 do not attempt to capture this information from authors. However, based on time-lag to publication, and the age of the articles, we expect that multiple versions of each software lie between when the authors ran their programs and when our team ran the programs. For instance, Matlab updates their software distribution twice a year - for an article published in 2010, it is likely that the version of Matlab used by the author was released in 2008 or 2009, at least six years before we replicated the article.

One way to address the issue of software versioning and other issues such as ambiguity in the documentation of programs would be to reach out to authors directly for confirmation. Unlike Dewald et al. (1986) and McCullough and Vinod (2003), however, who requested data and programs from the original authors, we only tried to obtain data that was available on the journal's archive, and replicators were explicitly instructed not to contact 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. 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 replication was marked as "not reproducible." There was no time limit for running programs.

[13] We captured the modified programs created by the replicators in the VCS. We could, therefore, capture objective measures of code changes using, for instance, the number of code lines changed. We have yet to do this.

Figure 1: Example of change to author-provided file

```
> svn diff -r961:HEAD $SVNURL/10.1257/app.5.4.92/replication-xxx/Data/
 do_files/main_analysis.do
Index: main_analysis.do
===================================================================
--- main_analysis.do (revision 961)
+++ main_analysis.do (revision 3425)
@@ -6,7 +6,7 @@
     version 11.2

 *place path here:
-global path "C:\"
+global path "\\rschfs1x\usercl\spring\xxx\Data"
 cd            "$path\output"
 use "$path\data\thefts_sales.dta",clear
 cap log close
```

2.3 Report on Reproducibility

Once the replication attempt terminated, replicators completed a questionnaire-based report on outcomes ("Exit Questionnaire") to capture information about the success or failure of the reproduction attempt and other descriptive information.[14] We asked replicators to describe the clarity and helpfulness of the documentation provided with the supplementary materials of the article. Although this information is also gathered at the assessment stage described in Section 2.1 to assign a subjective measure of ex-ante replication difficulty, a full understanding of an article's documentation quality is best left to a replicator who has gone through a replication attempt. We also captured the qualitative nature of the code changes (if any).

2.4 Other data related to the articles

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. We queried the Web of Science (Thomson-Reuters, 2016) database for each of up to five authors per article, and recorded their h-index (Hirsch, 2005) and the number of citations for each author by year, which is the raw data underlying the calculation of the h-index, as well as the search criteria used to find the author. In some cases, a simple search by author name does not yield a unique person (e.g., "Smith, Adam"), and sometimes, the metadata in Web of Science contained errors.[15] We also obtained citation statistics for each article.

[14] The print version of the online questionnaire is provided in Appendix D.

[15] For example, we only found one article for "Lawrence E. Katz" in Web of Science as of January 2016, namely the article in AEJ:AE, but did find quite a few more for "Lawrence F. Katz." While we initially thought this to be the result of some inside joke for senior economists, even the AEJ:AE website lists the author of the article as "Lawrence F. Katz," and we have no explanation for how this error could persist in Web of Science. Our search criteria adjusted for this error.

2.5 Recruitment of Replication Lab Members

Over the course of five summers, we recruited undergraduate students (typically but not always seniors) for the replication work 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 language and acceptable 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 replication 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).

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 replication can potentially be.

3 Are data policies enough to ensure full reproducibility?

In total, 303 AEJ:AE articles were assessed[16]. Table 2 presents the by-year breakdown of these articles. We first document the compliance with the data availability policy by gathering information about data availability, documentation clarity and a subjective measure of replication difficulty. We then turn to the replication results.

3.1 Descriptive statistics

3.1.1 Data compliance

While all articles had some supplementary materials, not all articles were accompanied by the datasets necessary for replication. Table 3 details whether the initial assessment declared an article to be eligible because the necessary data was present, based on the README and other materials provided by the authors. In general, if the data was not present, it was due to the confidentiality of the data. 80 articles stated that they used confidential or proprietary data and were therefore not considered for replication, along with the 14 articles with missing data for which no explanation was provided. We note that this is based on an ex-ante assessment, not based on an attempt to actual reproduce the analysis (see Table 8 for causes of reproduction failure due to datasets being missing).

Most papers had programmed codes and stored data in proprietary softwares format. The vast

[16]There were 342 assessments for these 303 articles. Articles could be assessed multiple times for a number of reasons, including explicit double-coding, "onboarding" of new replicators, and operator error. At this stage, we consolidated duplicates, which will be removed later in the analysis to keep the most successful outcomes when several outcomes were available for a given article.

Table 2: Articles by Year

2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
23	32	36	40	10	40	24	36	42	20	303

Table 3: Was Data Provided?

Reason	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total	Percent
Confidential Data	5	10	8	10	1	15	2	11	11	7	80	26.4
Data was Provided	16	22	28	27	9	23	21	23	29	11	209	68.98
No Data or Reason	2			3		2	1	2	2	2	14	4.62
Total	23	32	36	40	10	40	24	36	42	20	303	100

majority of articles used the Stata programming language for at least some portion of analysis (Table 4). This preponderance 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 relevant version of a package. 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). Again, the Stata format has proven to be quite robust, as newly released versions of Stata maintain backward compatibility to all previous 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 format allows the embedding of richer metadata, which is not feasible for CSV formats. We did not verify that metadata (variable labels, sane variable names, etc.) complied with modern data curation standards.

Table 4: Programming Languages and Data Formats

Software	Programming Language	Data Format
Stata	281	203
Not Reported	14	76
Matlab	11	3
SAS	9	
R	6	
SPSS	2	1
Excel	1	8
Eviews	0	
Fortran	0	0
Mathematica	0	
CSV		12
RDS		2
txt		2

Totals are not equal to the total number of articles because the articles could use more than one programming language or data format.

3.1.2 Documentation of replication packages

Overall, the documentation of data and code were considered clear enough to enable reproduction attempts. Indeed, the subjective measure of "reproducibility" (described in Table 1), presented in Table 5, shows a reasonably even distribution of articles across the rating scale.

Table 5: Replication Difficulty Assessment

Difficulty Rating	Number of Articles	Percent
1	64	21.12
2	64	21.12
3	66	21.78
4	37	12.21
5	72	23.76

Previous authors have pointed toward the need to improve the documentation of submitted data and programs (Chang and Li, 2015; McCullough et al., 2006). Replication attempts are made significantly easier when replicators are not required to resolve ambiguity and each figure is well documented for reproduction. Table 6 presents a summary of the documentation quality of the materials provided with the articles for which a reproduction attempt is made categorized by the year in which they were published (from the post-replication questionnaire). As evaluated by the replicators, the quality of documentation seems decent, with a majority of articles being 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%) only 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.

Table 6: Documentation Clarity

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total	Percent
Complete	10	8	17	21	4	18	15	15	19	6	133	73.89
Incomplete	4	5	6	5	3	5	3	5	7	2	45	25
No Info			1							1	2	1.11
Total	14	13	24	26	7	23	18	20	26	9	180	100

As can be seen in Table 6's totals, the post-completion response questionnaire captured information about 180 of the 209 eligible articles (see the table in appendix A for a description of articles at each stage of our analysis). However, there were 228 post-completion reports filed for these articles, 36 of which were duplicated articles resulting from multiple reproduction attempts. Of these duplicated articles, 21 arrived at different conclusions about the replication success of that article. In these cases, we kept the replication attempts that resulted in the more successful outcome. Specifically, we define an article to be a successful replication if at least one replicator was able to replicate the results. Similarly, if multiple replications of an article arrived at a "partially replicated" and "not replicated" conclusion (without a successful attempt), then we say it was partially replicated. This categorization gives the best chance for replication, results that we present in the next subsection. [17]

[17] Our final sample differs from the 209 eligible articles for two reasons. First, the completion response questionnaire captured information for some articles that were not recorded in the initial assessment questionnaire. We only kept articles for which we had both entry and exit questionnaire information. This is explained by a subsample of 2013 articles that were incorrectly

3.2 Replicability assessment

3.2.1 Extent and sources of non-replicability

We find a moderate replication success overall, regardless of the definition of reproducibility ratio we consider. We define the reproducibilty ratio as

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

where n_t is the number of articles that were either fully, partially or not replicated, with $t = (fail, partial, full)$. The denominator d_s stands for the number of articles, with $s = (assessed, eligible, nonconf)$. *Assessed* corresponds to the sample of articles for which we had both an entry and exit questionnaires, which means the assessor finished the replication exercise or assessment, as well as a unique record. This includes papers with missing or confidential data. *Eligible* articles are those identified as using non-confidential data in the preliminary analysis, and finally, *non-conf* is restricting the sample to papers amenable for replication, those that have the data[18]. During the replication exercise, results could differ in precision (small discrepancies, rounding errors) or coverage (all, some or few results being reproduced), so how do we distinguish between full and partial replication? Assessors were instructed to categorize articles as fully replicable if all numbers and figures matched up to some decimals (allowing for some minor rounding errors). Partial replication means that the replicators were able to execute the computer programs that produced the numerical values reported in the articles and that there were differences in the numerical values but they remained negligible.[19] In particular, the results were qualitatively silmilar, or the main results had to hold but other secondary results and robustness checks could still differ. It has to be noted that the categorization remained to some extent subjective to the assessor.

Table 7 presents the main results of the reproduction exercise. 68 of 180 replication attempts successfuly replicated the article's analysis (37.8%). The success rate of replication conditional on non-confidential data was 42% (68 out of 162). A further 69 (42.6%) were at least partially successful. We summarize our replication success depending on how we define the replicability ration, i.e., whether we consider assessed articles, eligible or articles with non-confidential data in figure 2. Considering all assessed articles for which we had complete records (entry and exit questionnaires), we only successfully replicated 24.8% of them. Restricting to the sample of articles eligible for replication because identified as using non-confidential data we get a higher success rate of 37.8%. Conditional on non-confidential data (after removing articles identified as using confidential data during the exercise, as it was not mentionned by authors), our success rate of 42% is thus higher.

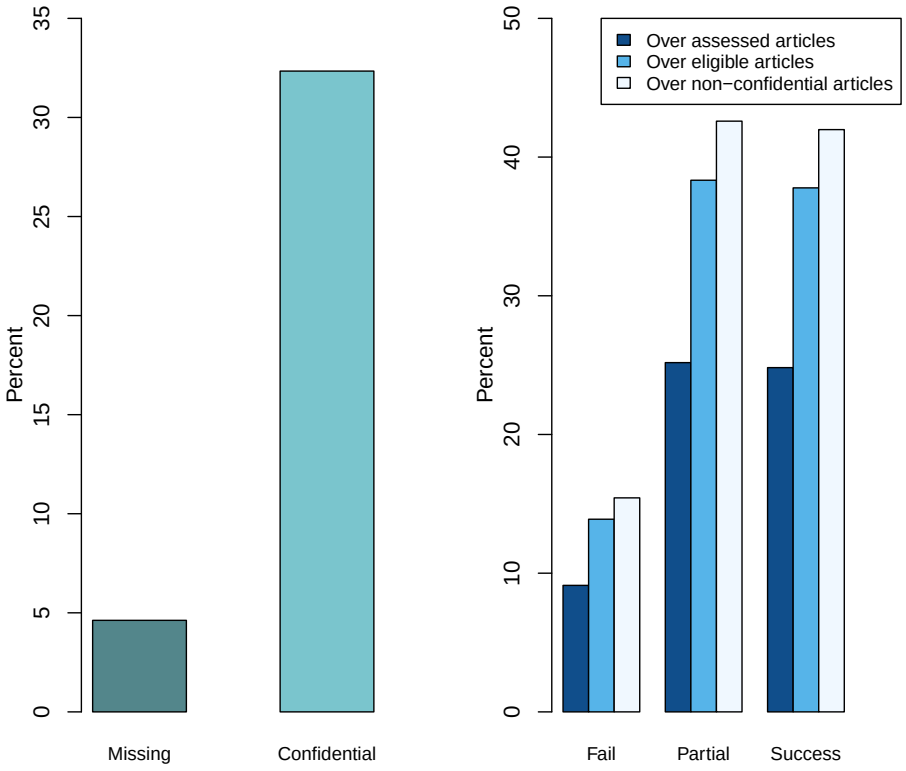
The main reason for unsuccessful reproductions is that the data used in the article was either confidential or proprietary, and therefore not available to replicators. Normally, this would imply that no replication attempt is undertaken. We have 18 cases where it would seem that assessors had missed the information that no data was available. The fact that these articles were not immediately recognized by

recorded, which explains the corresponding low 2013 number in Table 2. Second, it appears as if not all of the 209 eligible articles were attempted. However, these non-attempts were not assigned specifically so it should not affect the results.

[18] recall that this sample can differ from the previous one as assessors discovered confidential data while doing the analysis, as this was

[19] 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.

Figure 2: Summary of replicability success



Note: ratio of missing and articles with confidential data over the sample of assessed articles (on the left). Failure, partial and full reproducibility ratio (to the right) according to different definitions.

the assessor is itself an argument for better metadata on journal websites. Combining with the 80 articles identified as relying on confidential data at the initial assessment stage, a total of 98 out of 303 (32.3%) relied on confidential or proprietary data, and were thus not reproducible by this project.

Table 7: Reproduction Results

Year	Confidential Data	Unsuccessful	Successful	Partial	Total
2009	5	1	4	4	14
2010		3	7	3	13
2011		12	10	2	24
2012	2	3	8	13	26
2013			4	3	7
2014	2	2	7	12	23
2015	2		4	12	18
2016	3	1	8	8	20
2017	2	2	13	9	26
2018	2	1	3	3	9
Total	18	25	68	69	180
Percent	10	13.89	37.78	38.33	100

Table 7 lists a further 25 articles that were not able to be reproduced for other reasons. Table 8 breaks down the reasons for these unsuccessful reproductions. 4 articles did not provide the (non-confidential) data required to produce their results. Further investigation is needed to identify the reason for this apparent non-conformance to the AEJ:AE data availability policies, although we point out that some non-confidential data is 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 replication of 1 article[20], 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. For 16 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.

Table 8: Reason for Unsuccessful Reproduction

Year	Missing Data	Corrupted Data	Code Error	Software Unavailable	Other	Total
2009		1				1
2010					3	3
2011					12	12
2012	1	1	1			3
2014	1			1		2
2016	1					1
2017	1				1	2
2018		1				1
Total	4	3	1	1	16	25

While most unsuccessful replications were not due to programming errors but rather inconsistent numbers, even successful replications 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 replications required minimal work from the replicators. 44 of the 68 successful replications 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 of the

[20]For instance, one example of assessor's comment was "Could not replicate 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."

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 verions of software or operating systems. The fact that about 35.3% required complexe changed calls for at least better documenation in implementing these changes were they unavoidable, along with more robust coding practices.

Table 9: Manipulation of Code Required for Successful Reproductions

Year	Complex Change	Directory Change	No Change	Total
2009	1	1	2	4
2010	1	3	3	7
2011	4	2	4	10
2012	2	3	3	8
2013		1	3	4
2014	2	1	4	7
2015		3	1	4
2016	3	3	2	8
2017	9	2	2	13
2018	2		1	3
Total	24	19	25	68

3.2.2 Documentation and reproducibility

Good documentation is key to better reproducibility, as emphasized by many authors (Chang and Li, 2015; McCullough et al., 2006; Stark, 2018). In Table 10 we investigate whether better documentation is positively correlated with reproduction success. The results show a positive and statistically significant relationship between reproduction success of an article and the quality of its documentation, based on the Pearson and Kendall correlation tests.

Table 10: Correlation of Reproduction Success vs Documentation Quality

	Estimate	p-value
Pearson	0.30000	0.00060
Kendall	0.30000	0.00070

If clear documentation allows for better replicability, we then wonder if such "easiness" of use of replicable papers' material make these papers more widely used and cited.

3.3 Is there a citation bonus for replicable papers?

We would expect *a priori* that replicable papers provide research which can be easily built upon and that other researchers are thus more likely to use. Providing clear programs and data may increase the reputation of an author and his papers. This should lead to a higher citation count for these papers. To test for a citation bonus of papers successfully replicated, we captured bibliometric measures in 2017 for articles published through 2013, leaving a minimum of 3 years of post-publication years available to measure these metrics. We captured h-index for all authors of a paper, and computed the avarege per-paper h-index, as

well as the lowest and highest when multiple authors were present. We also computed the average annual citations of the paper. Table 12 presents a summary of these measures, categorized by reproduction success. Articles have an average of 2.2 authors and were cited, on average, 4.5 times per year. Papers with more authors tend to have more documentation (table 11).

Table 11: Clarity and Author Metrics

clarity	Number of Articles	Number of Authors
Complete	133	2.4
Incomplete	45	2.3
No Info	2	2

Table 12: Publication and Author Metrics

Outcome	Number of Articles	Avg h-index	Lowest h-index	Number of Authors	Avg Annual Citations
Unsuccessful	26	7.1	4.6	2	4.9
Partial	25	7	4.7	2.1	4
Successful	34	7.3	4.3	2.6	4.7

We investigate the relationship between the bibliometric measures, reproducibility measures and outcomes. We model the count of citations, conditional on h-index measures, reproduction outcome of the paper, the type of data used, and other covariates. In Table 13, we start by controlling for the h-index measures, interacted with an indicator whether the article used confidential data. Results indicate a positive but noisy citation bonus for papers with confidential data. Authors with a high h-index, an indicator of high citation count in the past, also seem to obtain more future citations, but there is no interaction with the use of confidential data.

Conditional on not using confidential data, how does the reproducibility of an article affect its future citation count? Tables 13 and 15 (in log terms) show that, controlling for h-indices, the ability to reproduce an article does not appear to play a significant role in the citation count. In columns (1) through (3), we control only for full reproduction, in columns (4) through (6), for full or partial reproduction. The only correlate with a strongly significant effect appears to be the authors' reputation as captured by h-index (column (2)). These results are worth highlighting as replicable papers should make it easier for other researchers to build on previous research, and therefore could be used (and quoted) more often.

4 Conclusion

In this paper, we carried out a large scale reproduction exercise of a journal that introduced a data availability policy. 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 replicated. In later work using the same journal, McCullough et al. (2006) found only 14 of 186 (7.5%) articles selected from the JMCB were reproducible. Using a different economic journal, we found a higher but nonethless moderate reproducibility rate of 38 % (42 % conditional on non-confidential data). Our main raison for failure of reproduction was the confidentiality or proprietary nature of data, but conditional on available data, the majority of failures stemmed from inconsistent numbers without convincing reasons.

Table 13: OLS: Citations vs Confidential Data

	Annual Citations		
	(1)	(2)	(3)
avghindex	3.000***		
	(0.700)		
tophindex		1.000***	
		(0.400)	
lowhindex			2.000
			(1.000)
confidential_data	17.000*	13.000*	8.000
	(9.000)	(8.000)	(9.000)
avghindex:confidential_data	−1.000		
	(1.000)		
tophindex:confidential_data		−0.500	
		(0.700)	
lowhindex:confidential_data			−0.200
			(2.000)
Constant	5.000	13.000***	20.000***
	(6.000)	(5.000)	(5.000)
N	119	119	119

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 14: OLS: Citations vs Reproduction Success

	Annual Citations					
	(1)	(2)	(3)	(4)	(5)	(6)
avghindex	4.000***			4.000**		
	(0.800)			(2.000)		
tophindex		2.000***			2.000*	
		(0.400)			(0.800)	
lowhindex			2.000*			2.000
			(1.000)			(2.000)
'Fully reproduced'	13.000	10.000	9.000			
	(10.000)	(9.000)	(10.000)			
avghindex:'Fully reproduced'	−2.000					
	(1.000)					
tophindex:'Fully reproduced'		−0.900				
		(0.700)				
lowhindex:'Fully reproduced'			−2.000			
			(2.000)			
'Full or Partial'				6.000	0.700	−3.000
				(15.000)	(11.000)	(12.000)
avghindex:'Full or Partial'				−1.000		
				(2.000)		
tophindex:'Full or Partial'					−0.200	
					(0.900)	
lowhindex:'Full or Partial'						−0.020
						(2.000)
Constant	−0.300	9.000	15.000**	−0.100	12.000	22.000**
	(7.000)	(6.000)	(7.000)	(13.000)	(10.000)	(11.000)
Observations	78	78	78	78	78	78

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 15: OLS: Log Citations vs Reproduction Success

	Annual Citations					
	(1)	(2)	(3)	(4)	(5)	(6)
avghindex	0.200***			0.200***		
	(0.030)			(0.060)		
tophindex		0.070***			0.070**	
		(0.020)			(0.030)	
lowhindex			0.100**			0.090
			(0.050)			(0.080)
'Fully reproduced'	0.600*	0.500	0.500			
	(0.400)	(0.300)	(0.400)			
avghindex:'Fully reproduced'	−0.070					
	(0.050)					
tophindex:'Fully reproduced'		−0.040				
		(0.030)				
lowhindex:'Fully reproduced'			−0.070			
			(0.080)			
'Full or Partial'				0.400	0.100	−0.100
				(0.500)	(0.400)	(0.500)
avghindex:'Full or Partial'				−0.060		
				(0.070)		
tophindex:'Full or Partial'					−0.020	
					(0.030)	
lowhindex:'Full or Partial'						0.000
						(0.090)
Constant	2.000***	2.000***	2.000***	2.000***	2.000***	3.000***
	(0.200)	(0.200)	(0.300)	(0.500)	(0.400)	(0.400)
Observations	78	78	78	78	78	78

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Such moderate replication success is not specific to this journal. 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 reproduced the results of only 22 of 67 (32.8%) papers. This seems to suggest that journal policies to enhance publications are helpful, but insufficient to foster replicabilty.

Most importantly, we showed that replicability of papers did not provide a citation bonus. This is surprising given that researchers can build on state of the art research more easily with transparent and easily replicable research. Part of the reason could be that it is sometimes easier to rewrite codes. Our analysis highlighted that complex changes to the code were frequently required to reproduce the papers, and documentation sometimes was lacking or inadequate. This is in line with the assessment of Trisovic et al. (2021) of the current state of research code. They analyzed more than 2000 replication datasets 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).

Data availability policies are thus necessary to reach transparency, but not sufficient to foster good coding practices. While our analysis sheds some light on good coding and data practices that facilitate replication, aiming at improving the way we do research, the fact that replicable papers carry no citation advantage emphasizes we will not get to this "good" equilibrium on our onw. Systematic reviews of datasets in journals with strict data sharing policy can help, as studies have shown that they result in higher re-execution (Trisovic et al., 2021).

References

Anderson, Richard G. and Areerat Kichkha (2017). "Replication, Meta-Analysis, and Research Synthesis in Economics." In: *American Economic Review* 107.5, pp. 56–59. DOI: 10.1257/aer.p20171033.

Anderson, Richard G. et al. (2005). *The Role of Data and Program Code Archives in the Future of Economic Research*. Working Paper 2005-014C.

Bell, Mark and Nicholas Miller (2013). *How to Persuade Journals to Accept Your Replication Paper*.

Berry, James et al. (2017). "Assessing the Rate of Replication in Economics." In: *American Economic Review* 107.5, pp. 27–31. DOI: 10.1257/aer.p20171119.

Bollen, Kenneth et al. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Tech. rep. Subcommittee on Replicability in Science, National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Burman, Leonard E., W. Robert Reed, and James Alm (2010). "A Call for Replication Studies." In: *Public Finance Review* 38.6, pp. 787–793. DOI: 10.1177/1091142110385210.

Camerer, Colin F et al. (2016). "Evaluating replicability of laboratory experiments in economics." In: *Science* 351.6280, pp. 1433–1436.

Chang, Andrew C. and Phillip 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." In: *American Economic Review* 107.5, pp. 60–64. DOI: 10.1257/aer.p20171034.

Christensen, Garret and Edward Miguel (2018). "Transparency, Reproducibility, and the Credibility of Economics Research." In: *Journal of Economic Literature* 56.3, pp. 920–980.

Christian, Thu-Mai Lewis et al. (2018). "Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals." In: DOI: 10.31235/osf.io/cfdba.

Clemens, Michael A. (2015). "THE MEANING OF FAILED REPLICATIONS: A REVIEW AND PROPOSAL." In: *Journal of Economic Surveys* 31.1, pp. 326–342. DOI: 10.1111/joes.12139.

Coffman, Lucas C., Muriel Niederle, and Alistair J. Wilson (2017). "A Proposal to Organize and Promote Replications." In: *American Economic Review* 107.5, pp. 41–45. DOI: 10.1257/aer.p20171122.

Dewald, William G, Jerry G Thursby, and Richard G Anderson (1986). "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." In: *American Economic Review* 76.4, pp. 587–603.

Duflo, Esther and Hilary Hoynes (2018). "Report of the Search Committee to Appoint a Data Editor for the AEA." In: *AEA Papers and Proceedings* 108, p. 745. DOI: 10.1257/pandp.108.745.

Duvendack, Maren, Richard W. Palmer-Jones, and W. Reed (2015). "Replications in Economics: A Progress Report." In: *ECON Journal Watch* 12.2, pp. 164–191.

Duvendack, Maren, Richard Palmer-Jones, and W. Robert Reed (2017). "What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics?" In: *American Economic Review* 107.5, pp. 46–51. DOI: 10.1257/aer.p20171031.

Frisch, Ragnar (1933). "Editor's Note." In: *Econometrica* 1.1, pp. 1–4.

Fuentes, Montse (2016). *Reproducible Research in JASA*. http://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/. Accessed: 2017-4-4.

Hamermesh, Daniel S. (2007). "Viewpoint: Replication in economics." In: *Canadian Journal of Economics/Revue canadienne d'économique* 40.3, pp. 715–733. DOI: 10.1111/j.1365-2966.2007.00428.x.

Hamermesh, Daniel S. (2017). "Replication in Labor Economics: Evidence from Data and What It Suggests." In: *American Economic Review* 107.5, pp. 37–40. DOI: 10.1257/aer.p20171121.

Hirsch, J. E. (2005). "An index to quantify an individual's scientific research output." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.46, pp. 16569–16572. DOI: 10.1073/pnas.0507655102.

Höffler, Jan H. (2017a). "Replication and Economics Journal Policies." In: *American Economic Review* 107.5, pp. 52–55. DOI: 10.1257/aer.p20171032.

— (2017b). "ReplicationWiki: Improving Transparency in Social Sciences Research." In: *D-Lib Magazine* 23.3/4. DOI: 10.1045/march2017-hoeffler.

International DOI Foundation (IDF) (2012). *The Digital Object Identifier System Home Page*.

Jacoby, William G., Sophia Lafferty-Hess, and Thu-Mai Christian (2017). *Should Journals Be Responsible for Reproducibility?* en.

King, Gary (1995). "Replication, Replication." In: *PS: Political Science and Politics* 28.3, pp. 443–499.

McCullough, B D and Hrishikesh D. Vinod (2003). "Econometrics and Software: Comments." In: *Journal of Economic Perspectives* 17.1, pp. 223–224.

McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison (2006). "Lessons from the JMCB Archive." In: *Journal of Money, Credit and Banking* 38.4, pp. 1093–1107.

Mueller-Langer, Frank et al. (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).

Pesaran, Hashem (2003). "Introducing a replication section." In: *Journal of Applied Econometrics* 18.1, pp. 111–111. DOI: 10.1002/jae.709.

Pollard, Tom J. and J. Max Wilkinson (2010). "Making Datasets Visible and Accessible: DataCite's First Summer Meeting." In: *Ariadne* 64.

Stark, Philip B. (2018). "Before reproducibility must come preproducibility." In: *Nature* 557.7707, pp. 613–613. DOI: 10.1038/d41586-018-05256-0.

Stodden, Victoria, Jennifer Seiler, and Zhaokun Ma (2018). "An empirical analysis of journal policy effectiveness for computational reproducibility." In: *Proceedings of the National Academy of Sciences* 115.11, pp. 2584–2589. DOI: 10.1073/pnas.1708290115.

Sukhtankar, Sandip (2017a). "Replications in Development Economics." In: *American Economic Review* 107.5, pp. 32–36. DOI: 10.1257/aer.p20171120.

— (2017b). "Replications in development economics." In: *American Economic Review* 107.5, pp. 32–36.

Sun, S. X., L. Lannom, and B. Boesch (2010). *Handle System Overview*. Corporation for National Research Initiatives.

Thomson-Reuters (2016). *Web of Science*.

Trisovic, Ana et al. (2021). *A large-scale study on research code quality and execution*. Tech. rep.

Vinod, Hrishikesh D. (2005). "Evaluation of Archived Code with Perturbation Checks and Alternatives." In: *Meetings of the American Economic Association*.

A Appendix

Table 16: Summary of data

	Count
Assessed articles	303
Assessed with complete records	274
Articles with non-confidential / non missing data	209
Eligible articles with non confidential data, complete and unique records)	180
Amenable for replication, after removing confidential data articles identified during replication	162

B Acronyms Used

AEA American Economic Association

AEJ:AE American Economic Journal: Applied Economics

AEJ:EP American Economic Journal: Economic Policy

AJPS American Journal of Politial Science

DOI Digital Object Identifier

EJ Economic Journal

JASA Journal of the American Statistical Association

JMCB Journal of Money, Credit and Banking

JPE Journal of Political Economy

JEEA Journal of the European Economic Association

OS operating system

ReStat Review of Economics and Statistics

URL Uniform Record Locator

VCS version control system

C Assessment Questionnaire

ReplicationDataQuestionnaire

Please fill out the form to the best of your abilities.

* Required

1. Please enter your NetID

2. **DOI** *
   What is the DOI (not the URL!) of the article you are reviewing? (This was sent to you by email: simply copy it here)

3. **TypeOfArticle** *
   Does the article contain empirical work, simulations, or experimental work?
   Mark only one oval.
   - ◯ Yes
   - ◯ No *After the last question in this section, stop filling out this form.*

4. **OnlineAppendix** *
   Does the article have an online Appendix?
   Mark only one oval.
   - ◯ Yes *Skip to question 5.*
   - ◯ No *Skip to question 7.*

Information on online materials

5. **OnlineAppendixURL**
   Enter the URL of the online Appendix

6. **OnlineAppendixDOI**
   Enter the DOI of the online Appendix (this is often the case if the journal provides a DOI to the article, and hosts the appendix)

Online Data

In the next few sections, we consider the following types of data: (i) input data are data as collected by the authors or another agency (examples: "CPS" or "my survey data" (ii) analysis data are the post-processed and clean data underlying specific regressions. Think of the basic workflow [input data] -> [preparation programs] -> [analysis data] -> [regression programs] -> results. If available, we will request information on up to three input datasets, and one analysis dataset.

7. **OnlineData** *
   Does the online appendix link to one or more downloadable dataset?
   Mark only one oval.
   - ◯ Yes *Skip to question 9.*
   - ◯ No

8. **OnlineDataInside** *
   Does the article itself mention where to obtain the final analysis data? (for instance, because the data are confidential or proprietary, or because there is a public-use download site for the data)
   Mark only one oval.
   - ◯ Yes
   - ◯ No *Skip to question 43.*

Information on online datasets

Please describe the first INPUT dataset.

9. **DataRunClean** *
   Does the article and/or its appendices allow you to identify the data needed to start from scratch? (Original input datasets)
   Mark only one oval.
   - ◯ Yes
   - ◯ No *Skip to question 43.*

Input dataset 1

10. **OnlineDataDOI**
    Please enter the DOI of the downloadable dataset (notation: doi://)

    .....

11. **OnlineDataHandle**
    Please enter the Handle of the downloadable dataset (notation: hdl://)

    .....

12. **OnlineDataURL**

Please enter the URL of the downloadable dataset. (this may duplicate one or the other of DOI or HDL, but is a more general way to describe it. notation: http://)

13. **DataAvailability**

Are the data available without restriction (can be downloaded or requested by anybody without restriction)? [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data]
Mark only one oval.

- Yes
- No
- DK

14. **DataAvailabilityAccess**

Do the data require users to apply for access, purchase, or otherwise sign agreements to access the data? (This should be mentioned in the Readme PDF or in the article) [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data.]
Mark only one oval.

- Yes
- No
- DK

15. **DataAvailabilityExclusive**

Are the data accessible only to the authors? [Answer yes if the authors clearly state that the data are only available to them. Answer No if there is clear evidence that others can access to the data, albeit with restrictions. Answer DK if you can't figure it out from the article.]
Mark only one oval.

- Yes
- No
- DK

16. **OtherNotes**

Any notes for this dataset that was not covered by the questions above.

17. **Do you want to describe another dataset?** *

Mark only one oval.

- ◯ Yes
- ◯ No *Skip to question 34.*

Input dataset 2

18. **OnlineDataDOI2**

Please enter the DOI of the downloadable dataset (notation: doi://)

.....

19. **OnlineDataHandle2**

Please enter the Handle of the downloadable dataset (notation: hdl://)

.....

20. **OnlineDataURL2**

Please enter the URL of the downloadable dataset. (this may duplicate one or the other of DOI or HDL, but is a more general way to describe it. notation: http://)

.....

21. **DataAvailability2**

Are the data available without restriction (can be downloaded or requested by anybody without restriction)? [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data]

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

22. **DataAvailabilityAccess2**

Do the data require users to apply for access, purchase, or otherwise sign agreements to access the data? (This should be mentioned in the Readme PDF or in the article) [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data.]

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

23. **DataAvailabilityExclusive2**

Are the data accessible only to the authors? [Answer yes if the authors clearly state that the data are only available to them. Answer No if there is clear evidence that others can access to the data, albeit with restrictions. Answer DK if you can't figure it out from the article.]
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

24. **OtherNotes2**

Any notes for this dataset that was not covered by the questions above.

.....

.....

.....

.....

.....

25. **Do you want to describe another dataset?**

Mark only one oval.

- ◯ Yes
- ◯ No *Skip to question 34.*

Input dataset 3

26. **OnlineDataDOI3**

Please enter the DOI of the downloadable dataset (notation: doi://)

.....

27. **OnlineDataHandle3**

Please enter the Handle of the downloadable dataset (notation: hdl://)

.....

28. **OnlineDataURL3**

Please enter the URL of the downloadable dataset. (this may duplicate one or the other of DOI or HDL, but is a more general way to describe it. notation: http://)

.....

29. **DataAvailability3**

Are the data available without restriction (can be downloaded or requested by anybody without restriction)? [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data]
Mark only one oval.

- Yes
- No
- DK

30. **DataAvailabilityAccess3**

Do the data require users to apply for access, purchase, or otherwise sign agreements to access the data? (This should be mentioned in the Readme PDF or in the article) [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data.]
Mark only one oval.

- Yes
- No
- DK

31. **DataAvailabilityExclusive3**

Are the data accessible only to the authors? [Answer yes if the authors clearly state that the data are only available to them. Answer No if there is clear evidence that others can access to the data, albeit with restrictions. Answer DK if you can't figure it out from the article.]
Mark only one oval.

- Yes
- No
- DK

32. **OtherNotes3**

Any notes for this dataset that was not covered by the questions above.

.....

.....

.....

.....

.....

Analysis datasets

33. **DataRunFinal**

Does the article and/or its appendices allow you to identify the data needed to run the final models?
Mark only one oval.

- ◯ Yes
- ◯ No *Skip to question 41.*

Analysis dataset

34. **OnlineFinalDataDOI**

Please enter the DOI of the downloadable dataset (notation: doi://)

.....

35. **OnlineFinalDataHandle**

Please enter the Handle of the downloadable dataset (notation: hdl://)

.....

36. **OnlineFinalDataURL**

Please enter the URL of the downloadable dataset. (this may duplicate one or the other of DOI or HDL, but is a more general way to describe it. notation: http://)

.....

37. **FinalDataAvailability**

Are the data available without restriction (can be downloaded or requested by anybody without restriction)? [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data]
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

38. **FinalDataAvailabilityAccess**

Do the data require users to apply for access, purchase, or otherwise sign agreements to access the data? (This should be mentioned in the Readme PDF or in the article) [Answer DK (Don't know) if it is not clear from the article how users can access non-downloadable data.]
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

39. **FinalDataAvailabilityExclusive**

Are the data accessible only to the authors? [Answer yes if the authors clearly state that the data are only available to them. Answer No if there is clear evidence that others can access to the data, albeit with restrictions. Answer DK if you can't figure it out from the article.]
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

40. **OtherNotesFinal**

Any notes for this dataset that was not covered by the questions above.

.....

.....

.....

.....

.....

Data formats

Considering all of the datasets described above, please check all boxes that apply.

41. **DataFormatInputs**

What format are the original input datasets in?
Check all that apply.

- ☐ Stata
- ☐ CSV
- ☐ R
- ☐ Matlab
- ☐ SPSS
- ☐ SAS
- ☐ Other: .....

42. **DataFormatAnalysis**

What format are the final analysis datasets in?
Check all that apply.

- [ ] Stata
- [ ] CSV
- [ ] R
- [ ] Matlab
- [ ] SPSS
- [ ] SAS
- [ ] Other: .....

Information on programs

43. **OnlinePrograms**

Does the online appendix have information on the programs used to run the analysis?
Mark only one oval.

- ◯ Yes *Skip to question 45.*
- ◯ No

44. **OnlineProgramsInside**

Does the article itself mention where to obtain the programs needed to replicate the study (for instance, at a data or code repository)?
Mark only one oval.

- ◯ Yes
- ◯ No

Information on online programs

45. **OnlineProgramsDOI**

Please enter the DOI of the downloadable programs (notation: doi://

.....

46. **OnlineProgramsHDL**

Please enter the Handle of the downloadable programs (notation: hdl://

.....

47. **OnlineProgramsURL**

Please enter the URL of the downloadable programs (notation: http://

.....

Documentation

48. **DocReadmePresent**

Does the downloadable data/program archive include a Readme PDF or TXT file, or some other generic instruction file?
Mark only one oval.

- ◯ Yes
- ◯ No

49. **DocReadmeContent**

Does the Readme PDF (or generic instruction file) list all included files, document the purpose and format of each file provided, and provides instruction to a user on how replication can be conducted?
Check all that apply.

- ☐ lists all included files
- ☐ documents the purpose of each file
- ☐ documents the format of each file
- ☐ provides instructions for replication

Program details

50. **ProgramFormat**

What format are the programs in?
Check all that apply.

- ☐ Stata
- ☐ R
- ☐ Matlab
- ☐ SPSS
- ☐ SAS
- ☐ Other: .....

51. **ProgramSequence**

Does the Readme PDF or one of the other included documents (including one of the programs) provide enough detail to run all the programs?
Mark only one oval.

- Yes
- No

52. **ProgramsDocumentation**

Are the programs themselves clearly documented? (There are comments throughout the program that briefly describe what is done at each step)
Mark only one oval.

- Yes
- No

53. **ProgramsHeaderAuthor**

Do the programs have a header that identifies the author? (Program metadata)
Mark only one oval.

- Yes
- No

54. **ProgramsHeaderInfo**

Do the programs have a header that identifies when they were created and/or modified (Program metadata)
Mark only one oval.

- Yes
- No

55. **ProgramsStructureManual**

Do the instructions require the user to do manual modifications to data or programs?
Mark only one oval.

- Yes
- No

56. **GeneralNotes**

General notes on this article, that wasn't captured by the questions

.....

.....

.....

.....

.....

57. **How difficult do you think replicating the article will be?** *
Mark only one oval.

	1	2	3	4	5	
easiest	◯	◯	◯	◯	◯	hardest

D Exit Questionnaire

Exit_Questionnaire_Draft

Please fill out the form to the best of your abilities.

* Required

1. Please enter your NetID

2. **DOI** *
   What is the DOI (not the URL!) of the article you reviewed?

3. **Code_Success** *
   Did you manage to eventually get all the programs to run successfully?
   Mark only one oval.
   - ◯ Yes
   - ◯ No *Skip to question 11.*

Original Program

4. **Program_Run_Clean** *
   Did the programs run "as is" without needing to make ANY changes?
   Mark only one oval.
   - ◯ Yes, no changes were necessary. *Skip to question 7.*
   - ◯ No, I needed to make changes in the code.

Changes to Program Code

5. **Directory_Change**
   Were the changes restricted to simply redirecting file/folder paths?
   Mark only one oval.
   - ◯ Yes
   - ◯ No, the changes to the code were more involved.

6. **Code_Changes**
   If the changes were more involved, briefly describe what changes you had to make.

Program vs Paper Discrepancies

7. **Output_Accuracy**
   Do the numbers produced by your program exactly match their corresponding values in the paper?
   Mark only one oval.

   - ◯ Yes
   - ◯ No, some of the numbers are different.

8. **Discrepancy_Location**
   If there are values that do not match, please list their location (ie. table number, column, page).

Skip to question 11.

Reason for replication failure.

9. Which of the following apply?
   Check all that apply.

   - ☐ Missing data set
   - ☐ Error in the program code
   - ☐ Other:

10. Briefly describe the reason why you could not replicate.

.....

.....

.....

.....

.....

Software Issues

11. **Software_Extensions**
Did you have to load any software extensions? (Eg. In matlab, the optimization toolkit is required to run the fmincon command. In Stata, outreg2 needs to be installed before running the command.)
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

12. **Software_Version**
Did the authors use a different version of software (ie. Stata11 instead of Stata13)?
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ DK

13. **First_Replicator**
Are you the first replicator?
Mark only one oval.

- ◯ Yes *Skip to question 17.*
- ◯ No

Previous Replicator Questions

14. **Common_Issues**
Did you encounter the same issues as the previous replicator?
Mark only one oval.

- ◯ Yes
- ◯ No

15. **Overcome_Issues**
Were you able to overcome any problems faced by the previous replicator?
Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ N/A. The previous replicator had no issues.

16. **Replication_Helpfulness**
Describe the usefulness of the previous replicator's notes. Did you add to them?

.....

.....

.....

.....

.....

Original Author

17. How complete was the original author's readme/generic instruction file?
Mark only one oval.

- ◯ Complete. Provided all information required to run the programs.
- ◯ Incomplete. Was ambiguous or left out crucial steps.
- ◯ No readme file was provided.

18. What actions could the authors have made to make the replication exercise easier? (Eg. correctly point to folder names)

.....

.....

.....

.....

.....

Overall Rating

19. **How difficult do you think the replication exercise was?** *
Mark only one oval.

	1	2	3	4	5	
easiest	◯	◯	◯	◯	◯	hardest

20. If this differs from the initial assessment, why?

21. **GeneralNotes**
General notes on this article/replication, that wasn't captured by the questions

E Replication Team

The following members of the Replication Lab provided valuable assistance:

Hautahi Kingi Alice Elaine Chou, Haeyong Shin, Yaxian Xie, Nathan Allan Bach, Cindy Vincens, Yuxin Chen, Flavio Stanchi, Sarah Jane Harrison, Yiwen Jiang, Jack Wendler, Jose Fernandez, Joran Isenberg, Sarah Harrison, Koonj Vekaria, Charley Chen, Yang Guo, Yiwen (Evelyn) Jiang, Noah Kwicklis, Madeline Kwicklis, Koonj Vekaria, Robin Wang, Jack Q. Wendler, Qianyan Yao, Joao Vitor Costa, Evan Shapiro, Yudi (Grace) Wang, Christopher Chang, Chuhan Liu, Daniel Kim, Kassandra Madulka, Robert Goldberg, Xinyi Wan, Siming Zou, Yu Gao, Andrew Wink, Matthew Salazar, Naomi Li, Anderson Park, Carina Chien, Nick Swan, Vendela Norman, Hayley A. Timmons, Jack VanSlyke, Gabriel Bond, Wenxin (Andee) Cao, Mcrid Wang, John Park, Xueshi Su, Sam Mbugua, Jiazhen Tan.