Draft Report on Reproducibility and Data Policy – Economica

(In progress – may contain error.)

November 3, 2023

Contents

Executive Summary

- In recent years, there has been a momentum on the part of economic journals towards adopting a data policy or upgrading the previous one. Currently, around 72% of the top 100 journals, and 76 % of the top 50 journals have some sort of data policy (for the ranking used and more details see Section 3).

- The first version of a standard for code and data availability policies, prepared by a team of data editors of economic journals, was published in December 2022. *DCAS* seems to be the first important collective step towards homogenising the reproducibility and data policy of the economic journals and is expected to be endorsed by more and more journals in the coming months and years (for more information on DCAS, see Appendix A).

- Endorsing DCAS and building a data repository "community" on Zenodo (or another comparable open data repository host) have near zero costs for any journal and seems to be a feasible contribution towards widening adoption of reproducibility policies even for the smaller journals with lower budgets.

- Enforcement, which consists of checking the documentation and the data and code package as well as re-running the codes to ensure the consistency between the paper and the linked code/data (aka "computational reproducibility") is costly, but can be performed on different levels depending on the budget constraints. The recent experience of the *Economic Inquiry* journal and *Canadian Journal of Economics* offer useful insights for smaller journals to reconcile the increasingly demanding requirements of reproducibility with their lower levels of resources.

Abbreviations

- **ICPSR**: Inter-University Consortium of Political and Social Research
- **DCAS**: Data and Code Availability Standard
- **RADE**: Restricted Access Data Environment
- **ARP**: Author Responsibility Policy ([Vlaeminck et al., 2015])
- **DAP**: Data Availability Policy ([Vlaeminck et al., 2015])

Definitions

- **Research Data**: Contains both data and code (not only code of the main analysis but also codes for preparatory stages of final datasets). (Elsevier definition),
- **Narrow Reproducibility**: obtaining consistent results using the same data and code as the original study. [Whited, 2023]
- **Reproducibility**: "the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator. So in an attempt to reproduce a published statistical analysis, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis to determine whether they yield the same results." [Bollen et al., 2015].
  Sub-categories according to [Dreber and Johannesson, 2023]
  - **Computational Reproducibility**: To what extent results in original studies can be reproduced based on data and code posted or provided by the original authors. [Dreber and Johannesson, 2023]
  - **Recreate Reproducibility**: To what extent results in original studies can be reproduced based on the information in the papers and access to the same raw data or data source, but without having access to the analysis code of the original study and/or the data set it was applied to.
  - **Robustness Reproducibility**: To what extent results in original studies are robust to alternative plausible analytical decisions on the same data.
- **Replicability**: " the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected." [Bollen et al., 2015].
  - Sometimes encompasses **generalisability**, extension of the scientific findings to other populations, contexts, and time frames. [Vilhuber, 2020]
  - [Hamermesh, 2007] calls this "scientific replication." Robustness tests performed by researchers have aspects of self-replication,
- **Grey literature**: documents such as technical reports, working papers, and so on, that are typically not subject to peer-review but are of sufficient quality to be worth preserving and citing [Vilhuber, 2020].

1 Introduction– 2023: A Year of Replications and Retractions

2023 has been a year of increasing focus on credibility, in particular replicability, of economic research. *Review of Economic Studies* retracted [Giuliano and Spilimbergo, 2014], the first retraction ever in its history, and *American Economic Review* had its second retraction ever by retracting [Boissel and Matray, 2022]. Less details were provided by *REStud* regarding their retraction: " While the original codes and data sets are no longer available, new analysis with a markedly similar data set does not support the original results." ([REStud, 2023]). In the second case, authors of the falsifying comment which led to the retraction of the paper around one year after its publication, discover an error in the submitted code and, more crucially, show that main results of the published paper are sensitive to reasonable alternative specifications [Bach et al., 2023].

These retractions, along with revived reports around well-known works of Dan Ariely and other researchers have raised concerns, particularly among younger researchers in the profession, about credibility, and public perception of credibility, of economic research. Are we doing enough with respect to replication and reproduction of published articles?

One might also wonder how these events should be thought of in the light of the increasing investment of top economic journals in improving reproducibility and replicability standards in the recent years. A plausible explanation for the higher rate of retractions and withdrawals in economic (and finance) journals since early 2010s, in particular incidence of a handful of high level retractions in the past 1-2 years, is the increased transparency and more extensive and intensive enforcement of data availability standards. Greater transparency and data sharing requirements leading to less-costly scrutiny and more retractions should be deemed as a step forward for credibility of economic research.

The fact that availability of data and code is a first yet crucial step towards replicability of economic research has long been acknowledged. This should be what Ragnar Frisch had in mind when he wrote, in the editorial note of the first issue of *Econometrica*:

> In statistical and other numerical work presented in Econometrica the original raw data will, as a rule, be published, unless their volume is excessive. This is important in order to stimulate criticism, control, and further studies. [Frisch, 1933]

The same idea was recently echoed in the response of the editorial team of *Journal of Finance* when they, in response to questions about their retraction of the award-winning paper [Rampini et al., 2020], stated that "almost non-existence" of

retractions in economic and finance journals in the past has unlikely been due to absence of errors, but most probably caused by difficulty of replication [due to unavailability of the original data and code] as well as low incentives for replication efforts. They then explain how the recent introduction of mandatory data and code sharing by *JFK* contributed to lower cost of replication by other authors and faster retraction of the original paper.

In this report, we are going to have a quick review on the *status quo* of data policies in economic research, particularly with the aim of what can be done (and what cost) for journals which has not yet adopted a data policy (due to resource constraints). The organisation of this report is as follows. In Section 2 we will review the evolution of data policies in economic journals over the past two decades. In Section 3 we offer a glimpse into the current landscape of data policy in the top-100 economic journals by looking at some descriptive statistics on types and components of data policy adopted by those journals. In Section 4 we aim to think about costs and benefits of upgrading data policy for journals with lower resources and conclude [1].

[1]I thank Joan Llull for his helpful comments, particularly with regard to different "models" currently in practice.

2 Data Policies in Economic Journals Since Early 2000s

2.1 *JPE* and Failure of the "Replication Market"

It goes without saying that replicability is necessary for credibility of findings in economics. The question is whether evolution and expansion of modern applied and empirical economic research since 1970s, has been accompanied and solidified by a proportionate number of replication exercises. Odds are that you answer in the negative to this question, as the majority of the handful of studies on replicability and reproducibility of economic papers give grounds for such a negative evaluation. [Dewald et al., 1986], one of the earliest serious investigations of reproducibility of economic research, undertook the task of verifying reproducibility of all empirical articles published or accepted between 1980-1984 (or still under review in 1984) in the *Journal of Money, Credit and Banking (JMCB)* (one of the pioneering journals with regard to data archiving and data policy). After contacting the authors twice (in more than 6 months), out of the total 154 article, they could obtain sufficiently complete data just for 8 of the articles. Their attempt in replicating 8 articles succeeded only for 5 of them[2]. More recently, [Chang and Li, 2015] attempt to replicate 67 published articles in "13 well-regarded economics journals using author-provided replication files that include both data and code." Excluding 8 papers which use confidential data or proprietary not-easily-accessible software, they succeeded in replicating no more than 29 (7 of these only with further assistance by authors), that is less than 49% leading them to conclude that economic research is more often not replicable. Even as late as early July 2023 and despite all the progress that have been made in the past couple of years, Joan Lull, the newly appointed data editor of the Econometric Society, stated that still around 60% of data and code archives submitted would not reproduce the results (before verification).

Note that these dismal statistics pertain to the less ambitious concept of (computational) reproducibility: being able to successfully run a set of programs used by the authors of original papers on the original datasets, to obtain the same quantitative results as in the original papers. But reproducibility is hardly an end in itself; even with full computational reproducibility, the codes or data transformations might still contain errors, and even if they are error-free, they might simply be deemed as poor matches to the theoretical variables. Indeed, the criterion which is more closely-linked to credibility of findings in (economics) research is "replicability"[3] which deals with threats that are deeper and more complicated than those facing reproducibility, such as publication bias, specification searching and excessive occurrence of type-1 error ([Christensen and Miguel, 2018]). In the light of the evidence that errors in

[2]The authors mention their private correspondence with editor of another top journal confirming their belief that these sound like good estimates for other major journals, not peculiar to JMCB.

[3][Maniadis et al., 2017] develop a model of replication to find conditions for replications to succeed in safeguarding the credibility of economic research.

code and data are quite common (though in many cases not consequential for the results at stake), if a significant portion of publications in economics still fail to pass the looser requirements of reproducibility, one might wonder, how would they fare with respect to replicability and its implications for credibility of economic research.

Again, these concerns are not new: around half a century ago, [Feige, 1975] published a comment in the *Journal of Political Economy*, warning against common practice of economic journals in discounting (if not discarding) "non-significant" results and how this, on top of a lack of replicability exercises, can lead to "data-grubbing" and specification-hunting, thereby leading to a proliferation of Type-I error in the published articles. To avoid such an outcome, Feige argued, the journals need "as a minimum standard" to require authors to fully report their procedures and data, but also to filter articles more based on research design rather than final results (more similar to the way grants are awarded to research proposals). The editors of *JPE*, while corroborating that these concerns are now widely acknowledged, denounced the proposed solution of Feige, on the basis its being both extremely costly and incapable of producing the right incentives, and propsed an alternative solution:

> We believe that the true remedy is resort to the powerful force of competition. We believe that journals should be prepared to accept alternative statistical tests of a hypothesis, in which either the confirmation or the contradiction of the author's statistical tests is reported. For this task to be reasonably economical, any author should be willing to provide his underlying data to other scholars (at cost). Indeed, this behavior is a requirement for responsible scholarship.

They thus announced that they would add a new section to the *JPE* called "Confirmations and Contradictions" which would publish brief yet comprehensive replication exercises (with no submission fees).

But, results of introducing this replication section in *JPE* (and another journal, *Journal of Consumer Research* ([Mayer, 1980]) was not encouraging. From 1976 to 1999, a total of 6 replication notes were published in the replication section of *JPE* (5 of these belong to the period before 1987, of which only one was "successful" in replicating the original results,[Duvendack et al., 2015]). The section was quietly left to die thereafter.

Why such poor results? On the one hand replication is costly, in particular when the replicators do not have access to the original data and code. One the other hand, there is an incentive problem, replication is not generally regarded as a reward-worthy exercise. To quote [Dewald et al., 1986]:

> Thomas Kuhn (1970) emphasized that replication-however valuable in the search for knowledge-does not fit within the "puzzle-solving" paradigm

> which defines the reward structure in scientific research. Scientific and professional laurels are not awarded for replicating another scientist's findings. Further, a researcher undertaking a replication may be viewed as lacking imagination and creativity, or of being unable to allocate his time wisely among competing research projects. In addition, replications may be interpreted as reflecting a lack of trust in another scientist's integrity and ability, as a critique of the scientist's findings, or as a personal dispute between researchers. Finally, ambiguities and/or errors in the documentation of the original research may leave the researcher unable to distinguish between errors in the replication and in the original study.

In other words, replication exercises are public goods with significant positive externalities that at the same time incur large costs to individual researchers who are going to conduct them, without much benefits accruing to them as scholars. Facing such roadblocks, further progress seemed to be contingent upon changing the reward structure in a way that it becomes more conducive to replication exercises, as well as finding ways to cut the costs of replication and reproduction.

One way to achieve the last point could be to revisit the editorial policy suggested by Ragnar Frisch and quoted above: journals can make it mandatory for the authors to make their data and code publicly available. Before their papers get accepted, authors would have significant incentives to provide their data if it is a requirement for their paper getting published, and it will be much less costly for them to do so while conducting their research than for other researchers trying to redo all the similar steps on their own. This can hugely cut the costs of reproducing results of the original paper, which can make it much easier to further replicate those findings. Furthermore, the authors who make their data and code available, can make it easier for other researchers to build upon their results and methods, and get rewarded by getting cited more [4]. So data availability policy, appeared to be a naturally next step, and this was exactly the step that editors of *Journal of Money, Credit and Banking (JMCB)* took in 1982.

2.2 *JMCB* and Data Availability Policy

JMCB put into force their "data availability policy" in 1982; adopted an editorial policy of requesting from authors the programs and data used in their articles and making these programs and data available to other researchers on request. Results were encouraging; when as a sequel, they attempted to analyse the effect of this

[4] [Renfro, 2004] quotes Stephen Hall:"Another good example is the contrast between the Quandt disequilibrium models that were around in the 70s and 80s and the Hamilton Markov switching. These two models are actually very closely related but the Quandt models never caught on because essentially Dick Quandt never gave out the software to implement them. Hamilton on the other hand gave out GAUSS code for everyone to do it and so created a whole industry."

editorial policy on practical availability of data and code for reproduction, the submission rate of data and code for articles published before the announcement of the data policy was around 34%, whereas for articles published after the announcement it rose to 74%([Dewald et al., 1986]). More importantly, a mandatory data and code policy seems to have led to less errors on the part of authors:

> Our findings suggest that the existence of a requirement that authors submit to the journal their programs and data along with each manuscript would significantly reduce the frequency and magnitude of errors. We found that the very process of authors compiling their programs and data for submission reveals to them ambiguities, errors, and oversights which otherwise would be undetected.

But at the same time, the experience of *JMCB* showed some limitations of such a data policy. The first apparent problem was compliance. In the second phased of the project, when a team of researchers including an editor of the journal attempted to obtain data and code for replication, there were still 26% of the authors who failed to submit their data and code [5]. The second issue was that, even among those who submitted their data and code, only 8 out of the 54 data sets were adequately documented. The most frequent problem was failure to precisely identify the source of the data. Replication issues were further exacerbated in many cases, by the mere fact that both the original data (think about *Survey of Current Business* and software packages used to analyse data are both subject to regular revisions with no guarantee for backward compatibility. This would call for more careful citation and versioning practices to be adopted by economists[6].

William Dewald, editor of *JMCB* from 1975-1983 played a pivotal role in adoption of a data policy and subsequent analysis of the results ([Dewald et al., 1986]). He was of course aware of the importance of permanence of data and code archives, so attempted to set up an on-premise archive using the good old floppy disks ([McCullough et al., 2006]). Unfortunately, subsequent events crashed his hopes of permanence, of not just the data he had archived but the very policy itself: after a change of the editorial team, not only the data policy was abandoned but also the files on floppy disks were discarded. This does not mean that the replication project pursued by Dewald was totally futile; in the same issue of *AER* in which [Dewald et al., 1986] was published, editors of *AER* announced a policy of requiring authors of the accepted articles to document their data and make it available upon request to any prospective replicator, though their policy fell short of requiring a mandatory data archive

[5] In one case, author of a paper which was submitted after the imposition of the new data policy but was still under review, failed to submit the data and code, saying that they have already lost or destroyed the data.

[6] [McCullough et al., 2006] has a more extensive discussion of lessons of the JMCB experience.

[Ashenfelter et al., 1986]. Few other journals such as *Journal of International Economics* and *Journal of Human Resources* followed suit. Furthermore, after Dewald was appointed director of research at the Federal Reserve Board of St. Louise in 1992, he again implemented a data/code archive for the *Review* and started to conduct an in-house replication verification of each article prior to publication. *Federal Reserve Bank of St. Louis Review* switched to a web-based archiving system in 1995 ([Anderson et al., 2008]).

2.3 Progress Since Early 2000s

2.3.1 Data and Code Availability Policy: *AEA* Takes the Lead

The *JMCB* itself started again to implement a mandatory data and code availability policy in 1996, with online archiving now available thanks to the growth of world wide web. More importantly, its experiment, with accompanied successes and failures well reported and analysed by [Dewald et al., 1986], could have been an instructive case to be followed up by more robust and better designed policies on the part of economic journals. However, while *Federal Reserve Bank of St. Louis Review* drew upon *JMCB*'s experiment to adopt a stronger data policy, apparently, these lessons were slow to diffuse and not sufficiently paid attention to on a more general scale. In their *AER* paper, [McCullough and Vinod, 2003], after demonstrating what can go wrong with nonlinear solvers of some of the most widely used pieces of statistical software and warning against the naiveté displayed by even some of the top econometricians with regard to the software, put the issue of "software-dependency" of some of the published results in applied economics in the more broader context of replicability and reproducibility. They first complain about an utter lack of data policy in some of the top journals:

> What is difficult to believe is that 17 years after Dewald et al. (1986), most economics journals have no such policy, e.g., Journal of Political Economy, Review of Economics and Statistics, Journal of Financial Economics, Econometrica, Quarterly Journal of Economics, and others. One cannot help but wonder why these journals do not have replication policies. Even in the qualitative discipline of history, authors are expected to make available their data...

Moreover, they criticise *AER* and few other journals which had a mere replication policy (APR), for inefficiency of their policy. While the *AER* had implemented a policy of "Details of computations sufficient to permit replication must be provided," [McCullough and Vinod, 2003] reported results of their own investigation of applied economic articles in the June 1999 issue of this journal to show that about half of the authors would not honour this policy. They proposed that the economic journals had better implement a mandatory data and code availability policy stored in an archive managed by the journal itself, something that has become possible and easily

affordable thanks to the expansion of the internet. They stated that at the time of their writing, only 3 journals had such a policy in place: *Federal Reserve Bank of St. Louis Review*, *JMCB* and *Macroeconomic Dynamics*, and a fourth journal, *Journal of Applied Econometrics* that had a similar policy but only for data (excluding the code).

In response to [McCullough and Vinod, 2003] the *American Economic Association (AEA)* announced its new (mandatory) 'data availability policy' in 2003, implemented it in 2004, and extended it to the new domain-specific journals in 2009–2012. The *JPE* announced its policy in 2004 and implemented it in 2005 and other top journals followed in the footprint of *AEA* ([Vilhuber, 2020]). *AEA* has since spearheaded a significant progress in data and code policy of the economic journals. While a data and code policy sounds, *prima facie* simple and straightforward to implement, there are intricacies to deal with. Apart from the more problematic issue of confidential and proprietary data (which are themselves different from the so-called "secret data"), there are ambiguities and misunderstanding about some of the publicly accessible data sets and how one should properly cite and archive them in order to make reproduction of the results, less daunting. For example, [Vilhuber, 2020] explains:

> Some widely used data sets are accessible by any researcher, but the license they are subject to prevents their redistribution and thus their inclusion as part of data deposits. This includes nonconfidential data sets from the Health and Retirement Study (HRS) and the Panel Study of Income Dynamics (PSID) at the University of Michigan and data provided by IPUMS at the Minnesota Population Center. The typical user will create a custom extract of the PSID and IPUMS databases through a data query system, not download specific data sets. Thus, each extract is essentially unique. Yet that same extract cannot be redistributed, or deposited at a journal or any other archive. In 2018, the PSID, in collaboration with ICPSR, has addressed this issue with the PSID Repository, which allows researchers to deposit their custom extracts in full compliance with the PSID Conditions of Use. [7]

[Vilhuber, 2020], [Vilhuber et al., 2023] and [Vilhuber, 2023] provide more details on the progress made and in particular, on current best practices.

2.3.2 Data Archiving

We already mentioned how the *JMCB* experience showcased the importance of a data archive for journals seeking a sound data policy. One of the earliest attempts

[7] For IPUMS, extracts from population samples (e.g., the 5% sample of the U.S. population census) rather than full population censuses (the 100% file) can be provided to journals for the purpose of replication. (end-note 19 of [Vilhuber, 2020])

was by *JMCB*'s Dewald trying to build a floppy-disk based data archive; an attempt which failed but was replicated by Dewald in his capacity as research director of the Federal Reserve Board of St. Louise. This last attempt later on transcended to a web-based archive starting 1995. In parallel and in-between the two attempts by Dewald, *the Journal of Applied Econometrics* set up a data archive in 1988. However, these used to be exceptions rather than norm. Even by as late as early 2000s, most journals, including those few which had set up some kind of replication policy, used to discard having a proper archive. Was it because of the costs? Probably not, because according to [Anderson and Dewald, 1994] even when National Science Foundation offered journals a free archive, editors refused to make use of it. The reluctance of economic journals to adopt and enforce a data and code archive policy for such a long time, is still a bit puzzling. Whatever the reason, it took about 2 decades for the most prestigious journals in the field of economics, to move away from that equilibrium.

It took several years for *QJE* and other top journals to adopt the *AER*'s policy. Even then, most of the journals preferred to use their own websites to archive the data and code. However, during the past several years, an increasing number of economic journals have started to set up their archives on proper data repositories, such as ICPSR, Harvard Dataverse and Zenodo, which seem to offer not just probably more longevity, but also a unique digital identifier to data and code, making citation and reproduction much less troublesome([Vilhuber, 2020]. [Currier et al., 2021] use content analysis to provide a more detailed review of current status and enforcement of data preservation policies in economic journals.

2.3.3 Documentation

As mentioned above, [Dewald et al., 1986] report lack of proper documentation to be a big issue for reproduction exercises. With data sets having become larger and larger over time, and codes and programs commensurately having got increasingly complicated, navigating the data and code of a typical article wanting proper documentation has become evermore daunting. This was confirmed by later attempts at reproduction, e.g., by [McCullough and Vinod, 2003] when they set to investigate the degree of compliance of authors who submit to *AER* with the announced (ARP) replication policy. They recount

> A third author, after several months and numerous requests, finally supplied us with six diskettes containing over 400 files—and no README file. Reminiscent of the attorney who responds to a subpoena with truckloads of documents, we count this author as completely noncompliant.

Similarly, in their attempt to replicate the papers in the *JMCB* archive, [McCullough et al., 2006] report various cases where lack of documentation of data or code makes it virtually impossible to reproduce the research, e.g.,

> One author provides no readme file and two data files with no column headers: we are supposed to guess the names of the variables!

Therefore they recommended journals to require authors to provide ReadMe files for the whole data and code files, documentation for the code and a code-book for the data.

Yet the most advanced and recent progress in this aspect is the recommendations put forth by authors of the DCAS standard, a joint effort of a small team of leading data editors of the economic journals which we will talk about more below. They have some recommendations for data/code sharing including documentation, which we have reproduced in the Appendix section A.1, as well as a template ReadMe file which we have included in Appendix section A.2.

2.3.4 Data Availability Statement

Last but not least, inclusion of a data availability statement gradually proved to be a useful component for reproduction exercises. [McCullough et al., 2006] complain that for some of papers they were attempting to reproduce, there were no datasets but also no indication or statement that the data is confidential though they were sure some of those paper had used confidential data. More importantly and contemporaneously, data availability statements are nowadays supposed to provide sufficiently detailed guidance for other researchers who seek access to the original data, including any limitations and the expected monetary and time cost of data access [Koren et al., 2022].

2.3.5 DCAS: Data and Code Availability Standard

A promising and welcome development towards a more comprehensive, thoughtfully designed and agreed-upon data policy standard for publications in economics (and social science at large) emerged in mid-December 2022, by launching of the first version of *DCAS*, the "Data and Code Availability Standard"[8]. This is the result of a joint effort by data editors of economic journals Miklós Koren (*Review of Economic Studies*), Marie Connolly (*Canadian Journal of Economics*), Joan Llull (*Economic Journal* and *Econometrics Journal*; since July 2023 *The Econometric Society*) and Lars Vilhuber (*AEA*) to come up with a well-designed standard for sharing research data and code.

As of today (12th of August of 2023), DCAS is endorsed by

- AEA Journals
- Canadian Journal of Economics

[8] https://datacodestandard.org/

- Econometric Society Journals
- Royal Economic Society Journals
- Economic Inquiry
- Review of Economic Studies

2.4 Compliance: Data Policy Verification as an Editorial Function

We already discussed how, despite the positive effects of introducing an author responsibility type of policy (as with Ear's initial replication policy) or even a stronger mandatory archive policy, compliance remains far from perfect. There are disincentives for economists to spend time on ensuring replication, compiling the code and data to get them readily available for independent researchers to conduct reproduction exercises, thoughtfully documenting everything, conceivably conducting more checks before send them out to journals are all time consuming (see [Feigenbaum and Levy, 1993] for a more detailed discussion of why economists might dislike reproducible research, for good reasons!)

[Dewald et al., 1986] point out one of the limitations of the first policy of *JMCB*:

> enforcing matters, even for the 65 authors whose manuscript were under review, only 49 responded for a request for submitting of Data and Code while they had accepted it as mandatory, with a mean response time of 130 days; and from this 49, a further 18 did not submit the codes and data, in one case based on the reason that they had already lost or destroyed the data (before a decision has been made on their manuscript.)

Even after upgrading of the *JMCB* data policy to mandatory data archiving, [McCullough et al., 2006] conclude their paper aimed at evaluating results of this policy by complaining that they could at the end just reproduce 22% of the candidate papers that were submitted under the new policy. They conclude that more checks should be done on the part of journals.

It is not surprising then that gradually, pioneering journals moved towards introducing a formal verification process as part of the editorship functions. Today about a dozen of top journals have formal enforcement processes, and this with varying degrees, which may include editorial monitoring of the contents of the supplementary materials, reexecution of computer code (verification of computational reproducibility), assessing the feasibility of data access reproduction is an integral part of a data editor's tasks, and improved archiving of data [Vilhuber, 2020]. However, having a data editor seem to be one of the more expensive components of data policy, in particular for the smaller journals.

In the next section we will have a closer look at the current state of data policies in a larger sample of journals.

3 A Glimpse into the Current Landscape of Data Policy in Economic Journals

In this section we will have a quick look into the *status quo* of research data policies of economic journals. To this end, we summarise the data policy of the top 100 journals in Economics, based on surveying of the relevant sections on their websites, such as "authors' guideline" or "data policy" (In many cases journals have two websites, one of the publisher and another of their own) [9].

While there is no universally accepted way to rank and choose journals, in this section, we follow the most recent publication on ranking of economic journals ([Ham et al., 2021]). The goal is to have a criterion for cutting the large number of journals that can be included in the statistical analysis, to get a sense of how the data policy "moments" change with the percentile (top-5, top-25, etc.), and what the current situation is for journals which are more or less on the same "tier" as *Economica*. We chose 100 to limit the size of the sample, because economica seems to be ranked between 40-70 in a couple of more widely used new rankings (it ranks 45 according to the aforesaid ranking). Note that this set of journals do not contain survey-type journals such as *JEP* or *JEL*, nor the finance journals.

3.1 Summary of the *Status Quo* of Data Policies

We investigate the data policy of journals on six dimensions:

1. Is there any kind of data and replication policy? And what kind?
2. Is there a mandatory data/code sharing policy?
3. Does it also contain mandatory documentation?
4. Is data availability statement mandatory?
5. What kind of repository do they employ (if any)?
6. Does the editors perform any verification to enforce the policies?

Table 1 shows a summary of the results. Out of the top 100 journals, 73% have some kind of data policy, 13% of them have data editors, 43% require authors to submit a data availability statement, 59% require authors to share data and code, and 41% explicitly require authors to have documentation of data/code included. Among the top 50 journals, around 80% have some kind of data policy and 70% require authors of accepted papers to share their data and code. In the following subsections, we dig into the details of each component.

[9] A caveat is, in some cases the wording is more or less vague, and journals with very similar wording of the same component of data policy, might significantly differ in how they implement that part.

Table 1: Summary of Data Policy in the Top 100 Econ Journals(%)

	Any Data Policy	Data Editor	Data Availability Statement	Data & Program Sharing	Documentation (ReadMe)
Top 5 (%)	100	60	80	100	100
Top 25 (%)	88	40	56	84	76
Top 50 (%)	76	24	42	68	60
Top 75 (%)	74.7	17.3	41.3	64	49.3
Top 100 (%)	72	13	42	58	41

3.2 Types of Data Policies

Different types of data policies currently used by journals appear to fall into 4 main categories: DCAS, AER policy, own data policies specific to the journal, and data policies of the publisher.

Table 2 shows the distribution of different types of data policies. As of 12 August of 2023, three out of the type 5 journals have endorsed DCAS, including *AER* and all journals of the AEA family. The remaining two, *QJE*[10] and *JPE* are still following the previous version of *AER* data policy. Over the whole 100 top journals, those that have adopted DCAS are:

- AEA Journals
- Canadian Journal of Economics
- Econometric Society Journals
- Royal Economic Society Journals
- Economic Inquiry
- Review of Economic Studies

For now there seems to be no journal out of top-100 that have adopted DCAS. There are also more ambiguous cases. For example, *Journal of European Economic Association* recommends DCAS and states on part of its website that their policy is "compatible with DCAS," but it is not clear whether they really require it or not.

Note that some journals seem to have adopted a data policy as suggested by their publishers. In particular Elsevier, and Springer seems to have data policies of

[10]QJE declares its data policy to be the AER data availability policy. However, it recommends the ReadMe Template of the "Social Science Data Editors" website.

different orders. For example, Most of the economic journals published by Springer, who do not follow DCAS or older AER nor they have their own data availability policy, seem to follow Springer's Type 3 Research Data Policy which mandates provision of a data availability statement but only encourages sharing of data and code (e.g., Journal of Economic Growth, Journal of Risk and Uncertainty). Occasionally They follow Type 1 or Type 2 (e.g., Public Choice, and Review of World Economy). However, Springer has planned to rule out Type 1 and 2 from Summer 2023. Visit the appendix C For more details on legacy and new Springer data policies

Regarding Elsevier, most of our journals with an Elsevier data policy, require sharing of data and code after an article has got accepted, before publication (e.g., Journal of Public Economics and International Journal of Industrial Organisation). What was not clear from the webpages, was that whether it is really enforced or not (a cursorily search for a couple of instances did not lead to any particular data archive.) Few of them require authors to submit a data availability statement. More importantly, it seems that Elsevier is trying to "push" Mendeley as the data archive of choice for journals that want to do it, though there are few cases, such as *Explorations in Economic History* in which journals have prioritised Open-ICPSR or other trusted repositories. All in all, it seems like in the majority of cases where there is no strong commitment of journal editors to DCAS or some particular data policy, the end result might be a combination of publishers' practices and editorial team preferences (these all should be taken with a grain of salt as we do not have strong evidence for them!)

Table 2: Types of Data Availability Policies in the Top 100 Econ Journals(%)

	DCAS	AER	Own Policy	Publisher's
Top 5 (%)	60	40	0	0
Top 25 (%)	44	16	20	12
Top 50 (%)	24	10	28	18
Top 75 (%)	18.7	6.7	25.3	28
Top 100 (%)	14	5	29	29

3.3 Data Editors

Few journals have specific data editors at the moment. These include Lars Vilhuber for AEA journals, Marie Connolly for the *Canadian Journal of Economics*, Joan Llull for the *Econometrics Society* journals, Miklos Koren for the *REStud*, and Florian Oswald for the *Royal Economic Society* journals.

However, there are other cases which are a bit fuzzy. For example, there are indications in the website of *Review of Economic Dynamics*, that Christian Zimmermann performs some of the functions which are usually part of data editors' job (archiving and some sort of quick checking?) Also, the journal of *Economic Inquiry* do not name any data editor on their website, but state that "a member of the editorial board will review the data archive to ensure that it meets the journal requirements."

3.4 Data and Program Sharing Requirements and Archiving

Most journals that have mandatory data and code sharing policy, manage an archive of the linked data and code of their published articles, either on their own website or on a public repository. In few cases, journals have announced a mandatory data and code sharing while it seems like they do not have a specific archive. Their guides usually recommends or suggests that authors put their data and code on a public repository (or one from the lists supported by the publisher) and include a link, e.g., in their data availability statement.

One ambiguity regarding the mandatory sharing of data and code that we alluded to above is for those journals that adopt Elsevier's data policy and announce a mandatory data sharing policy, though it is not clear whether they enforce this or not. We also mentioned that it seems like Elsevier is pushing for Mendeley Data to become the main archive for journals published by them.

Table 3: Data Repository in the Top 100 Econ Journals(%)

	ICPSR, Zenodo or Harvard Dataverse	"Public Repositories"	Journal's Website	Mendeley	Publisher's List
Top 5 (%)	100	0	0	0	0
Top 25 (%)	52	4	16	4	20
Top 50 (%)	38	10	12	12	30
Top 75 (%)	32	6.7	12	9.3	36
Top 100 (%)	26	13	9	7	33

All in all there are some journals which archive they data on one of the "trusted repositories" (see [Connolly et al., 2023] for an update and detailed guide on the importance and choice of repositories.) Overall, of the top-100 journals, 5 use a Harvard Dataverse Repository, 8 use the Open-ICPSR (including 5 AEA journals) and 6 use Zenodo. *Journal of Applied Econometrics* uses ZBW Journal Data and

Canadian Journal of Economics uses the Dataverse on Borealis. 8 journals use their own websites to archive data. See Table 3 for more details of the distribution of the repository policies. Note that the columns are not necessarily mutually exclusive. E.g., Mendeley is also an instance of Publisher's List, but we put it in a separate column to single out cases where Mendeley is the only archive named/used by the journal. Also in some cases, there are combinations like "Mendeley or a comparable public repository," etc.

4 Conclusion: Trade-offs for Smaller Journals

Replicability of research findings in a discipline is a key to its credibility. But replicability is very costly to verify, and its verification often goes beyond findings of a single manuscript and "involves agreements and disagreements among research papers which form a body of literature" around some given question or result ([Vilhuber, 2023]). What is less costly to verify and, in turn, a key prerequisite for verifying replicability, is the limited notion of reproducibility which is something that can be verified and enforced by the journals (with varying degrees of intensity depending on the budget) for any single manuscript.

Despite being a more modest goal, reproducibility, even in the narrower sense of computational reproducibility, is still a costly good to produce. It is costly for the authors because it takes time to compile the code and to document the data up to the required standards. These costs are conceivably increasing over time as both processing of data and programs become more and more complicated. There are also additional costs if the underlying data is confidential, or when the analysis is high-dimensional. Finally, there might be some equity concerns as reproducibility requirements are particularly burdensome for authors who cannot afford research assistance. But reproducibility policies, or even the more limited mandatory data and code sharing policies have benefits. They not only benefit the whole discipline, but also the individual authors by decreasing the likelihood of error on their part. To quote [Dewald et al., 1986]:

> Our findings suggest that the existence of a requirement that authors submit to the journal their programs and data along with each manuscript would significantly reduce the frequency and magnitude of errors. We found that the very process of authors compiling their programs and data for submission reveals to them ambiguities, errors, and oversights which otherwise would be undetected.

Reproducibility checks are also costly for journals. Checking the documentations to see whether they conform to the required standards, verifying that the data availability statement makes sense and can direct other researchers towards obtaining the data, and last but not least, re-running the computer programs to see whether they can produce the same graphs and tables as in the original paper are all costly to conduct. This pecuniary costs can be substantial if a journal is going to conduct checks thoroughly and systematically, as it involves hiring data editors and well-trained research assistants, while many academic journals run on tight budgets [11].

[11] While the pecuniary costs can differ substantially between journals, depending on what data policy they have and how thoroughly they are going to check the submitted packages, Lars Vilhuber gives an estimate of the total reproducibility budget to be around twice as much as an editor gets paid ([Labor Dynamics Institute, 2022])

Such costs can be particularly concerning for smaller journals, not only the pecuniary costs but also side damages such as a possible cut to submissions if the prospective submitting authors consider the data policy as burdensome. However, there are still low-cost steps that can be adopted by journals to contribute to the collective movement towards higher reproducibility. It was previously suggested that adopting the DCAS and setting up a repository on Zenodo have almost no cost. Moreover, side damages should probably not be a big concern based on the recent report of the editor of *Economic Inquiry* who that drop in the submission rate was not high (it was around 15% in the subsequent year after introducing the policy and extra checks.) Even this low cost might mean revert to zero as more and more journals move towards adopting data policies [12].

To wrap up, adopting and enforcing a data policy have clear benefits, while pecuniary costs can range from a minimum of around zero to a maximum of about twice the payment of an editor. The near zero cost policy could be to just endorse the DCAS and build a community on Zenodo or another similar trusted repository, and require authors to submit their data and code packages to the linked depository without any systematic checking and verification for the moment. The intermediate option is to check the data and documentation without rerunning all the code. The more complete solution involves a full checking which might be too costly for a small journal with lower resources. Journal of *Economic Inquiry* and the *Canadian Journal of Economics* offer useful insights into how to conduct a less demanding and basic, yet financially feasible, level of verification. For example, *Canadian Journal of Economics* checks that packages are complete and include all the documentation and the ReadMe file passes the minimum requirements, but does not run reproducibility checks [13].

[12] See [Whited, 2023], [Chang and Li, 2015] and [Anderson et al., 2008] for more detailed discussion of costs and benefits of data and replication policies.

[13] There are other options such as outsourcing to CASCAD which might have benefits for smaller journals.

A DCAS: Data and Code Availability Standard

A.1 DCAS Table

	Data	
1	Data Availability Statement	A Data Availability Statement is provided with detailed enough information such that an independent researcher can replicate the steps needed to access the original data, including any limitations and the expected monetary and time cost of data access.
2	Raw data	Raw data used in the research (primary data collected by the author and secondary data not otherwise available) is made publicly accessible. Exceptions are explained under Rule 1.
3	Analysis data	Analysis data is provided as part of the replication package unless they can be fully reproduced from accessible data within a reasonable time frame. Exceptions are explained under Rule 1.
4	Format	The data files are provided in any format compatible with commonly used statistical package or software. Some journals require data files in open, non-proprietary formats.
5	Metadata	Description of variables and their allowed values are publicly accessible.
6	Citation	All data used in the paper are cited.
	Code	
7	Data transformation	Programs used to create any final and analysis data sets from raw data are included.
8	Analysis	Programs producing the computational results (estimation, simulation, model solution, visualization) are included.
9	Format	Code is provided in source format that can be directly interpreted or compiled by appropriate software.

	Supporting materials	
10	Instruments	If collecting original data through surveys or experiments, survey instruments or experiment instructions as well as details on subject selection are included.
11	Ethics	If applicable, details are shared about ethics approval.
12	Pre-registration	If applicable, pre-registration of the research is identified and cited.
13	Documentation	A README document is included, containing a Data Availability Statement, listing all software and hardware dependencies and requirements (including the expected run time), and explaining how to reproduce the research results. The README follows the schema provided by the Social Science Data Editors' template README.
	Sharing	
14	Location	Data and programs are archived by the authors in the repositories deemed acceptable by the journal.
15	License	A license specifies the terms of use of code and data in the replication package. The license allows for replication by researchers unconnected to the original parties.
16	Omissions	The README clearly indicates any omission of the required parts of the package due to legal requirements or limitations or other approved agreements.

https://datacodestandard.org
Version 1.0 (December 15, 2022)
DOI: 10.5281/zenodo.7436134
Journal policies may vary.

A.2 DCAS Template for the README File

DCAS ReadMe Template

Template README and Guidance

> INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with `> INSTRUCTIONS`). Please ensure that a PDF is submitted in addition to the chosen native format.

Overview

> INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Data Availability and Provenance Statements

> INSTRUCTIONS: Every README should contain a description of the origin (provenance), location and accessibility (data availability) of the data used in the article. These descriptions are generally referred to as "Data Availability Statements" (DAS). However, in some cases, there is no external data used.

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

> If box above is checked and if no simulated/synthetic data files are provided by the authors, please skip directly to the section on Computational Requirements. Otherwise, continue.

> INSTRUCTIONS: - When the authors are **secondary data users** (they did not generate the data), the provenance and DAS coincide, and should describe the condition under which (a) the current authors (b) any future users might access the data. - When the data were generated (by the authors) in the course of conducting (lab or field) **experiments**, or were collected as part of **surveys**, then the description of the provenance should describe the data generating

process, i.e., survey or experimental procedures: - Experiments: complete sets of experimental instructions, questionnaires, stimuli for all conditions, potentially screenshots, scripts for experimenters or research assistants, as well as for subject eligibility criteria (e.g. selection criteria, exclusions), recruitment waves, demographics of subject pool used. - For lab experiments specifically, a description of any pilot sessions/studies, and computer programs, configuration files, or scripts used to run the experiment. - For surveys, the whole questionnaire (code or images/PDF) including survey logic if not linear, interviewer instructions, enumeration lists, sample selection criteria.

The information should describe ALL data used, regardless of whether they are provided as part of the replication archive or not, and regardless of size or scope. The DAS should provide enough information that a replicator can obtain the data from the original source, even if the file is provided.

For instance, if using GDP deflators, the source of the deflators (e.g. at the national statistical office) should also be listed here. If any of this information has been provided in a pre-registration, then a link to that registration may (partially) suffice.

DAS can be complex and varied. Examples are provided here, and below.

Importantly, if providing the data as part of the replication package, authors should be clear about whether they have the **rights** to distribute the data. Data may be subject to distribution restrictions due to sensitivity, IRB, proprietary clauses in the data use agreement, etc.

NOTE: DAS do not replace Data Citations (see Guidance). Rather, they augment them. Depending on journal requirements and to some extent stylistic considerations, data citations should appear in the main article, in an appendix, or in the README. However, data citations only provide information **where** to find the data, not **how to access** those data. Thus, DAS augment data citations by going into additional detail that allow a researcher to assess cost, complexity, and availability over time of the data used by the original author.

Statement about Rights

- ☐ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Appropriate permission are documented in the LICENSE.txt file.

(Optional, but recommended) License for Data

INSTRUCTIONS: Most data repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt file, separately from the README, possibly combined with the license for any code. Some data may be subject to inherited license requirements, i.e., the data provider may allow for redistribution only if the data is licensed under specific rules - authors should check with their data providers. For instance, a data use license might require that users - the current author, but also any subsequent users - cite the data provider. Licensing can be complex. Some non-legal guidance may be found here. For multiple licenses within a data package, the `LICENSE.txt` file might contain the concatenation of all the licenses that apply (for instance, a custom license for one file, plus a CC-BY license for another file).

NOTE: In many cases, it is not up to the creator of the replication package to simply define a license, a license may be *sticky* and be defined by the original data creator.

Example: The data are licensed under a Creative Commons/CC-BY-NC license. See LICENSE.txt for details.

Summary of Availability

☐ All data **are** publicly available.
☐ Some data **cannot be made** publicly available.
☐ **No data can be made** publicly available.

Details on each Data Source

INSTRUCTIONS: For each data source, list the file that contains data from that source here; if providing combined/derived datafiles, list them separately after the DAS. For each data source or file, as appropriate,

- Describe the format (open formats preferred, but some software-specific formats OK if open-source readers available): `.dta`, `.xlsx`, `.csv`, `netCDF`, etc.
- Provide a data dictionairy, either as part of the archive (list the file name), or at a URL (list the URL). Some formats are self-describing *if* they have the requisite information (e.g., `.dta` should have both variable and value labels).
- List availability within the package

- Use proper bibliographic references in addition to a verbose description (and provide a bibliography at the end of the README, expanding those references)

A summary in tabular form can be useful:

Data.Name	Data.Files	Location	Provided	Citation
"Current Population Survey 2018"	cepr_march_2018.dta	data/	TRUE	CEPR (2018)
"Provincial Administration Reports"	coast_simplepoint2.csv; rivers_simplepoint2.csv; RAIL_dummies.dta; railways_Dissolve_Simplify_point2.csv	Data/maps/	TRUE	Administration (2017)
"2017 SAT scores"	Not available	data/to_clean/	FALSE	College Board (2020)

where the `Data.Name` column is then expanded in the subsequent paragraphs, and `CEPR (2018)` is resolved in the References section of the README.

Example for public use data collected by the authors

> The [DATA TYPE] data used to support the findings of this study have been deposited in the [NAME] repository ([DOI or OTHER PERSISTENT IDENTIFIER]). [1]. The data were collected by the authors, and are available under a Creative Commons Non-commercial license.

Example for public use data sourced from elsewhere and provided

> Data on National Income and Product Accounts (NIPA) were downloaded from the U.S. Bureau of Economic Analysis (BEA, 2016). We use Table 30. Data can be downloaded from https://apps.bea.gov/regional/downloadzip.cfm, under "Personal Income (State and Local)", select CAINC30: Economic Profile by County, then download. Data can also be directly downloaded using https://apps.bea.gov/regional/zip/CAINC30.zip. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafile: `CAINC30__ALL_AREAS_1969_2018.csv`

Example for public use data with required registration and provided extract

> The paper uses IPUMS Terra data (Ruggles et al, 2018). IPUMS-Terra does not allow for redistribution, except for the purpose of repli-

cation archives. Permissions as per https://terra.ipums.org/citation have been obtained, and are documented within the "data/IPUMS-terra" folder. > Note: the reference to "Ruggles et al, 2018" would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: `data/raw/ipums_terra_2018.dta`

Example for free use data with required registration, extract not provided

The paper uses data from the World Values Survey Wave 6 (Inglehart et al, 2019). Data is subject to a redistribution restriction, but can be freely downloaded from http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Choose `WV6_Data_Stata_v20180912`, fill out the registration form, including a brief description of the project, and agree to the conditions of use. Note: "the data files themselves are not redistributed" and other conditions. Save the file in the directory `data/raw`.

Note: the reference to "Inglehart et al, 2018" would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: `data/raw/WV6_Data_Stata_v20180912.dta` (not provided)

Example for confidential data

INSTRUCTIONS: Citing and describing confidential data, in particular when it does not have a regular distribution channel or online landing page, can be tricky. A citation can be crafted (see guidance), and the DAS should describe how to access, whom to contact (including the role of the particular person, should that person retire), and other relevant information, such as required citizenship status or cost.

The data for this project (DESE, 2019) are confidential, but may be obtained with Data Use Agreements with the Massachusetts Department of Elementary and Secondary Education (DESE). Researchers interested in access to the data may contact [NAME] at [EMAIL], also see www.doe.mass.edu/research/contact.html. It can take some months to negotiate data use agreements and gain access to the data. The author will assist with any reasonable replication attempts for two years following publication.

Example for confidential Census Bureau data

All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata,

> follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: https://www.census.gov/ces/rdcresearch/howtoapply.html. You must request the following datasets in your proposal: 1. Longitudinal Business Database (LBD), 2002 and 2007 2. Foreign Trade Database – Import (IMP), 2002 and 2007 [. . .]

(adapted from Fort (2016))

Example for preliminary code during the editorial process

> Code for data cleaning and analysis is provided as part of the replication package. It is available at https://dropbox.com/link/to/code/XYZ123ABC for review. It will be uploaded to the [JOURNAL REPOSITORY] once the paper has been conditionally accepted.

Dataset list

> INSTRUCTIONS: In some cases, authors will provide one dataset (file) per data source, and the code to combine them. In others, in particular when data access might be restrictive, the replication package may only include derived/analysis data. Every file should be described. This can be provided as a Excel/CSV table, or in the table below.
>
> INSTRUCTIONS: While it is often most convenient to provide data in the native format of the software used to analyze and process the data, not all formats are "open" and can be read by other (free) software. Data should at a minimum be provided in formats that can be read by open-source software (R, Python, others), and ideally be provided in non-proprietary, archival-friendly formats.
>
> INSTRUCTIONS: All data files should be fully documented: variables/columns should have labels (long-form meaningful names), and values should be explained. This might mean generating a codebook, pointing at a public codebook, or providing data in (non-proprietary) formats that allow for a rich description. This is in particular important for data that is not distributable.
>
> INSTRUCTIONS: Some journals require, and it is considered good practice, to provide synthetic or simulated data that has some of the key characteristics of the restricted-access data which are not provided. The level of fidelity may vary - it may be useful for debugging only, or it should allow to assess the key characteristics of the statistical/econometric procedure or the main conclusions of the paper.

DCAS ReadMe Template

Data file	Source	Notes	Provided
`data/raw/lbd.dta`	LBD	Confidential	No
`data/raw/terra.dta`	IPUMS Terra	As per terms of use	Yes
`data/derived/regression_input.dta`	Above inputs	Combines multiple data sources, serves as input for Table 2, 3 and Figure 5.	Yes

Computational requirements

INSTRUCTIONS: In general, the specific computer code used to generate the results in the article will be within the repository that also contains this README. However, other computational requirements - shared libraries or code packages, required software, specific computing hardware - may be important, and is always useful, for the goal of replication. Some example text follows.

INSTRUCTIONS: We strongly suggest providing setup scripts that install/set up the environment. Sample scripts for Stata, R, Julia are easy to set up and implement. Specific software may have more sophisticated tools: Python, Julia.

Software Requirements

INSTRUCTIONS: List all of the software requirements, up to and including any operating system requirements, for the entire set of code. It is suggested to distribute most dependencies together with the replication package if allowed, in particular if sourced from unversioned code repositories, Github repos, and personal webpages. In all cases, list the version *you* used.

- Stata (code was last run with version 15)
  - `estout` (as of 2018-05-12)
  - `rdrobust` (as of 2019-01-05)
  - the program "`0_setup.do`" will install all dependencies locally, and should be run once.
- Python 3.6.4
  - `pandas` 0.24.2
  - `numpy` 1.16.4
  - the file "`requirements.txt`" lists these dependencies, please run "`pip install -r requirements.txt`" as the first step. See https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability for further instructions on creating and using the "`requirements.txt`" file.

- Intel Fortran Compiler version 20200104
- Matlab (code was run with Matlab Release 2018a)
- R 3.4.3
  - `tidyr` (0.8.3)
  - `rdrobust` (0.99.4)
  - the file "`0_setup.R`" will install all dependencies (latest version), and should be run once prior to running other programs.

Portions of the code use bash scripting, which may require Linux.

Portions of the code use Powershell scripting, which may require Windows 10 or higher.

Controlled Randomness

INSTRUCTIONS: Some estimation code uses random numbers, almost always provided by pseudorandom number generators (PRNGs). For reproducibility purposes, these should be provided with a deterministic seed, so that the sequence of numbers provided is the same for the original author and any replicators. While this is not always possible, it is a requirement by many journals' policies. The seed should be set once, and not use a time-stamp. If using parallel processing, special care needs to be taken. If using multiple programs in sequence, care must be taken on how to call these programs, ideally from a main program, so that the sequence is not altered.

☐ Random seed is set at line _____ of program _____

Memory and Runtime Requirements

INSTRUCTIONS: Memory and compute-time requirements may also be relevant or even critical. Some example text follows. It may be useful to break this out by Table/Figure/section of processing. For instance, some estimation routines might run for weeks, but data prep and creating figures might only take a few minutes.

Summary Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

☐ <10 minutes
☐ 10-60 minutes
☐ 1-2 hours
☐ 2-8 hours
☐ 8-24 hours
☐ 1-3 days
☐ 3-14 days
☐ > 14 days
☐ Not feasible to run on a desktop machine, as described below.

DCAS ReadMe Template

Details The code was last run on a **4-core Intel-based laptop with MacOS version 10.14.4**.

Portions of the code were last run on a **32-core Intel server with 1024 GB of RAM, 12 TB of fast local storage**. Computation took 734 hours.

Portions of the code were last run on a **12-node AWS R3 cluster, consuming 20,000 core-hours**.

> INSTRUCTIONS: Identifiying hardware and OS can be obtained through a variety of ways: Some of these details can be found as follows:
>
> - (Windows) by right-clicking on "This PC" in File Explorer and choosing "Properties"
> - (Mac) Apple-menu > "About this Mac"
> - (Linux) see code in tools/linux-system-info.sh‘

Description of programs/code

> INSTRUCTIONS: Give a high-level overview of the program files and their purpose. Remove redundant/ obsolete files from the Replication archive.

- Programs in `programs/01_dataprep` will extract and reformat all datasets referenced above. The file `programs/01_dataprep/main.do` will run them all.
- Programs in `programs/02_analysis` generate all tables and figures in the main body of the article. The program `programs/02_analysis/main.do` will run them all. Each program called from `main.do` identifies the table or figure it creates (e.g., `05_table5.do`). Output files are called appropriate names (`table5.tex`, `figure12.png`) and should be easy to correlate with the manuscript.
- Programs in `programs/03_appendix` will generate all tables and figures in the online appendix. The program `programs/03_appendix/main-appendix.do` will run them all.
- Ado files have been stored in `programs/ado` and the `main.do` files set the ADO directories appropriately.
- The program `programs/00_setup.do` will populate the `programs/ado` directory with updated ado packages, but for purposes of exact reproduction, this is not needed. The file `programs/00_setup.log` identifies the versions as they were last updated.
- The program `programs/config.do` contains parameters used by all programs, including a random seed. Note that the random seed is set once for each of the two sequences (in `02_analysis` and `03_appendix`). If running in any order other than the one outlined below, your results may differ.

DCAS ReadMe Template

(Optional, but recommended) License for Code

> INSTRUCTIONS: Most journal repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt file, separately from the README, possibly combined with the license for any data provided. Some code may be subject to inherited license requirements, i.e., the original code author may allow for redistribution only if the code is licensed under specific rules - authors should check with their sources. For instance, some code authors require that their article describing the econometrics of the package be cited. Licensing can be complex. Some non-legal guidance may be found here.

The code is licensed under a MIT/BSD/GPL [choose one!] license. See LICENSE.txt for details.

Instructions to Replicators

> INSTRUCTIONS: The first two sections ensure that the data and software necessary to conduct the replication have been collected. This section then describes a human-readable instruction to conduct the replication. This may be simple, or may involve many complicated steps. It should be a simple list, no excess prose. Strict linear sequence. If more than 4-5 manual steps, please wrap a main program/Makefile around them, in logical sequences. Examples follow.

- Edit `programs/config.do` to adjust the default path
- Run `programs/00_setup.do` once on a new system to set up the working environment.
- Download the data files referenced above. Each should be stored in the prepared subdirectories of `data/`, in the format that you download them in. Do not unzip. Scripts are provided in each directory to download the public-use files. Confidential data files requested as part of your FSRDC project will appear in the `/data` folder. No further action is needed on the replicator's part.
- Run `programs/01_main.do` to run all steps in sequence.

Details

- `programs/00_setup.do`: will create all output directories, install needed ado packages.
  - If wishing to update the ado packages used by this archive, change the parameter `update_ado` to `yes`. However, this is not needed to successfully reproduce the manuscript tables.
- `programs/01_dataprep`:

  - These programs were last run at various times in 2018.
  - Order does not matter, all programs can be run in parallel, if needed.
  - A `programs/01_dataprep/main.do` will run them all in sequence, which should take about 2 hours.
- `programs/02_analysis/main.do`.
  - If running programs individually, note that ORDER IS IMPORTANT.
  - The programs were last run top to bottom on July 4, 2019.
- `programs/03_appendix/main-appendix.do`. The programs were last run top to bottom on July 4, 2019.
- Figure 1: The figure can be reproduced using the data provided in the folder "2_data/data_map", and ArcGIS Desktop (Version 10.7.1) by following these (manual) instructions:
  - Create a new map document in ArcGIS ArcMap, browse to the folder "2_data/data_map" in the "Catalog", with files "provinceborders.shp", "lakes.shp", and "cities.shp".
  - Drop the files listed above onto the new map, creating three separate layers. Order them with "lakes" in the top layer and "cities" in the bottom layer.
  - Right-click on the cities file, in properties choose the variable "health"... (more details)

List of tables and programs

INSTRUCTIONS: Your programs should clearly identify the tables and figures as they appear in the manuscript, by number. Sometimes, this may be obvious, e.g. a program called "`table1.do`" generates a file called `table1.png`. Sometimes, mnemonics are used, and a mapping is necessary. In all circumstances, provide a list of tables and figures, identifying the program (and possibly the line number) where a figure is created.

NOTE: If the public repository is incomplete, because not all data can be provided, as described in the data section, then the list of tables should clearly indicate which tables, figures, and in-text numbers can be reproduced with the public material provided.

The provided code reproduces:

☐ All numbers provided in text in the paper
☐ All tables and figures in the paper
☐ Selected tables and figures in the paper, as explained and justified below.

Figure/Table #	Program	Line Number	Output file	Note
Table 1	02_analysis/table1.do		summarystats.csv	
Table 2	02_analysis/table2and3.do	12	table2.csv	
Table 3	02_analysis/table2and3.do	245	table3.csv	

DCAS ReadMe Template

Figure/Table #	Program	Line Number	Output file	Note
Figure 1	n.a. (no data)			Source: Herodus (2011)
Figure 2	02_analysis/fig2.do		figure2.png	
Figure 3	02_analysis/fig3.do		figure-robustness.png	Requires confidential data

References

> INSTRUCTIONS: As in any scientific manuscript, you should have proper references. For instance, in this sample README, we cited "Ruggles et al, 2019" and "DESE, 2019" in a Data Availability Statement. The reference should thus be listed here, in the style of your journal:

Steven Ruggles, Steven M. Manson, Tracy A. Kugler, David A. Haynes II, David C. Van Riper, and Maryia Bakhtsiyarava. 2018. "IPUMS Terra: Integrated Data on Population and Environment: Version 2 [dataset]." Minneapolis, MN: *Minnesota Population Center, IPUMS*. https://doi.org/10.18128/D090.V2

Department of Elementary and Secondary Education (DESE), 2019. "Student outcomes database [dataset]" *Massachusetts Department of Elementary and Secondary Education (DESE)*. Accessed January 15, 2019.

U.S. Bureau of Economic Analysis (BEA). 2016. "Table 30:"Economic Profile by County, 1969-2016." (accessed Sept 1, 2017).

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

Acknowledgements

Some content on this page was copied from Hindawi. Other content was adapted from Fort (2016), Supplementary data, with the author's permission.

B AER

B.1 AER Data and Code Availability

American Economic Association
Data and Code Availability

Corresponding Author: _____

Manuscript Title: _____

Manuscript ID: _____

Does this submission contain empirical work, simulations, or experimental work?
☐ Yes ☐ No

If yes, please complete the rest of this form.

☐ I have read the AEA's Data and Code Availability Policy, and understand my commitments under that policy.

☐ All data sources are cited and referenced *in the manuscript* as per AEA guidance.

☐ Data and code have been deposited in a trusted repository.

Please complete ONE option below:

AEA Data and Code Repository
Project Number: _____
☐ I have followed the AEA Repository deposit instructions.

Other trusted repository
Deposit DOI: https://doi.org/_____

☐ The README uses the AEA/Economics Standards template.

☐ The replication package has a master script that creates all tables, figures, and in-text numbers without manual intervention.

Is any of the data used in this manuscript subject to access restrictions? (Note: any access restrictions, including web registration requirements, must be described in the "Data Availability Statement" section of the README.)
☐ Yes ☐ No

If yes, select all that apply:
Replication package uses
- ☐ data that can be made available to the Data Editor (privately).
- ☐ data that cannot be provided, but can be obtained by replicators within a short time frame.
- ☐ data that cannot be easily accessed.

Signature: _____

Date: _____

Version 1.2 (2022-01)

C Springer Old and New Research Data Policy

C.1 Legacy Data Policy

Springer Data Policy

All Springer Nature journals are moving to a policy that requires data availability statements for primary research articles. This is already in place for BMC, Nature and SpringerOpen titles and is being progressively adopted by Springer and Palgrave Macmillan journals.

While the implementation is underway, certain Springer and Palgrave Macmillan journals will retain the older data policy types, as outlined below. The specific data policy of each journal is stated in the submission guidance. 'Type 3' is equivalent to our new, standardised data policy. Our legacy data policies

Policy Type Policy summary:
Type 1: Data sharing and data citation is encouraged
Type 2: Data sharing and evidence of data sharing encouraged
Type 3: Data sharing encouraged and statements of data availability required
Type 4: Data sharing, evidence of data sharing and peer review of data required

C.2 New Research Data Policy

https://www.springernature.com/gp/authors/research-data-policy

Research data policy

At Springer Nature we advance discovery by publishing trusted research, supporting the development of new ideas and championing open science. We also aim to facilitate compliance with research funder and institution requirements to share data.

To help accomplish this we have established a standard research data policy for our journals, based on transparency around supporting data. This policy applies to all datasets that are necessary to interpret and replicate the conclusions reported in a research article.

1. **All original articles must include a data availability statement.**

   Data availability statements should include information on what data are available, where these can be found, and any applicable access terms. This applies to both original and reused data, and whether or not data can be shared publicly. See our guidance on data availability statements for more information.

2. **We strongly encourage that all datasets supporting the analysis and conclusions of the paper are made publicly available at the time of publication, and we mandate the sharing of community-endorsed**

data types.

We encourage authors to deposit their supporting data in publicly available repositories, or failing this within the manuscript or additional supporting files. See our repository guidance for more information.

For a number of data types, submission to a community-endorsed, public repository is mandatory. See our list of mandated data types.

3. **Peer reviewers are entitled to request access to underlying data (and code) when needed to perform their evaluation of a manuscript.**

4. **We recognise it is not always possible to share research data publicly, for instance when privacy of research participants could be compromised. In such instances data availability should still be stated in the manuscript along with any conditions for access.**

   A large number of our journals already support this policy, including Nature Portfolio, BMC and many Springer and Palgrave titles. We are in the process of implementing this policy across the remainder of our portfolio in stages.

D References

[Anderson and Dewald, 1994] Anderson, R. G. and Dewald, W. G. (1994). Replication and scientific standards in applied economics a decade after the journal of money, credit and banking project. *Review*, 76.

[Anderson et al., 2008] Anderson, R. G., Greene, W. H., McCullough, B. D., and Vinod, H. D. (2008). The role of data/code archives in the future of economic research. *Journal of Economic Methodology*, 15(1):99–119.

[Ashenfelter et al., 1986] Ashenfelter, O., Haveman, R., Riley, J., and Taylor, J. (1986). Editorial statement. *The American Economic Review*, 76(4):v–v.

[Bach et al., 2023] Bach, L., Bozio, A., Guillouzouic, A., and Malgouyres, C. (2023). Dividend taxes and the allocation of capital: Comment. *American Economic Review*, 113(7):2048–2052.

[Boissel and Matray, 2022] Boissel, C. and Matray, A. (2022). Dividend taxes and the allocation of capital. *American Economic Review*, 112(9):2884–2920.

[Bollen et al., 2015] Bollen, K., Cacioppo, J., Kaplan, R., Krosnick, J., and Olds, J. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the subcommittee on replicability in science, advisory committee to the national science foundation directorate for social, behavioral, and economic sciences. *National Science Foundation*.

[Chang and Li, 2015] Chang, A. C. and Li, P. (2015). Is economics research replicable? sixty published papers from thirteen journals say'usually not'. Technical report, FEDS Working Paper.

[Christensen and Miguel, 2018] Christensen, G. and Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–980.

[Connolly et al., 2023] Connolly, M., Koren, M., Llull, J., Morrow, P., and Vilhuber, L. (2023). A journal's guide to choosing a repository for replication packages.

[Currier et al., 2021] Currier, B., Butler, C., and Lillard, K. (2021). Safeguarding research: A review of economics journals' preservation policies for published code and data files. *Federal Reserve Bank of Kansas City Working Paper*, (21-14).

[Dewald et al., 1986] Dewald, W. G., Thursby, J. G., and Anderson, R. G. (1986). Replication in empirical economics: The journal of money, credit and banking project. *The American Economic Review*, pages 587–603.

[Dreber and Johannesson, 2023] Dreber, A. and Johannesson, M. (2023). A framework for evaluating reproducibility and replicability in economics. *Available at SSRN 4458153*.

[Duvendack et al., 2015] Duvendack, M., Palmer-Jones, R. W., Reed, W. R., et al. (2015). Replications in economics: A progress report. *Econ Journal Watch*, 12(2):164–191.

[Feige, 1975] Feige, E. L. (1975). The consequences of journal editorial policies and a suggestion for revision. *Journal of Political Economy*, 83(6):1291–1296.

[Feigenbaum and Levy, 1993] Feigenbaum, S. and Levy, D. M. (1993). The market for (ir) reproducible econometrics. *Accountability in Research*, 3(1):25–43.

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

[Giuliano and Spilimbergo, 2014] Giuliano, P. and Spilimbergo, A. (2014). Retracted: Growing up in a recession. *Review of Economic Studies*, 81(2):787–817.

[Ham et al., 2021] Ham, J. C., Wright, J., and Ye, Z. (2021). New rankings of economics journals: Documenting and explaining the rise of the new society journals. *Available at SSRN 3606030*.

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

[Koren et al., 2022] Koren, M., Connolly, M., Llull, J., and Vilhuber, L. (2022). Data and code availability standard [version 1.0]. `https://datacodestandard.org`.

[Labor Dynamics Institute, 2022] Labor Dynamics Institute (2022). Should journals verify reproducibility? `https://www.youtube.com/watch?v=-dc4xxCIeqQ`.

[Maniadis et al., 2017] Maniadis, Z., Tufano, F., and List, J. A. (2017). To replicate or not to replicate? exploring reproducibility in economics through the lens of a model and a pilot study.

[Mayer, 1980] Mayer, T. (1980). Economics as a hard science: Realistic goal or wishful thinking? *Economic Inquiry*, 18(2):165.

[McCullough et al., 2006] McCullough, B. D., McGeary, K. A., and Harrison, T. D. (2006). Lessons from the jmcb archive. *Journal of Money, Credit and Banking*, pages 1093–1107.

[McCullough and Vinod, 2003] McCullough, B. D. and Vinod, H. D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review*, 93(3):873–892.

[Rampini et al., 2020] Rampini, A. A., Viswanathan, S., and Vuillemey, G. (2020). Retracted: Risk management in financial institutions.

[Renfro, 2004] Renfro, C. G. (2004). Econometric software: The first fifty years in perspective. *Journal of Economic and Social Measurement*, 29(1-3):9–107.

[REStud, 2023] REStud (2023). Retraction of: Growing up in a Recession. *The Review of Economic Studies*, 90(2):1009–1009.

[Vilhuber, 2020] Vilhuber, L. (2020). Reproducibility and replicability in economics. *Harvard Data Science Review*, 2(4):1–39.

[Vilhuber, 2023] Vilhuber, L. (2023). Reproducibility and transparency versus privacy and confidentiality: Reflections from a data editor. *Journal of Econometrics*.

[Vilhuber et al., 2023] Vilhuber, L., Schmutte, I., Michuda, A., and Connolly, M. (2023). Reinforcing Reproducibility and Replicability: An Introduction. *Harvard Data Science Review*, 5(3). https://hdsr.mitpress.mit.edu/pub/l8dmf3cm.

[Vlaeminck et al., 2015] Vlaeminck, S., Herrmann, L.-K., et al. (2015). Data policies and data archives: A new paradigm for academic publishing in economic sciences? In *ELPUB*, pages 145–155.

[Whited, 2023] Whited, T. (2023). Costs and Benefits of Reproducibility in Finance and Economics. *Harvard Data Science Review*, 5(3). https://hdsr.mitpress.mit.edu/pub/mnhzk8gq.