Report for 2018 by the AEA Data Editor

By Lars Vilhuber*

Your abstract here, please. Keywords: reproducibility; replicability; science of science

The purpose of scientific publishing is the dissemination of robust research findings, exposing them to the scrutiny of peers. Key to this endeavor is documenting the provenance of those findings. For theoretical articles, these are the proofs of theorems and the like that the authors provide. For empirical articles, the foundations on which the findings reside are external to the article, and often to the journal, in which they are published. Many scientists, journals, learned societies, and funding agencies have called for greater transparency of research practices, and more assurance that published research is reproducible (Stodden et al., 2016; Fuentes, 2016; Moffitt, 2016; Camerer et al., 2016*a*; Bollen et al., 2015; Joskow, 2015; Christensen and Miguel, 2018). Our scientific community faces increasingly complex issues of privacy and confidentiality that prevent "open" access to those same sources (Anderson and Seltzer, 2009; Abowd and Schmutte, forthcoming). Large and private databases (often both at the same time) are being used to analyze economic phenomena, with subsequent publications (Baker, Gibbs and Holmstrom, 1994; Lazear, 2000; Bailey et al., 2018; Chen et al., 2017; Hall and Krueger, 2018), yet few such data are available for replication exercises Jeng and Lerner (2016). To ensure the credibility of the scientific endeavor, transparency of the methods and data used are critical. Various studies have shown that too few studies are (easily) reproducible (McCullough, 2007; McCullough, McGeary and Harrison, 2006; Anderson et al., 2008; Anderson and Dewald, 1994). There is a need to properly cite the digital inputs to our published output and to properly curate those inputs.

In January 2018, I was appointed as the first Data Editor of the American Economic Association, with the mission to "design and oversee the AEA journals strategy for archiving and curating research data and promoting reproducible research" (Duflo and Hoynes, 2018). This first report by a Data Editor describes my efforts over the past year to advance that mission. It also highlights some of the short- and medium-term changes that economists might expect when publishing their research.[1]

* Vilhuber: Cornell University, lars.vilhuber@cornell.edu.

[1] A variety of replication concepts have been defined in economics (Hamermesh, 2007; Clemens, 2017). In this article, we adopt the definitions articulated by Bollen et al. (2015), among others. *Reproducibility* refers to "the ability [. . .] to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator," and is related to the "narrow" sense of replication of Pesaran (2003). Use of the "same procedures" may imply using the same computer code or re-implementing the statistical procedures in a different software package. Hamermesh (2007) calls this "pure replication". Christensen and Miguel (2018, p. 942) argue that this is the "basic standard [that] should be expected of

I. The current environment

The American Economic Association (AEA)'s data and code posting policy (American Economic Association, 2008), as well as that of other societies and journals, are a reaction to earlier calls to increase transparency (McCullough, McGeary and Harrison, 2006; Anderson et al., 2008), and are intended to create a minimal framework from which to replicate empirical findings, by requiring the data and code to be available to others. In practice, enough reproduction and replication attempts fail (Camerer et al., 2016*b*; Chang and Li, 2015, 2017), not just in economics (Baker, 2015; Collaboration, 2015) (I will comment on our own efforts later). It remains an open question who should be tasked with conducting a "replication" in the first place - should the editorial team verify reproducibility during the editorial process (Jacoby, Lafferty-Hess and Christian, 2017), should the referees be able to do this, or should they be required to do this? Or should the readers of the articles, and the broader scientific community, attest to the replicability and ultimately the generalizability of the findings (Hamermesh, 2017)? Related is the question whether enough replications are being published (Berry et al., 2017; Burman, Reed and Alm, 2010; Coffman, Niederle and Wilson, 2017; Duvendack, Palmer-Jones and Robert Reed, 2017; Höffler, 2017).

Very few journals have implemented verification of submitted code and data during the editorial process. In political science, the American Journal of Political Science in collaboration with the Odum Institute for Research in Social Science (Christian et al., 2018) has been conducting data curation and code verification. The Journal of the American Statistical Association performs a "broad evaluation of quality and potential for usability of the code and data" since 2016 (Stodden et al., 2016).

No journal currently does an adequate job of providing information about restricted-access data.[2] This is not only the fault of the journals: Most restricted-access data centers do not provide structured information about existence, modalities of access, or even data landing pages for the datasets they provide access to.[3] None of these solutions are widespread, and standards are only now being developed.

all published economics research, and hope this expectation is universal among researchers." *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 (Hamermesh, 2017, : "scientific replication")

[2]Elsevier journals have experimented with "Data Descriptions", but while the form is machine-readable, it is essentially free-form text, and checking the box "confidential data" essentially stops the process of filling in any information.

[3]Restricted-access data hosted on ICPSR and possibly Harvard Dataverse are notable exceptions.

II. The Mission, if You Choose to Accept It

With the mission outlined above in mind, the Data Editor's long-term tasks are

1) Elaborate a data and code availability policy that is modern, responsive, and imposes the lowest burden on authors and readers that is commensurate with the overall goals;
2) Creating technical, human, and organizational infrastructure at the AEA journals to support all aspects of implementing the data and code availability policy;
3) Working with other providers of scientific infrastructure to improve support for documenting provenance and replicability;
4) Working with the economics community to enhance and broaden education on replicable science;
5) Conducting research and participating in experiments in the intersection of publication, replication, and provenance documentation

In particular, a revised data and code posting policy should maximize credibility and trustworthiness of research findings, and address the following goals:

1) to encourage and reward incorporating basic principles of replicability into researchers' workflow;
2) to prioritize linking to existing data and code repositories, as the primary mechanism of providing source materials, with a journal-sanctioned repository as a fall-back archive;
3) to require and facilitate proper documentation of restricted-access data;
4) to enforce a limited measure of verification;
5) balance the previous goals with the need to *reduce* the burden on authors, not increase it.

III. Implementing improved transparency of research

In the first year, we have moved a few tasks forward.

A modern data and code availability policy should support both reproducibility and replicability, by supporting accurate and transparent description of the provenance of the scientific results. In particular, a functional implementation of those concepts suggests that both data and code need to be subject to the Findable, Accessible, Interoperable, Re-usable (FAIR) principles FORCE11 (2016): findable, accessible, interoperable, and re-usable. In this context, we interpret the "interoperability" of code as "code that works, and the workings of which are comprehensible by a third party" (CITE??).

A. Goal 1: Improved findability of data used in research articles

Under this goal, we start by abolishing journal-specific "supplementary materials" as the primary repository of data and code that are part of the provenance chain of an article. As currently implemented at most journals, including the AEA's journals, they lack findability, proper citability as first-class objects, and are somewhat opaque (packaged as ZIP files). Historical materials will be migrated to a new curated archive at ICPSR. The AEA's "Data and Code Archive @ ICPSR" will display the full contents of the materials as deposited by authors in the past, without the need to download ZIP files. The materials will receive their own citable Digital Object Identifier (DOI). Through the DOI registrars, we automatically leverage the ability to link and associate the archives with their original articles.

On the AEA's journal websites, the links to "supplementary" materials will initially appear to be the same (although pointing to the new locations), but future enhancements will allow for greater visibility or transparency of the associated materials. However, by separating the hosting of data archives (for historical materials, at ICPSR) from the referencing of those archives, we open the door to a more consistent model of linking to and citing data artifacts associated with published articles.

For future submissions, we will allow researchers to reference supplementary materials on a wide list of data archives or repositories.[4] In particular, while some journals are already curating a list of recommended *open* data archives (Nature Scientific Data, 2016) (ALSO PLOS, F1000), we will also allow authors to reference materials in reliable *restricted-access* data repositories. What constitutes a "reliable" data repository? For one, it needs *persistence*, and *accessibility*. Properly managed repositories have a *preservation policy* - they commit to maintaining deposits for a defined duration (often, but not always, in perpetuity), and only under very restrictive circumstances will remove deposits. Such repositories will also have a policy about access - who can obtain data deposited at the institution, and under what conditions. This characterization applies homogeneously to open and restricted-access repositories. A recommended repository will have been vetted (by the AEA Data Editor, or a reliable third party) to have acceptable and credible policies.

Authors submitting their work to the AEA journals will be affected in several ways. First, those authors who already deposit their (open access) code and data at known repositories will not have to do so again - a simple reference (and citation!) of the previously archived materials is sufficient. Authors who use data provided through institutional providers (Panel Study of Income Dynamics (PSID), Health and Retirement Study (HRS), the U.S. Census Bureau, and international equivalents), and who in some cases cannot deposit the data, will also reference the persistent location where they obtained their data from, and

[4] Do I need to distinguish the two? is there a difference?

where others can do so as well. In the case of restricted-access, a better description of access procedures will be requested from authors, who in turn should ask their data providers to provide such procedural descriptions, in the form of web pages and (persistent and citable) documents.

If replicability is truly part of the research scientist's workflow, then by the time she submits an article to any journal, the intermediate and final data products as well as the code used for an article have already been deposited at appropriate repositories and archives. If all such repositories and archives are of sufficient quality, then the additional deposit at a journal is duplicative at best, and perturbative to the provenance chain at worst. The right solution is to reference those other repositories, not copy them. Of course, for those that have not used repositories and simply wish to provide a replication zip archive of the files on their laptop, an adequate deposit solution should also exist. By fundamentally relying on references to repositories instead of deposits, it also becomes possible to put public-use and restricted-access data on a comparable footing at the journal with regards to potential replicability. Well documented location of data (through DOI), access protocols (implicit or click-through license, contracts, etc.), and access mechanism (direct download, delivery of physical media or controlled download, sign-on to controlled secure access, etc.) are then available for any data.

IV. Goal 2: Improved Reliability of Replication Materials

?

V. Data Citations

Properly referencing data goes beyond just reproducibility - it is also proper scientific writing style. In the same way that we use bibliographic references to "printed" resources, we should also be using such references for data resources, to give and receive credit where credit is due. Not referencing an article or book is at best an oversight, and at worst plagiarism - and the same should apply to data objects. Numerous guides and tutorials exist (DataONE, 2011; ICPSR, 2018; Martone, 2014).

The AEA uses the Chicago style for citations and bibliographies (American Economic Association, 2018). However, the Chicago Style Manual (Chegg, 2018; Chicago Manual of Style Online, 2018) does not provide examples for data citations, and neither does the Citation Style Language[5] used by applications like Zotero[6] and Mendeley Desktop[7].

As part of our activities, the AEA prepress department has started the process

[5] https://citationstyles.org/
[6] https://www.zotero.org/
[7] https://www.mendeley.com/download-desktop/

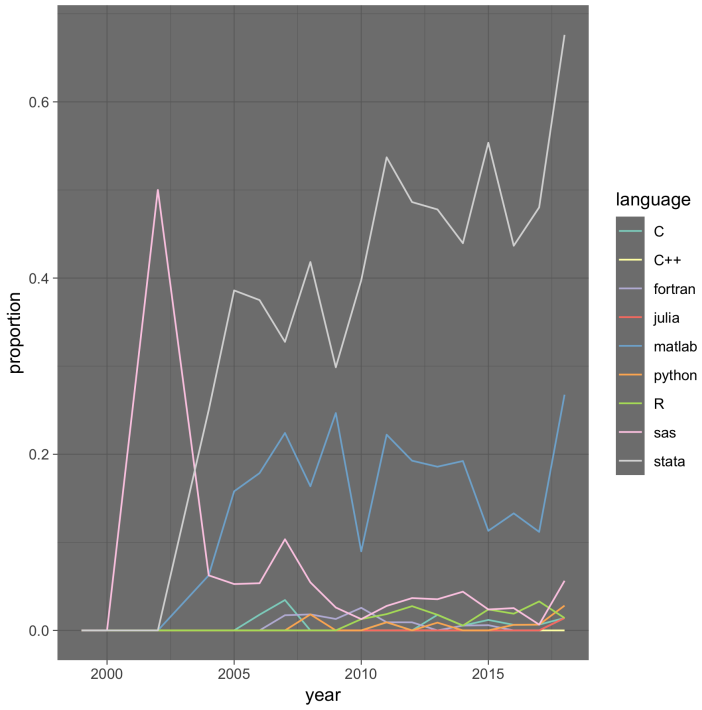

FIGURE 1. POPULARITY OF STATISTICAL SOFTWARE IN THE AER

Figure provided by Patrick Baylis (UBC), based on filename extensions in ZIP files of replication materials on the AEA website.

of updating AEA templates available through such software.[8] Some guidance for data citations is provided at

VI. Future activities

Keeping an eye on novel techniques **??** REgistered reports: **??**

REFERENCES

Abowd, John M., and Ian M. Schmutte. forthcoming. "An Economic Analysis of Privacy Protection and Statistical Accuracy as Social Choices." *American Economic Review*.

[8] For the technically inclined, this process involves updating an existing style or creating a new style on https://citationstyles.org/ and https://github.com/citation-style-language/styles, from where it propagates to a large number of software packages.

American Economic Association. 2008. "Data Availability Policy." *https://www.aeaweb.org/journals/policies/data-availability-policy*, (accessed: 2017-04-06).

American Economic Association. 2018. "Sample References." *https://www.aeaweb.org/journals/policies/sample-references*, (accessed on 2018-08-10).

Anderson, Margo J., and William Seltzer. 2009. "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues." *Journal of Privacy and Confidentiality*, 1(1).

Anderson, Richard G., and William G. Dewald. 1994. "Replication and Scientific Standards in Applied Economics A Decade After the Journal of Money, Credit and Banking Project." *Federal Reserve Bank of St. Louis Review*, 76(6).

Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2008. "The Role of Data/Code Archives in the Future of Economic Research." *Journal of Economic Methodology*, 15(1): 99–119.

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives*, 32(3): 259–280.

Baker, George, Michael Gibbs, and Bengt Holmstrom. 1994. "The Wage Policy of a Firm." *The Quarterly Journal of Economics*, 109(4): 921–955.

Baker, Monya. 2015. "Over Half of Psychology Studies Fail Reproducibility Test." *Nature*.

Berry, James, Lucas C. Coffman, Douglas Hanley, Rania Gihleb, and Alistair J. Wilson. 2017. "Assessing the Rate of Replication in Economics." *American Economic Review*, 107(5): 27–31.

Bollen, Kenneth, John T Cacioppo, Robert M Kaplan, Jon A Korsnick, and James L Olds. 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." 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." *Public Finance Review*, 38(6): 787–793.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016*a*. "Evaluating replicability of laboratory experiments in economics." *Science*, 351(6280): 1433–1436.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016*b*. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, aaf0918.

Chang, Andrew C, and Phillip Li. 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"." Board of Governors of the Federal Reserve System (U.S.).

Chang, Andrew C., and Phillip Li. 2017. "A Preanalysis Plan to Replicate Sixty Economics Research Papers That Worked Half of the Time." *American Economic Review*, 107(5): 60–64.

Chegg. 2018. "Citation Machine: Chicago Manual Of Style 17th Edition (Author Date) format citation generator for journal article." *http://www.citationmachine.net/chicago-17-author-date*, (accessed on 2018-10-02).

Chen, M. Keith, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen. 2017. "The Value of Flexible Work: Evidence from Uber Drivers." National Bureau of Economic Research Working Paper 23296.

Chicago Manual of Style Online. 2018. "Author-Date: Sample Citations." *https://www.chicagomanualofstyle.org/tools_citationguide/citation-guide-2.html*, (accessed on 2018-10-02).

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

Christian, Thu-Mai, Sophia Lafferty-Hess, William Jacoby, and Thomas Carsey. 2018. "Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals."

Clemens, Michael A. 2017. "The Meaning of Failed Replications: A Review and Proposal." *Journal of Economic Surveys*, 31(1): 326–342.

Coffman, Lucas, Muriel Niederle, and Alistair J Wilson. 2017. "Replications: A Proposal to Increase their Visibility and Promote them."

Collaboration, Open Science. 2015. "Estimating the Reproducibility of Psychological Science." *Science*, 349(6251): aac4716–aac4716.

DataONE. 2011. "DataONE Tutorial on Data Citation." *http://www.dataone.org/sites/all/documents/L09_DataCitation.pptx*, (accessed on 2018-08-10).

Duflo, Esther, and Hilary Hoynes. 2018. "Report of the Search Committee to Appoint a Data Editor for the AEA." *AEA Papers and Proceedings*, 108: 745.

Duvendack, Maren, Richard Palmer-Jones, and W Robert Reed. 2017. "What is Meant by 'Replication' and Why Does It Encounter Resistance in Economics?"

FORCE11. 2016. "THE FAIR DATA PRINCIPLES."

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

Hall, Jonathan V., and Alan B. Krueger. 2018. "An Analysis of the Labor Market for Uber's Driver-Partners in the United States." *ILR Review*, 71(3): 705–732.

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

Hamermesh, Daniel S. 2017. "Replication in Labor Economics: Evidence from Data and What It Suggests." *American Economic Review*, 107(5): 37–40.

Höffler, Jan H. 2017. "ReplicationWiki: Improving Transparency in Social Sciences Research." *D-Lib Magazine*, 23(3/4).

ICPSR. 2018. "Data Citations." *https://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp*, accessed on 2018-08-10.

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

Jeng, Leslie, and Josh Lerner. 2016. "Making Private Data Accessible in an Opaque Industry: The Experience of the Private Capital Research Institute." *American Economic Review*, 106(5): 157–160.

Joskow, Paul L. 2015. "President's Letter, Alfred P. Sloan Foundation Annual Report 2014." Alfred P. Sloan Foundation.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346–1361.

Martone, M. (ed.). 2014. "Data Citation Synthesis Group: Joint Declaration of Data Citation Principles."

McCullough, B D. 2007. "Got Replicability? The Journal of Money, Credit and Banking Archive." *Econ journal watch*, 4(3): 326–337.

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

Moffitt, Robert. 2016. "Report: American Economic Association Committee on Statistics (AEAStat)." *American Economic Review*, 106(5): 788–793.

Nature Scientific Data. 2016. "Nature Scientific Data recommended repositories." *figshare*.

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

Stodden, Victoria, and Isabel Reich. 2012. "Software Patents as a Barrier to Scientific Transparency: An Unexpected Consequence of Bayh-Dole." *7th Annual Conference on Empirical Legal Studies*.

Stodden, Victoria, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John P A Ioannidis, and Michela Taufer. 2016. "Enhancing reproducibility for computational methods." *Science*, 354(6317): 1240–1241.

Reviewers