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.

II. The Mission, if You Choose to Accept It

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

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.

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, though the visible impact on the Association's journals or practices is yet to come. In this section, I lay out both vision, actions, and plans for various goals and tasks.

A. Task 1: Improved data and code availability policy

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." Furthermore, this should be true for *all* data and code, not just code that is open-source and data that is open-access.

Replacing ZIP files. — Under this goal, we are working to abolish, or reduce the recourse to, journal-specific "supplementary materials" as the primary repository of data and code. As currently implemented at most journals, including the AEA's journals, "supplementary materials" are packaged as ZIP files and attached to a web page. Provided in this way, they lack findability, proper citability as first-class objects, and are somewhat opaque. To replace such supplementary materials, we are laying the groundwork in three ways.

Creating the AEA Data and Code Archive. — First, we will replace the ZIP files with deposit at a proper data archive, the "AEA Data and Code Archive." The archive will display the full contents of the materials as deposited by authors in the past, without the need to download ZIP files. The archive will be searchable, by JEL codes, keywords, and other characteristics of the data. 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, but also any other articles that use and cite the data, such as replication articles. Several other economics journals (Quarterly Journal of Economics, the Review of Economics and Statistics) already have similar archives. **We are currently finalizing contract terms and addressing technical improvements with such a data archive, and expect a decision to be made in time for the Meetings, and the repository to be available for new deposits in 2019Q2.** All historical data supplements will be migrated to the "AEA Data and Code Archive" as well, with the **migration expected to be completed by 2019Q3**.

Allowing third-party repositories. — However, 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, albeit in intermediate form. 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.[4] Such a policy is already standard practice in some other domains and publishing platforms (some Springer Nature[8] journals, F1000 Research[9], geosciences).

We will therefore allow authors to keep their supporting materials (data, code) at third-party repositories, as long as those repositories satisfy certain criteria

[4] I note here that some materials associated with the Association's journals have sometimes been deposited at sites such as personal websites, Github.com[5], Google Drive[6], Dropbox[7], and others. These sites are not *considered* data archives, for two key reasons. For one, however unlikely it may seem, these commercial companies are ephemerous, and do not have data preservation as a primary mission. More importantly, users who keep data on these sites can delete the data at any time, for any reason, possibly simply because they did not pay their monthly or annual fee. Such practices are incompatible with proper data curation standards.

[8] https://www.springernature.com/gp/authors/research-data-policy/data-policy-types/

[9] https://f1000research.com/for-authors/data-guidelines

in terms of accessibility and preservation. In particular, this policy will allow us to treat public-use data available through repositories such as ICPSR[10] symmetrically with restricted access data curated by survey institutes such as the confidential geodata at the Panel Study of Income Dynamics (PSID), national statistical offices such as the U.S. Census Bureau or Statistics Norway, private-sector research institutes such as the Private Capital Research Institute[11], or private companies such as Twitter[12] or Uber[13], as long as objective criteria in terms of *data curation* and *accessibility* are met. If authors do not have access to such an archive, or have not adopted replicable workflows that incorporate data curation, then they will continue to deposit their data and code in the "AEA Data and Code Archive." This new policy will be implemented in parallel with the deployment of the "AEA Data and Code Archive."

Transparent method to describe all data and code. — Finally, to support to third-party repositories, we are developing a *data description standard* (in the terms of the trade, a "metadata schema") to accurately and completely describe all materials related to a published article, including their access and preservation policies. The data description standard will be accompanied by tools to support this schema for authors and journal websites.[14] The goal of the schema and tools is to move away from idiosyncratic ways to describe and name the data, where to find data, and how to access the data. The ultimate result of implementing the schema will be to provide a more consistent way to display such information within articles or on journal web pages, provide such information in a machine-readable way, and to efficiently leverage the existing metadata that is already provided by numerous data archives. Furthermore, this schema will be implemented as an *information package* that data archives of all kinds can provide to users of their data, and which researchers can share and re-use. The information package is explicitly designed to handle both public-use data archives with well-defined DOI, access policies (simple download!), and preservation policies, as well as repositories of restricted-access data without persistent (public) identifiers, complex access policies with myriad rules, and hidden preservation policies. If the data provider does not provide the information package, the author will be able to fill out such a package on her own. I expect this to be a net reduction in time and effort for authors who already describe their data, and to be an increase in work only for those that do not provide complete descriptions of data provenance. In particular authors who use restricted-access data should see a large reduction in the time needed to describe their data access.

[10] https://icpsr.org

[11] http://www.privatecapitalresearchinstitute.org/

[12] https://twitter.com

[13] https://uber.com

[14] A paper describing the schema (Vilhuber and Lagoze, 2019) is scheduled to be presented at the International Digital Curation Conference (IDCC) in February 2019 and is currently being reviewed at the International Journal of Digital Curation.

The schema and tools are currently being developed and peer-reviewed, and are expected to be tested internally in 2019. Conditional on satisfactory performance both for journal managers and users, we plan to deploy it more widely in 2020.

I emphasize the term "standard" used above - our goal is not simply to have a schema of use for the Association's journals, but rather, to allow authors to use the data description we are developing at any journal, not just those published by the Association. We have already reached out to some journals and publishing platforms (see below), and will continue to do so in the next year.

Timing. — We expect to ask for the *information package* quite early in the submission process, possibly at the same time as manuscript submission itself. In the case of third-party archives, we also expect that authors will have deposited their materials before submission, but no later than the time they receive a revise-and-resubmit. All information will be verified prior to acceptance, rather than prior to publication, as is the case right now. However, we do not expect of authors that the data or code itself be made public prior to publication of the article.

Visible changes. — 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. 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 (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 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. Once the data description standard mentioned above is available, the information provided by data providers will be captured in that form, and through a web-based tool.

B. *Task 2: Creating infrastructure at the AEA journals*

Pre-publication verification of reproducibility. — One of the key findings of the recent literature has been that code provided by authors, even when deposited at journal as supplementary materials, may not always yield the advertised outcomes (Chang and Li, 2015, 2017; Höffler, 2017). To that end, we will reinforce the data and code availability policies above by conducting pre-publication verification of code reproducibility. To our knowledge, this has been systematically done only at the American Journal of Political Science (Christian et al., 2018), though the AJPS also does significant data curation tasks for the authors.

The sizeable challenge is how to conduct such pre-publication verification without unduly delaying the editorial workflow. Christian et al. (2018) report lengthy delays between acceptance and publication due to the combination of data curation and code verification. Several factors lead us to think that this is feasible.

For the past 4 years, I have been running a Replication Lab at Cornell with mostly undergraduates. As part of a summer activity (since Fall 2018 also during classes), students download articles and their supplemental data packages from the AEA website, and attempt to reproduce them ("post-publication verification"). The initial focus was on a complete census of the AEJ:AE, but since my appointment in January 2018, the focus has been broadened to AEJ:Macro, AEJ:Micro, and to some extent the other AEA journals. As of December 2018, about 900 articles have been assessed, and about 430 reproduction attempts have been conducted (a detailed report is being prepared as these lines are being written, and will be available by late December 2018).

Table 1—Software usage by journal

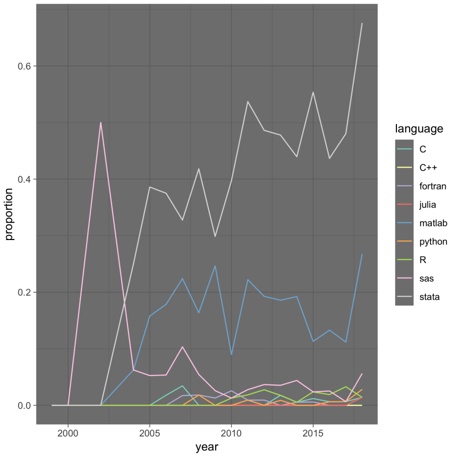
journal	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
American Economic Journal: Applied Economics	34	65	83	154	34	56	24	36	43	20	549
American Economic Journal: Macroeconomics				30	29	35	37	29	34		194
American Economic Journal: Microeconomics						10	10	43	38		101
American Economic Review									67		67

Note: Assessments conducted by Cornell Replication Lab. Some articles were assessed multiple times. As of December 2018. Preliminary numbers.

While we have not attempted to measure exact hourly productivity measures, we have some approximate numbers. Over a certain time period in the Summer of 2018, 10 students worked 774 hours. During that time, they assessed 147 articles, attempted 108 replications, and succeeded 49 of those. Thus, 5 students working each a 20-hour workweek can process 20 articles per week, and about 13 reproductions attempted. Note however that these numbers are not representative of all journals. All AEJ:AE articles have an empirical component, in contrast to most other journals. On the other hand, there may be systematic differences in the complexity of the analyses reported in the AEJ:AE relative to other journals. Nevertheless, we believe that an appropriately trained student lab can handle many if not most of the required reproduction attempts.

Second, the variety of software is relatively small. Figure 1 shows the distribution of software in on the AEA website. Table 2 identifies software 'mentions' by journal as assessed by the Cornell Replication Lab. An article might mention (or use) multiple software programs in their analysis package. Economists in general make heavy use of Stata, with a significant fraction using Matlab. In general, the

Cornell Replication Lab had to assign a graduate student (or faculty member) for non-standard software (Gauss, Fortran, etc.), but only very rarely.



AER

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

Figure 1. Popularity of statistical software

Table 2—Software usage by journal

Journal	Stata	Matlab	R	SAS
American Economic Journal: Applied Economics	86.46	3.31	1.66	2.49
American Economic Journal: Macroeconomics	36.21	40.33	1.23	3.29
American Economic Journal: Microeconomics	10.78	7.84	1.96	0
American Economic Review	50	20.51	1.28	1.28

Note: In percent of mentions for each journal, as assessed by the Cornell Replication Lab. An article might mention multiple software programs. Additional software not listed here includes Fortran, Mathematica, SPSS, Eviews, Excel. Preliminary numbers.

Third, we are explicitly not attempting to address data curation, though we will do simple checks for data documentation (can a reproducer understand what the file is for). Furthermore, by allowing for third-party archives, we expect that the authors' home institutions or third parties may increasingly offer assistance with this, regardless of where the author will be submitting the article.

C. Task 3: Working with other providers of scientific infrastructure to improve support for documenting provenance and replicability

We are engaging with other providers - both within the economics community, and elsewhere.

Restricted-access data and verification. — Naturally, given the preponderance of restricted-access data used in economics research, the verification of code that uses such data is challenging. For now, such code will not be verifiable. However, we are aware of efforts in various areas looking to provide verification services when data are restricted-access, either by third-parties with access rights, or the data providers themselves as part of their disclosure-avoidance procedures. We are actively engaging in and supporting such efforts. We have also explored in specific cases if members of our Replication Lab might obtain short-term access to such data, so far without success.

Journals. — We have had discussions with members of the board of the Review of Economic Studies. The board has recently appointed a Data Editor there as well, and we look forward to collaborations. We have also discussed with editors of the Journal of Labor Economics, the Industrial and Labor Relations Review, the Journal of Human Resources, and Sociological Sciences. At conferences such as those of the Research Data Alliance (RDA) and the FORCE11 group, we have participated in workshops, forums, and one-on-one discussions with publishers (Springer Nature) and data curators, and expect to be more formally engaged in working groups in the future.

D. Task 4: Working with the economics community to enhance and broaden education on replicable science

Training. — Implicit in setting up a scalable architecture for pre-publication verification using students is the ability to train those students. Explicit in making reproducibility part of the publication process is that researchers are trained early on to incorporate reproducibility into their workflow. Reproducibility is not absent from campus training, and there are many groups providing such training more broadly. Efforts worth mentioning include the Open Science Framework[15], Project TIER[16], BITSS[17], CodeOcean.com[18] and (Software/Data Carpentry). Many guides and tutorials by smaller groups are also available, including Gentzkow and Shapiro (2014); Wilson et al. (2016); Vilhuber (2018). Future

[15] https://osf.io
[16] http://www.projecttier.org/
[17] http://www.bitss.org/
[18] https://codeocean.com

activities may include encouraging incorporation of such training into curricula at undergraduate and graduate training levels.

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). However, few data citations actually occur in AEA journals.

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[19] used by applications like Zotero[20] and Mendeley Desktop[21].[22] We believe that data citations will be much more prevalent if (a) data providers consistently provide them in machine-readable formats that software can understand and (b) software can consistently provide it in the format required by the journal.

Together with the AEA editorial office, we have started the process of updating AEA templates available through such software.[23] We have also made minor adjustments to the data citation template (American Economic Association, 2018), to conform to newer guidelines, for instance for representing DOI.[24]

Furthermore, for two journals, we will start monitoring data citations during the refereeing process. The *Replication Lab* will do a short assessment of data used and created in articles, verify whether data citations are present, and provide a report to editors, who in turn can provide the report to authors as part of the usual editorial feedback. Authors should expect to see these reports in 2019Q1.

E. *Possible issues*

During our consultations with editors, authors, and other participants, we encountered a few possible issues worth mentioning.

[19] https://citationstyles.org/

[20] https://www.zotero.org/

[21] https://www.mendeley.com/download-desktop/

[22] The object type 'dataset' is defined within CSL, but not implemented in software. Zotero is expected to support an item type 'dataset' as of version 5.1. In December 2018, Zotero is at version 5.0.58.

[23] 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.

[24] I have also produced a draft update to the Bibtex style distributed by the AEA (`aea.bst`), see https://github.com/AEADataEditor/aea-de-guidance/blob/master/citations/aea-mod.bst to download.

Inability to provide information. — Some authors might not be able to provide information on their data provider's preservation policies, or might not be able to provide reproducible information on data and code location when access is restricted, for instance within corporate computer networks. We have consistently iterated that the policy enforcement remains in the hands of the editor responsible for the paper. The policies above are meant to provide more information than was available in the past, including about legitimate cases where no information is available.

Delays. — Many editors were worried that increased verification procedures might delay publication. We have attempted to address this concern by providing a flexible 'trigger' for the verification procedures, by collecting information on how fast assessments and verifications can be handled, and how to scale such tasks. The most important mechanism to avoid delays in the editorial and publication workflow is to have the information about data and code as early as possible, and yet to be able to adjust the time-intensive processes to fall into periods where long delays are normal, such as during a revise-and-resubmit stage. We remain attentive to such concerns, and monitor the situation.

Intellectual Property. — During the negotiations for the hosting of the "AEA Data and Code Archive," we identified several wrinkles related to intellectual property. When migrating toe the new Archive, an explicit license for re-use of the materials must be chosen. As many other journals, the AEA journals acquire copyright to the article and its associated materials, as stated on its website. However, the AEA has never provided an explicit license to the supplementary materials, such as a Creative Commons license. In theory, this means that any extended use of the materials, such as for replication purposes or re-use, is possibly an infringement of copyright. Clearly, this is not the intent of the AEA, and we are working with counsel to clarify this situation for the historical materials.

Going forward, new deposits in the Archive will not transfer copyright - the original authors will retain copyright, but choose an open license for re-use. What exactly those choices are is still being considered. Many archives have chosen the Creative Commons licenses Creative Commons (2017), varying in degree of commercial and derivative re-use allowed. However, Creative Commons specifically suggests that software - i.e., computer code - not be licensed under their licenses, and instead use open source licenses Open Source Initiative (2018). Additional guidance can be found in Stodden and Reich (2012), who suggest CC-BY for databases and modified BSD licenses for software. Mixed archives of data and code, as those traditionally provided by economists, thus will need mixed licenses, and we have been working with the hosting institution for the Archive to allow for such mixed licenses.

Finally, in allowing third-party repositories, the Data Editor will need to verify that the license provided for code and data is sufficiently open to allow for replication - a task facilitated by the *information package* described above.

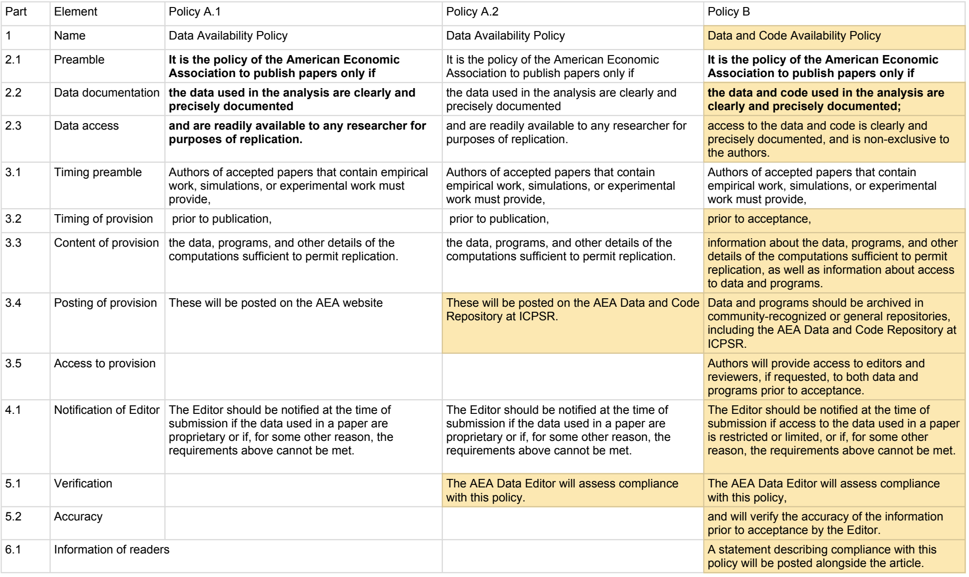
F. Future activities

Issuing a new data and code availability policy. — Once the various activities above coalesce into solid timelines, authors, editors, and referees must be made aware of new policies. In particular, while new deposit procedures to the "AEA Data and Code Archive" will not materially impact the research workflow of authors, allowing for third-party repositories might impact authors, and pre-publication verification of code reproducibility will most definitely impact the way authors prepare for submission of articles. We do not wish to unduly delay submission or publication, and will therefore roll out new submission requirements (new data and code availability policies) with sufficient advance notice so authors can incorporate these into their workflow. In crafting, we have sought guidance by discussion in the data publication community (Hrynaszkiewicz et al., 2017, e.g.). We foresee at least two variants of a data and code availability policies (Figure 2). Policy A.1 is the current policy, widely adopted by journals of the AEA as well as others. When introducing the "AEA Data and Code Archive," the policy will need to be adapted, without fundamentally changing its intent or enforcement (Policy A.2). In particular, provision of supplemental data will be expected post-acceptance. Implementing Policy A.2, however, will involve changes to the workflow for authors, editors, and staff, and will be subject to advance notice. Each policy comes with its set of instructions (Figure **??**). The policies can also be compared at this website, and comments can be sent to `mailto:dataeditor@aeapubs.org`. We note that the data and code availability policy itself will be curated and versioned, by publication in either the AER or the PP. Articles will have a note indicating under what policy the submission occurred.

RCT. — We have spent little time this year considering the improved inclusion of randomized control trial (RCT). The *information package* outlined above will handle better referencing of registrations, e.g. in the AEA RCT Registry[25]. However, we will also engage with MIT to consider how the AEA RCT Registry can potentially be improved.

Other activities. — Furthermore, as pointed out above, we continue to engage with the research data community, while keeping abreast of novel techniques (Brinckman et al., 2018, Codeocean.com), developments in the pre-publication phase (Butler and Kulp, 2018) and the new publication methods such as registered reports (Chris Chambers, 2014; Nosek and Lakens, 2014).

[25] `https://www.socialscienceregistry.org/`



Part	Element	Policy A.1	Policy A.2	Policy B
1	Name	Data Availability Policy	Data Availability Policy	Data and Code Availability Policy
2.1	Preamble	**It is the policy of the American Economic Association to publish papers only if**	It is the policy of the American Economic Association to publish papers only if	**It is the policy of the American Economic Association to publish papers only if**
2.2	Data documentation	**the data used in the analysis are clearly and precisely documented**	the data used in the analysis are clearly and precisely documented	**the data and code used in the analysis are clearly and precisely documented;**
2.3	Data access	**and are readily available to any researcher for purposes of replication.**	and are readily available to any researcher for purposes of replication.	access to the data and code is clearly and precisely documented, and is non-exclusive to the authors.
3.1	Timing preamble	Authors of accepted papers that contain empirical work, simulations, or experimental work must provide,	Authors of accepted papers that contain empirical work, simulations, or experimental work must provide,	Authors of accepted papers that contain empirical work, simulations, or experimental work must provide,
3.2	Timing of provision	prior to publication,	prior to publication,	prior to acceptance,
3.3	Content of provision	the data, programs, and other details of the computations sufficient to permit replication.	the data, programs, and other details of the computations sufficient to permit replication.	information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.
3.4	Posting of provision	These will be posted on the AEA website	These will be posted on the AEA Data and Code Repository at ICPSR.	Data and programs should be archived in community-recognized or general repositories, including the AEA Data and Code Repository at ICPSR.
3.5	Access to provision			Authors will provide access to editors and reviewers, if requested, to both data and programs prior to acceptance.
4.1	Notification of Editor	The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.	The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.	The Editor should be notified at the time of submission if access to the data used in a paper is restricted or limited, or if, for some other reason, the requirements above cannot be met.
5.1	Verification		The AEA Data Editor will assess compliance with this policy.	The AEA Data Editor will assess compliance with this policy,
5.2	Accuracy			and will verify the accuracy of the information prior to acceptance by the Editor.
6.1	Information of readers			A statement describing compliance with this policy will be posted alongside the article.

Figure 2. Proposed data and code availability policies

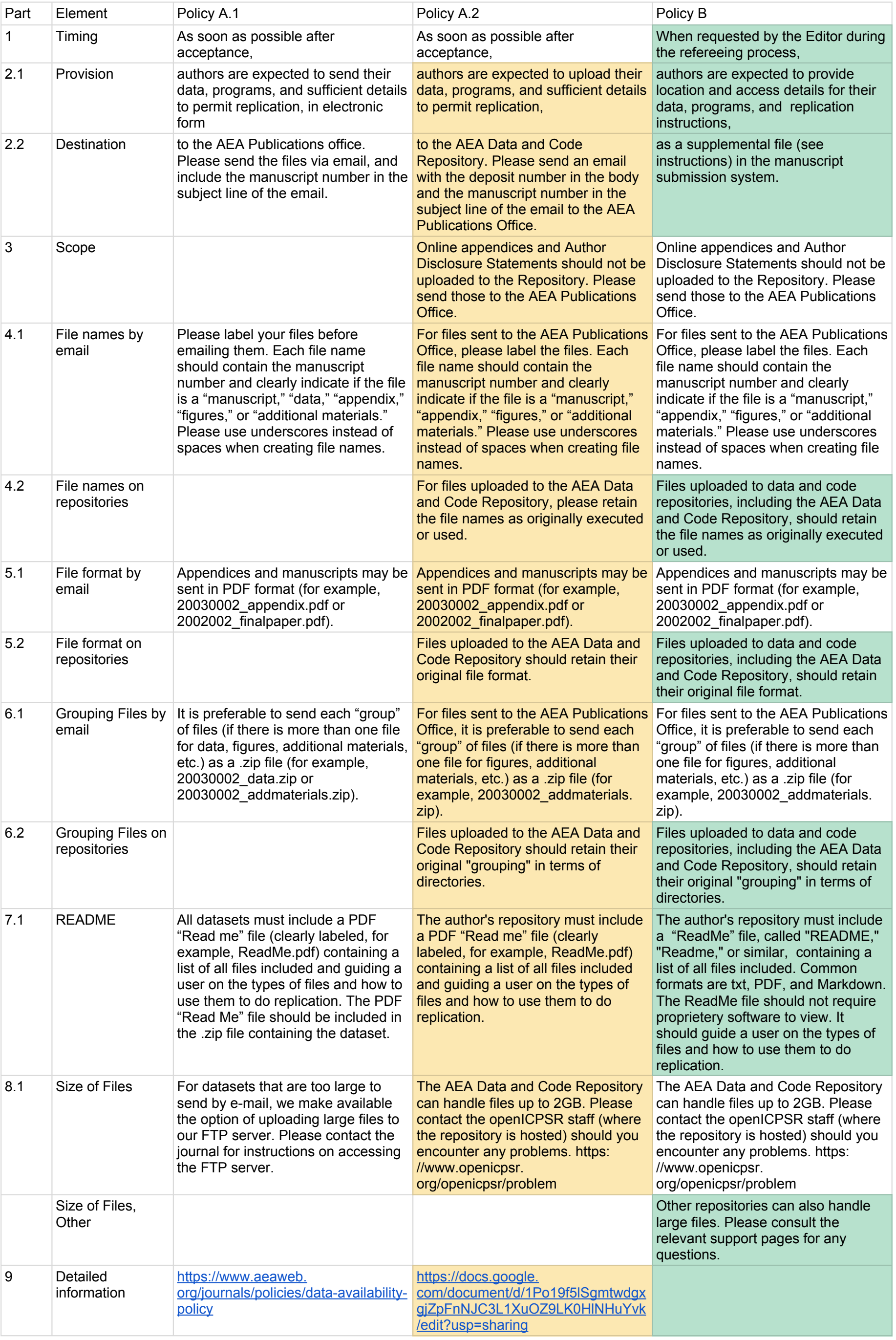
Note: Also available at https://docs.google.com/spreadsheets/d/1khrXxnmKC7Llj9vH17r1KEkN0hcTVqZnQloretc0wDA/edit?usp=sharing

IV. Some Concluding Remarks

If I were to choose a single mission statement sentence for the Data Editor, it would be

> The AEA Data Editor will be a leader in the efforts to provide the infrastructure to support assurances of replicability, at the AEA journals and in the broader economics and social science community.

However, I also believe that in an ideal future world, the job of a Data Editor has disappeared, the same way we do not have a "Reference List Editor" today. Researchers, anticipating that others will verify their empirical findings and publish the results of that verification, will create perfectly replicable archives prior to submission to journals. The current position will be seen, in retrospect, as a temporary bandaid. I have no timeline when that situation will come to pass.



Part	Element	Policy A.1	Policy A.2	Policy B
1	Timing	As soon as possible after acceptance,	As soon as possible after acceptance,	When requested by the Editor during the refereeing process,
2.1	Provision	authors are expected to send their data, programs, and sufficient details to permit replication, in electronic form	authors are expected to upload their data, programs, and sufficient details to permit replication,	authors are expected to provide location and access details for their data, programs, and replication instructions,
2.2	Destination	to the AEA Publications office. Please send the files via email, and include the manuscript number in the subject line of the email.	to the AEA Data and Code Repository. Please send an email with the deposit number in the body and the manuscript number in the subject line of the email to the AEA Publications Office.	as a supplemental file (see instructions) in the manuscript submission system.
3	Scope		Online appendices and Author Disclosure Statements should not be uploaded to the Repository. Please send those to the AEA Publications Office.	Online appendices and Author Disclosure Statements should not be uploaded to the Repository. Please send those to the AEA Publications Office.
4.1	File names by email	Please label your files before emailing them. Each file name should contain the manuscript number and clearly indicate if the file is a “manuscript,” “data,” “appendix,” “figures,” or “additional materials.” Please use underscores instead of spaces when creating file names.	For files sent to the AEA Publications Office, please label the files. Each file name should contain the manuscript number and clearly indicate if the file is a “manuscript,” “appendix,” “figures,” or “additional materials.” Please use underscores instead of spaces when creating file names.	For files sent to the AEA Publications Office, please label the files. Each file name should contain the manuscript number and clearly indicate if the file is a “manuscript,” “appendix,” “figures,” or “additional materials.” Please use underscores instead of spaces when creating file names.
4.2	File names on repositories		For files uploaded to the AEA Data and Code Repository, please retain the file names as originally executed or used.	Files uploaded to data and code repositories, including the AEA Data and Code Repository, should retain the file names as originally executed or used.
5.1	File format by email	Appendices and manuscripts may be sent in PDF format (for example, 20030002_appendix.pdf or 2002002_finalpaper.pdf).	Appendices and manuscripts may be sent in PDF format (for example, 20030002_appendix.pdf or 2002002_finalpaper.pdf).	Appendices and manuscripts may be sent in PDF format (for example, 20030002_appendix.pdf or 2002002_finalpaper.pdf).
5.2	File format on repositories		Files uploaded to the AEA Data and Code Repository should retain their original file format.	Files uploaded to data and code repositories, including the AEA Data and Code Repository, should retain their original file format.
6.1	Grouping Files by email	It is preferable to send each “group” of files (if there is more than one file for data, figures, additional materials, etc.) as a .zip file (for example, 20030002_data.zip or 20030002_addmaterials.zip).	For files sent to the AEA Publications Office, it is preferable to send each “group” of files (if there is more than one file for figures, additional materials, etc.) as a .zip file (for example, 20030002_addmaterials.zip).	For files sent to the AEA Publications Office, it is preferable to send each “group” of files (if there is more than one file for figures, additional materials, etc.) as a .zip file (for example, 20030002_addmaterials.zip).
6.2	Grouping Files on repositories		Files uploaded to the AEA Data and Code Repository should retain their original "grouping" in terms of directories.	Files uploaded to data and code repositories, including the AEA Data and Code Repository, should retain their original "grouping" in terms of directories.
7.1	README	All datasets must include a PDF “Read me” file (clearly labeled, for example, ReadMe.pdf) containing a list of all files included and guiding a user on the types of files and how to use them to do replication. The PDF “Read Me” file should be included in the .zip file containing the dataset.	The author's repository must include a PDF “Read me” file (clearly labeled, for example, ReadMe.pdf) containing a list of all files included and guiding a user on the types of files and how to use them to do replication.	The author's repository must include a “ReadMe” file, called "README," "Readme," or similar, containing a list of all files included. Common formats are txt, PDF, and Markdown. The ReadMe file should not require proprietery software to view. It should guide a user on the types of files and how to use them to do replication.
8.1	Size of Files	For datasets that are too large to send by e-mail, we make available the option of uploading large files to our FTP server. Please contact the journal for instructions on accessing the FTP server.	The AEA Data and Code Repository can handle files up to 2GB. Please contact the openICPSR staff (where the repository is hosted) should you encounter any problems. https://www.openicpsr.org/openicpsr/problem	The AEA Data and Code Repository can handle files up to 2GB. Please contact the openICPSR staff (where the repository is hosted) should you encounter any problems. https://www.openicpsr.org/openicpsr/problem
	Size of Files, Other			Other repositories can also handle large files. Please consult the relevant support pages for any questions.
9	Detailed information	https://www.aeaweb.org/journals/policies/data-availability-policy	https://docs.google.com/document/d/1Po19f5lSgmtwdgxgjZpFnNJC3L1XuOZ9LK0HlNHuYvk/edit?usp=sharing	

Figure 3. Instructions for proposed data and code availability policy

Note: Also available at https://docs.google.com/spreadsheets/d/1khrXxnmKC7Llj9vH17r1KEkN0hcTVqZnQloretcOwDA/edit?usp=sharing

REFERENCES

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

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.

Brinckman, Adam, Kyle Chard, Niall Gaffney, Mihael Hategan, Matthew B. Jones, Kacper Kowalik, Sivakumar Kulasekaran,

Bertram Ludäscher, Bryce D. Mecum, Jarek Nabrzyski, Victoria Stodden, Ian J. Taylor, Matthew J. Turk, and Kandace Turner. 2018. "Computing Environments for Reproducibility: Capturing the "Whole Tale"." *Future Generation Computer Systems.*

Burman, Leonard E, W Robert Reed, and James Alm. 2010. "A Call for Replication Studies." *Public Finance Review*, 38(6): 787–793.

Butler, Courtney, and Christina Kulp. 2018. "The Role of Data Supplements in Reproducibility: Curation Challenges."

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).

Chris Chambers. 2014. "Registered Reports: A Step Change in Scientific Publishing."

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.

Creative Commons. 2017. "About The Licenses."

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.

Gentzkow, Matthew, and Jesse M Shapiro. 2014. "Code and Data for the Social Sciences: A Practitioner's Guide." Stanford University Mimeo.

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. 2017. "What is Replication? The Possibly Exemplary Example of Labor Economics."

Hamermesh, Daniel S. 2007. "Viewpoint: Replication in Economics." *Canadian Journal of Economics*, 40(3): 715–733.

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

Hrynaszkiewicz, Iain, Aliaksandr Birukou, Mathias Astell, Sowmya Swaminathan, Amye Kenall, and Varsha Khodiyar. 2017. "Standardising and Harmonising Research Data Policy in Scholary Publishing." *International Journal of Digital Curation*, 12(1): 65–71.

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.

Nosek, Brian A., and Daniël Lakens. 2014. "Registered Reports: A Method to Increase the Credibility of Published Results." *Social Psychology*, 45(3): 137–141.

Open Source Initiative. 2018. "Open Source Licenses by Category."

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.

Vilhuber, Lars. 2018. "Computational Tools for Social Scientists Workshop."

Vilhuber, Lars, and Carl Lagoze. 2019. "Metajelo: A Metadata Package for Journals to Support External Linked Objects." *International Journal of Digital Curation*, submitted.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K Teal. 2016. "Good Enough Practices in Scientific Computing."

Replication Team

The following members of the Replication Lab provided valuable assistance to the Data Editor in one or more of his tasks:

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