Report for 2019 by the AEA Data Editor

By LARS VILHUBER[*], JAMES TURITTO[†], KEESLER WELCH[‡]

The American Economic Association (AEA) Data Editor's stated mission is to "design and oversee the AEA journals strategy for archiving and curating research data and promoting reproducible research" (Duflo and Hoynes, 2018). How to follow through on this mission was articulated in the 2018 Report by the Data Editor (Vilhuber, 2019).

Over the past year, we have worked on infrastructure to allow for greater reproducibility of empirical articles in economics, at the AEA journals and elsewhere. Since July 2019, the first elements of that infrastructure are emerging and being made visible and available to researchers and editors alike. In particular, we have updated the data and code availability policy (DCAP) (Section I), started comprehensive pre-publication verification of reproducibility (Section II.B), made historical supplemental archives more accessible and findable (Section I.A and II.A), and coordinated with other journals, societies, and registries on these topics (Section III).

I. Task 1: Updating the AEA Data and Code Availability Policy

The AEA's data availability policy was mostly unchanged since it was first published in 2008 (American Economic Association, 2008). In the meantime, other journals have strenghtened their policies, for instance to support the verification of reproducibility during the editorial process (Jacoby, Lafferty-Hess and Christian, 2017; Christian et al., 2018). Cross-disciplinary working groups in the scientific publishing community have created template data availability policies (Hrynaszkiewicz et al., 2017).

In Vilhuber (2019), we set out to modernize the AEA's data and code availability policy, in particular to bring the policy in line with the Findable, Accessible, Interoperable, Re-usable (FAIR) principles (FORCE11, 2016). A modern data and code availability policy should support an accurate and transparent description of the provenance of the scientific results. 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.

On July 16, 2019, the AEA announced an updated DCAP (American Economic Association, 2019), also published as American Economic Association (2020). Additional resources to guide researchers are made available through various frequently asked questions (FAQ) and guidance documents.[1] We summarize the key modifications here.

The policy now clearly applies to code as well as data, and explains how to proceed when data cannot be shared by an author. Materials must now be made available to the AEA prior to *acceptance*. Computer code should be provided for all stages of the data cleaning and data analysis (code for the data cleaning portion was previously optional). We also clarified that raw data must be uniformly made available, when

[*] Cornell University, lars.vilhuber@cornell.edu.
[†] J-PAL, jturitto@povertyactionlab.org.
[‡] J-PAL, keesler@povertyactionlab.org.

[1] See https://www.aeaweb.org/journals/policies/data-code/faq and https://aeadataeditor.github.io/aea-de-guidance/ for a list of such resources.

permissions allow. This is also true for author-collected survey data and for data from experiments. For restricted-access or proprietary data, to the extent permissible, the data must be made available to the AEA Data Editor for verification, even if the data cannot be published by the author. When that is not feasible, the AEA Data Editor will verify the access procedures. Where feasible within time and monetary constraints, we will obtain access to the data ourselves, in order to verify the reproducibility, and where this is not possible for us, we will actively seek out assistance from others.[2] xxx Although the ultimate responsibility lies with authors (see Section I.D on Intellectual Property), we will continue to verify that authors have the right to publish the data, check for obvious personally identifiable information (PII), and ask for licenses, written permissions, etc. as needed. We must regularly reject authors' offer to publish the data as part of a data and code supplement, because, after inspection of data use agreements, terms of use, and licenses, the authors actually do not have the rights to do so. In those cases, as in all other cases, we ask the author to provide detailed information on how others may also obtain usage rights to the data from the original rights holder. We have also encountered cases where the authors mistakenly thought they did not have redistribution rights, and after our inquiry, provided such data (and code) as part of the supplement.

To implement the policy, the AEA Data Editor will review README files, appendices to the article, and the contents of data and code archives, in order to assess whether the policy's criteria of "clearly and precisely documented" and "readily available" are met. The criterion "sufficient to permit replication" is verified by testing computational reproducibility, where possible. We describe the infrastructure to support such pre-publication verification in Section II.B.

A. *Improving Findability of Replication Materials at the AEA*

Since the implementation of the first data availability policy, accessibility of supplements associated with AEA publications, when provided, was quite good. Most were referenced in footnotes in the articles, pointing to the AEA website. ZIP files were accessible via links on each article's landing page, if data and code were shareable. Findability, however, was suboptimal. The ZIP files encapsulating datasets and code were opaque, with little or no immediate information (metadata) on what those packages contained. Interested researchers needed to download the ZIP file before being able to assess whether the replication materials contained any or all the data and code necessary to reproduce the article's and figures.

In July 2019, the AEA began using a data and code repository hosted by Inter-university Consortium for Political and Social Research (ICPSR). We describe the infrastructure supporting this change in the next section. All files that are part of data and code supplements are to be deposited at the new repository, and can be browsed and inspected with ease. ZIP files are no longer accepted as supplementary packages, other than as a convenient method to import files to the data and code repository.[3] Supplements are tagged with JEL codes as well as other keywords (e.g., "Current Population Survey" or "behavioral study"). In addition, authors are encouraged to provide additional methodological information, such as the time period or geographic region covered by the data collected, or the survey method used. All of this informa-

[2] At the time of this writing, we have availed ourselves successfully of third-party support for cases where state administrative and federal tax data were used, and some authors have been able to share data with the Data Editor that cannot be subsequently published.

[3] Due to technical limitations, a small number of supplements that provide more than 1,000 files will still be partially zipped.

tion is encoded into the archive's metadata, and is used by various search engines, such as the native search engine on ICPSR, through Google Dataset Search[4] or through Digital Object Identifier (DOI) registries such as DataCite[5]. We expect that the entries for data and code supplements will increasingly surface on more scholarly search engines such as Web of Science[6] and Google Scholar[7].

Deposits receive their own DOI, and can be cited on their own (see Figure 1). Authors can and should henceforth cite their data supplement. Conversely, the article that a supplement is associated with is clearly identified on the supplement's landing page. Future enhancements to the platform will be able to display other articles that also cite the supplement.

In addition to own supplements, we are also verifying that authors cite the datasets they have used and accessed. While this has been policy at the Association's journals for several years, enforcement has been difficult: In order to identify a missing data citation, the use of a dataset must be identified first. The verification process under the Data Editor now has the means to do so, and has started to verify that all authors comply with the AEA citation guidelines that require data citations. Data citations also substantially increase findability of data, allow data providers to receive proper credit, and align the Association with broader principles in the academic publishing world (Altman and Crosas, 2013; DataONE, 2011; Data Citation Synthesis Group, 2014; Cousijn et al., 2018). We have also updated the online Sample References[8] with refreshed examples.

[4] https://toolbox.google.com/datasetsearch
[5] https://search.datacite.org/
[6] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[7] https://scholar.google.com/
[8] https://www.aeaweb.org/journals/policies/sample-references

B. *Third-party repositories*

Many other repositories and archives exist, and can be linked to. Archives exist at research institutions (institutional repositories, Harvard Dataverse[9] and other Dataverse instances around the world, Zenodo[10]), as nonprofits (Dryad[11]), and as commercial companies (Figshare[12], Mendeley Data[13]). These can be open-access data archives created by authors of AEA articles (Clemens, 2017; Gentzkow, Shapiro and Sinkinson, 2011), code archives created by authors even before submitting to the AEA journals, or DOI for restricted-access data, for instance, at the German Research Data Center (FDZ) at the Institute for Employment Research (IAB) (Schmucker et al., 2018*a*,*b*) or at the National Archive of Criminal Justice Data (NACJD) (Campbell et al., 2013). By systematically linking out to other platforms, we can accommodate, in a principled fashion, other archives as well. By doing it homogeneously for all archives and repositories, we give the researcher the flexibility to initiate the process early in the research data lifecycle. The AEA DCAP allows the use of such archives in lieu of depositing the materials at the AEA Data and Code Repository, as long as the archive is deemed to satisfy certain criteria.[14] The final determination will be made by the AEA Data Editor. We note, however, that if the AEA Data Editor determines that data or code can be made publicly available, deposit in a restricted-access archive is not acceptable. For instance, restricted-access research environments such as

[9] https://dataverse.harvard.edu/
[10] https://zenodo.org/
[11] https://datadryad.org
[12] https://figshare.com/
[13] https://data.mendeley.com/
[14] CoreTrustSeal (2017) defines criteria for trusted archives, though not all reputable archives have such a seal. Several entities maintain lists of acceptable archives, including FAIRsharing.org (https://fairsharing.org/), Nature Scientific Data (2018), and various others.

3 **Romer, Christina D., and David H. Romer.** 2010. "Replication data for: The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Association* [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E112357V1.

Figure 1. : Supplement citation, as per AEA Style Guide

Supplement citation as per AEA "Sample References," accessed at https://www.aeaweb.org/journals/policies/sample-references on February 20, 2020.

the IAB FDZ and the Federal Statistical Research Data Centers (FSRDC) routinely release computer code, and all such code must therefore be deposited and made available in an open access repository, before the manuscript is accepted for publication.[15]

C. *Post-publication modifications*

At the time of publication, the article will be linked with one (or more) archived supplements, constituting the version of record. What happens when, despite pre-publication scrutiny, an error in the code emerges? For instance, while pre-publication verification uses the same version of software and packages as the authors, software bugs may later emerge. The AEA Data and Code Repository provides a means to do so. Authors can update their supplement, generating a new version, say "V2", that corrects for the error. The earlier version — "V1" – remains linked to the article as "the version of record," but the new version can be found and used by any reader of the article following the link to the supplemental code.

[15]It is worth reiterating that personal websites, Github.com (https://github.com), Google Drive (https://drive.google.com), Dropbox (https://dropbox.com), and others 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, including because they simply did not pay their monthly or annual fee. Such practices are incompatible with proper data curation standards.

D. *Intellectual Property*

In moving to author-initiated creation of archives, we also changed how intellectual property rights associated with the supplementary materials are handled. For new deposits at the AEA Data and Code Repository, authors retain the copyright. The default license for all repositories based at openICPSR is the Creative Commons Attribution (CC-BY) (Creative Commons, 2017). Based on guidance in Stodden and Reich (2012), and after consultation with the Association's counsel, we suggest CC-BY for databases and an open source license (Open Source Initiative, 2018) for software and code. Authors may choose to license their data and code under a different license, though such licenses will be vetted by the Data Editor for compliance with the DCAP. For historical archives, authors transferred the copyright to the AEA upon publication. In migrating to the AEA Data and Code Repository, we are re-licensing these under a mixed license, combining CC-BY for data and "modified BSD" license for code. This license allows for liberal re-use by others, while ensuring that the credit is given to the authors.

II. Task 2: Creating Infrastructure at the AEA Journals

As laid out in Duflo and Hoynes (2018), the second task consists in creating an infrastructure

for enhanced transparency at the AEA journals. This infrastructure has two key components: the AEA Data and Code Repository as a transparent, strongly curated repository, with expanded visibility onto supplements, and greater findability of those supplements through various channels. The second component is the infrastructure necessary to conduct pre-publication verification of computational reproducibility at scale.

Importantly, the use of a dedicated platform for data storage also allows us to simplify the process of channeling the materials from authors to the final data publication. Traditionally, authors have emailed ZIP files to the editorial office, or provided links to folders on shared storage platforms in a variety of ways. The editorial office then had to repackage those materials, before posting them on the AEA website. Going forward, authors upload their materials directly to the ICPSR platform, being able to preview what they will look like once published. The AEA Data Editor can access the materials prior to their publication. Once both authors and the AEA Data Editor have approved the contents of the data deposit, it is published simultaneously with the article.[16] Authors will also be the primary creators of the metadata about their supplements, ensuring accuracy.

Critically, prior to publication of manuscript, data, and code, members of the Data Editorial staff will download the materials for the purpose of verifying the completeness, accuracy, and computational reproducibility (see below). The Data Editor staff will also verify the completeness of the metadata provided on the ICPSR website, and correspond with the author if necessary to correct or augment the information.

A. *Migrating Historical Supplements*

The process above will apply to all new data supplements. However, we have also started the process of migrating all historical supplements to the new platform. Between Oct 11 and Oct 13, 2019, the staff at openICPSR ingested 2,552 historical supplements (473 Gb).[17] This was only the first part of the migration, as there are about 1,000 more archives that need to be migrated.[18]

While the median supplement is only 1.5 Mb large, and has 12 files, the largest supplements can be very large (see Figure 2). The largest in the current batch of migrated supplements has 30.5 Gb and 795 files. The ZIP files in which these supplements had been made available are now expanded, and files can be inspected and downloaded individually, as needed. This makes data, code, and README files easy to inspect. Migrated supplements come from a range of years (Figure 3). Most supplements (72.96 percent) use Stata at least partially, followed by Matlab (22.45 percent) (Table 1 and Figure 4). This distribution is quite persistent across years (Figure 5). Open-source software packages (R, Python) have only a small presence.

Metadata stemming from the articles, such as JEL codes, are incorporated into the new deposit in the AEA Data and Code Repository. Each migrated deposit obtains their own citable DOI, and links back to the article it is associated with. Links on the AEA website that used to point to the ZIP files have been adjusted to point to the new DOI. Every deposit is also findable through native search engine on ICPSR and through Google Dataset Search.

[16] Authors can always publish the supplements earlier.

[17] The data for this part of the analysis can be found at Vilhuber (2020*a*).

[18] Another 500 supplements were migrated in December 2019.

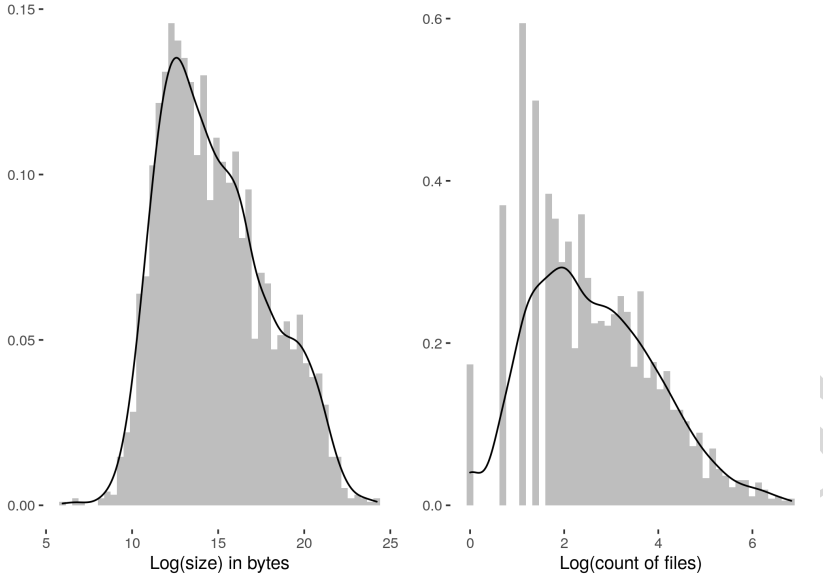

Figure 2. : Supplement statistics, files migrated in October 2019

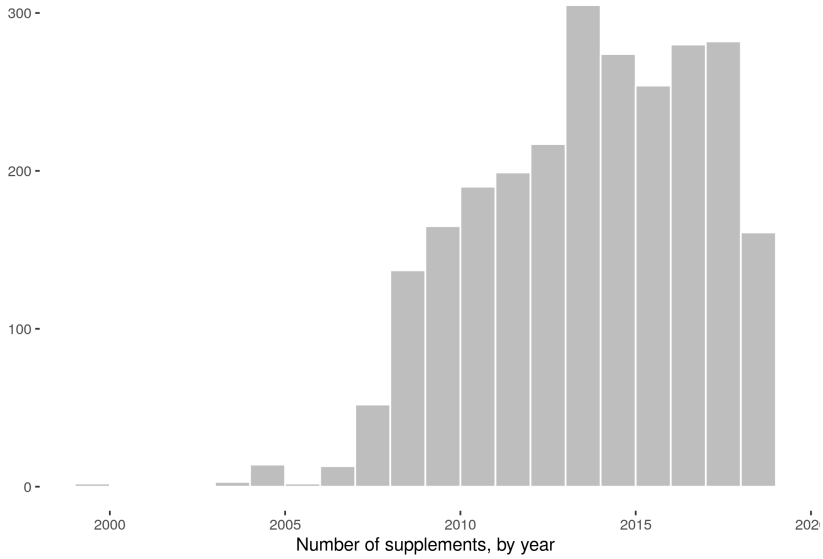

Figure 3. : Supplements by year, files migrated in October 2019

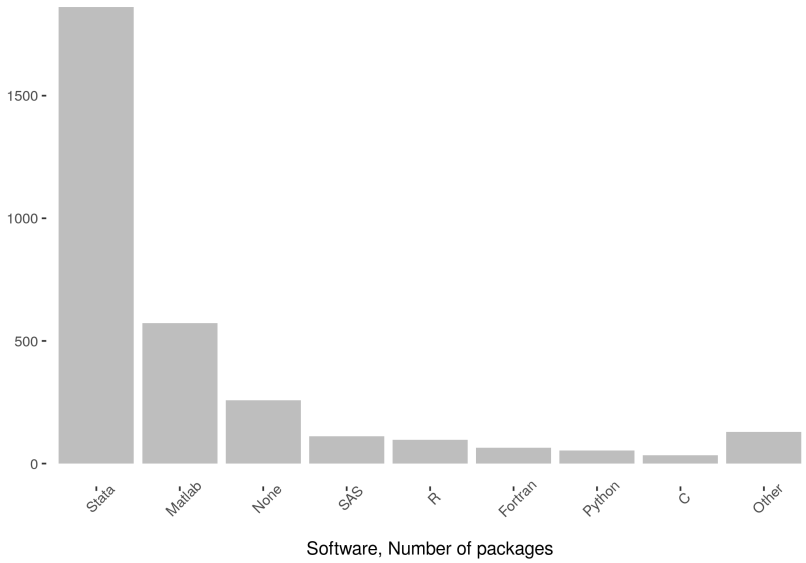

Figure 4. : Software by supplement, files migrated in October 2019

Table 1—: Software usage in supplements

Software	Supplements	Percent
Stata	1,862	73.0
Matlab	573	22.4
None	258	10.1
SAS	111	4.3
R	97	3.8
Fortran	64	2.5
Python	54	2.1

Software usage in supplements.
A supplement can use more than one software.
Software with less than 2 percent utilization not listed.

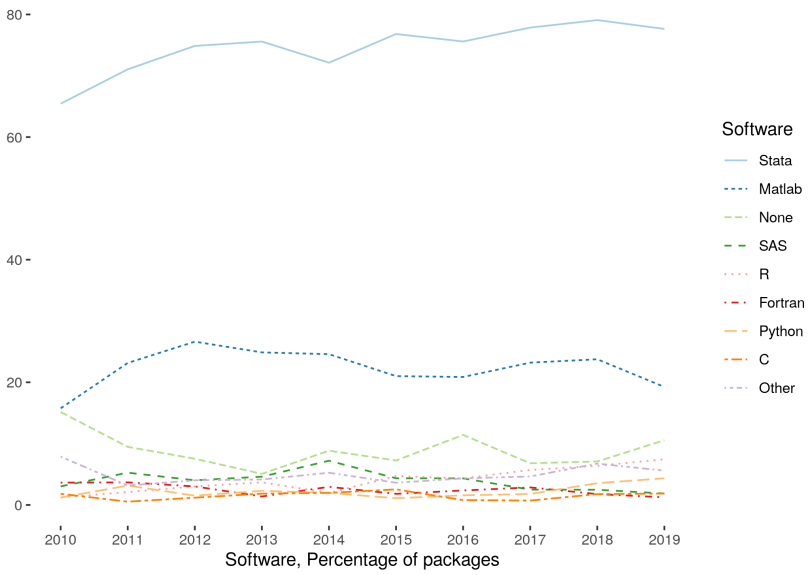

Figure 5. : Software by supplement, across years, files migrated in October 2019

B. *Pre-publication verification of computational reproducibility*

The other component of the infrastructure is a scalable implementation of pre-publication verification. We consulted existing pre-publication verifications, for instance at the American Journal of Political Science (AJPS)[19] and at the Econometric Journal, as well as with verification projects at universities (e.g., Cornell University's R-Squared Service[20]). We worked closely with the AEA editorial office to identify opportunities for integration.

We determined that the pre-publication verification service would do three distinct tasks. The first task involves identifying all data used. Article and deposited supplements are analyzed as to the data they contain or describe. Tables and figures are scrutinized as to the data used in their creation, and compared to the authors' description. This serves two purposes. First, to verify the presence of data citations, as required by AEA style guides[21]. Second, to identify whether all data necessary for verification could be made available, and whether the description of where datasets come from (provenance) and how they could be accessed (license, requests, etc.) are well-described. The second task involves analyzing the code provided. Even before executing any code, staff members determine whether the code is complete. Furthermore, they inspect code and README to determine whether the creation of tables and figures can be identified. Finally, the third task involves bringing all these elements together, and re-executing the computer code provided. The resulting tables and figures are compared to those in the paper.

[19] https://ajps.org/wp-content/uploads/2018/05/ajps_quantitative-data-verification-checklist.pdf

[20] https://ciser.cornell.edu/research/results-reproduction-r-squared-service/

[21] https://www.aeaweb.org/journals/policies/sample-references

There is little support for this workflow within existing academic review systems. We developed and continuously adapted a workflow, using online tools.[22] Together with the editorial team, we added a *conditional acceptance* stage that did not previously exist in the AEA workflow. During this stage, manuscripts get assigned to the AEA Data Editor, and the verification process described above is initiated. A member of the Replication Lab (see list at end of article) is assigned to a paper, and makes an assessment by moving through the workflow described above. The report they prepare is reviewed by the AEA Data Editor, and then submitted through the traditional manuscript workflow back to the author. This process is similar to, but distinct from the usual refereeing of manuscripts. The process can go through one or more assessment rounds, until the AEA Data Editor confirms that a manuscript and its supplement comply with the DCAP. The manuscript is then definitively accepted, and moves forward through the usual publication process. The deposit at the AEA Data and Code Repository is linked to the article once a DOI has been assigned to the article, and is then published.

Between July 16, 2019, and November 28, 2019, the AEA Data Editor team conducted 216 assessments for 138 manuscripts. For comparison, the AJPS conducts pre-publication assessments for about 65 manuscripts per year, based on the number of supplements published at https://dataverse.harvard.edu/dataverse/ajps. We collected metrics from the online system.[23] Figure 6 shows the distribution of assessments across journals. Figure 7 shows the number of rounds that the 138 completed manuscripts have gone through. Assessment take varying amounts of time, depending on the complexity of the paper and the code. Figure 8 shows the distribution of the time it takes to complete each round of assessment. The distribution of the total length of all revision rounds, from the first submission (to the Data Editor) to final acceptance, is shown in Figure 9. Figure 10 shows the component of the total length that is due to author response time, i.e., the time between the filing of a report by the AEA Data Editor requesting changes, and the time the manuscript is re-assigned to the AEA Data Editor.

[22] We used Jira software (https://www.atlassian.com/software/jira), together with private Git repositories. A detailed description will be available.

[23] Data can be found at Vilhuber (2020*b*).

DELAYS

A recurring concern expressed by editors and staff members was the potential for delays in publication, due to the verification process. The data presented here indicates that concerns of about a delay in the time-to-publication remain valid, though we have no evidence at this time that the overall time to publication has increased — the first articles that completed the entire pre-publication verification process were published in late January 2020, after the analysis in this article was completed.

We are addressing these concerns in a variety of ways. First, the current set of manuscripts moving through the assessment process did not anticipate the process — the manuscripts were written and refereed prior to the July announcement. They thus could not incorporate stronger guidance as easily as future submissions will be able to. Second, we are expanding guidance, as well as preparing template README and project structures that will assist future manuscript submissions to be more rapidly compliant with the DCAP, ideally in a first pass. This may include stronger requirements about the information to be provided by authors, including detailed information about the length of time it takes to execute the computer code. Third, there is anecdotal evidence that other factors affecting the time-to-publication have been reduced, due to the move to conditional acceptance. Thus, manuscript materials are being

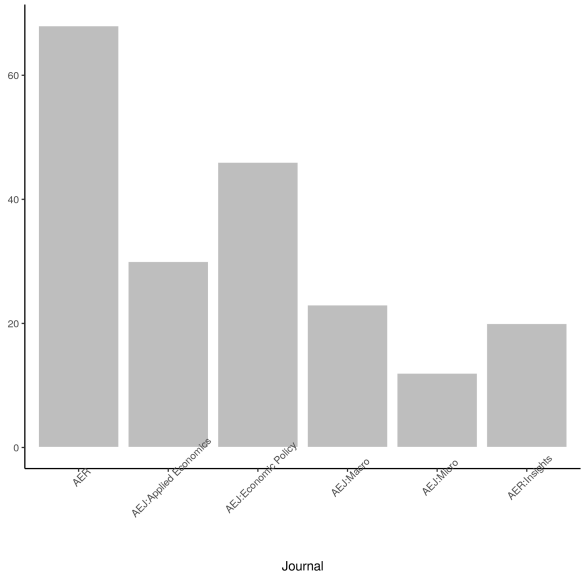

Figure 6. : Number of assessments

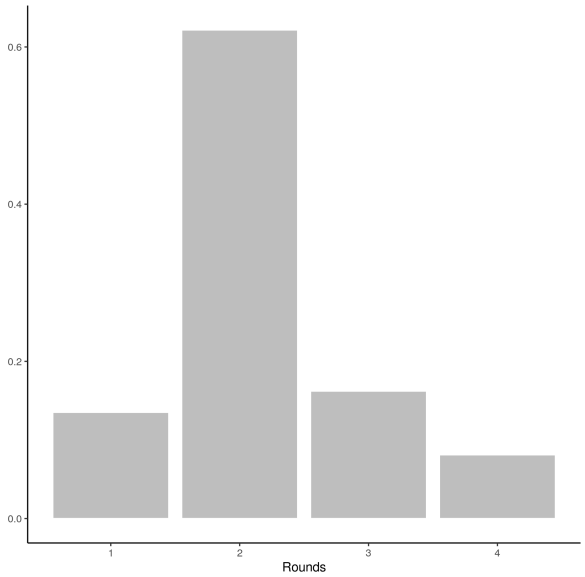

Figure 7. : Assessment rounds for completed manuscripts

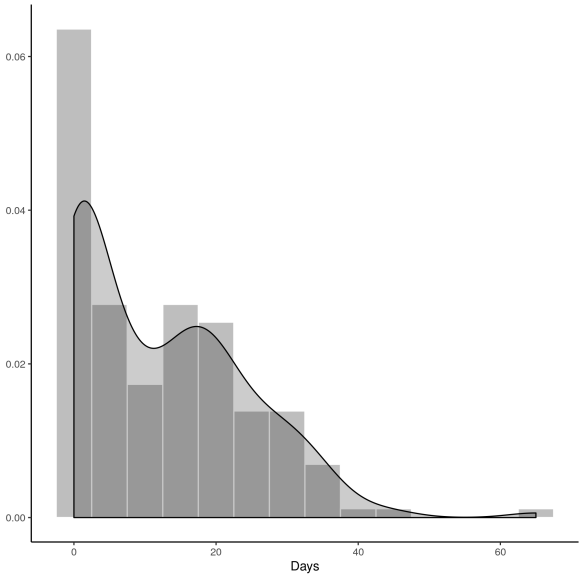

Figure 8. : Length of an assessment round in days

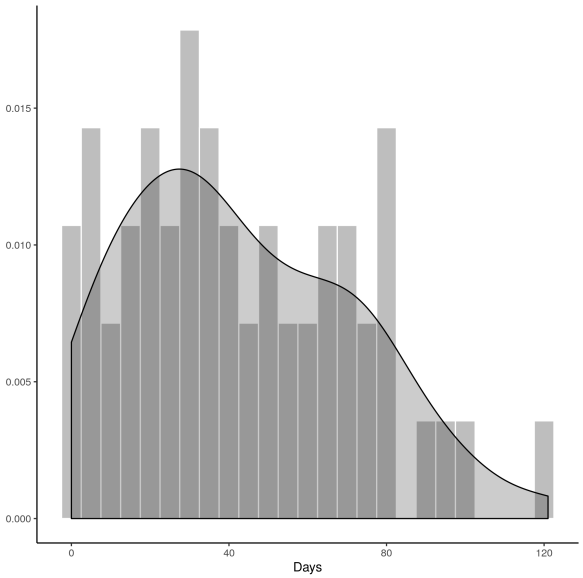

Figure 9. : Length of revisions for completed manuscripts

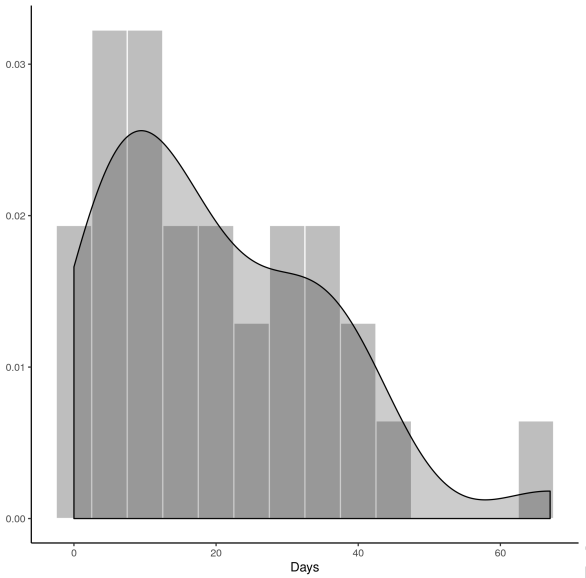

Figure 10. : Author response time

completed earlier than before.

We note that none of the 138 manuscripts we assessed had fundamental flaws — all problems identified so far have been fixable (and fixed).

III. Task 3: Working with Other Providers of Scientific Infrastructure to Improve Support for Documenting Provenance and Replicability

An important component of the AEA Data Editor's position is to interact with other providers of scientific infrastructure. This involves other publishers and journals, archives such as ICPSR, providers of restricted or proprietary data, the AEA RCT Registry, metadata harvesters, and third-party verification services.

A. Economics Journals

We have coordinated with other data editors conducting similar activities at other journals. In 2019, both the Review of Economic Studies (ReStud) and the Economic Journal (EJ) appointed data editors, tasked among other things with pre-publication verification. The Canadian Journal of Economics (CJE) is currently revising its DCAP. The Journal of the American Statistical Association (JASA) is expanding the scope of its pre-publication verification to all sections of the journal. In all cases, either the AEA Data Editor or an ad-hoc group of "Social Science Data Editors" initiate by the AEA Data Editor, was consulted. The "Social Science Data Editors" group includes editors from journals that do not (yet) have a pre-publication verification service. A website[24] provides examples, checklists, and links to training materials, to assist authors in improving data and code archives prior to submission, regardless of the journal they may be submitting to.

[24] socialsciencedataeditors.github.io

B. *Search services, text mining access*

As noted earlier, the move of supplements into the openICPSR-supported repository enables broad dissemination of metadata on supplements. Among others, Google Dataset Search harvests and then displays such metadata. Staff from openICPSR and the AEA Data Editor have been discussing informally with the Google Dataset Search team on how to correctly display metadata as it moves through the various scrapers.

Over the course of the past year, the AEA has been approached by two research projects that wished to use the full-text versions of AEA articles to enhance the linkage between articles and data. We worked with these teams to establish a policy by which they could legally access the full archive of articles, but also subsequently make their work available to others. AEA counsel is currently working on a full-fledged data and text mining policy.

C. *Third-party verification services*

We have started discussions with third-party verification services. In particular, we have trialled using their services, instead of our in-house team, to conduct assessments. This aligns with the discussions among editors about resources and scalability of assessments, and with the educational outreach, which some of these services also conduct. Part of the effort consists in aligning the criteria used by these services with those of the various journals.

D. *AEA RCT Registry*

The AEA RCT Registry and the AEA Data Editor regularly consult on data and publication-related issues. On August 12, 2019, the Registry announced that henceforth, registrations could be cited using the newly assigned DOIs. Furthermore, new submission guidance at the AEA journals[25] now requires that registrations be cited, not simply mentioned. As part of the AEA Data Editor's data and code review process, manuscripts are checked for compliance with that requirement. Sample RCT citations have been added to the Sample References[26] (see also Figure 11).

```
Zhang, Kelly. 2017. "Voter
Pessimism and Electoral
Accountability: Experimental
Evidence from Kenya." AEA
RCT Registry. May 02.
https://doi.org/10.1257/rct.5-8.0.
```

Figure 11. : Example of a RCT citation

Since its inception in 2012, the Registry has seen significant growth in the number of visitors, registrations, and the number of pre-registrations submitted by researchers in the social sciences. In October 2019, there were 4.4k visitors to the website. The Registry reached 3,000 entries in November 2019 (see Figure 12). The AEAs decision in early 2018 to *require* registration of all randomized controlled trials prior to submission for journal publication has been a major contributor to growth.

Two trends emerge from an analysis of the registry data. First, more studies are being registered before their intervention start date. Figure 13 compares the number of pre-registered studies to those registered after the intervention start date per quarter over time.

Second, the use of pre-analysis planss (PAPs) is becoming increasingly common in economics and other social sciences. Registering a pre-analysis plan is optional on the AEA RCT Registry, but Figure 14 shows that many researchers have registered pre-analysis plans along with their studies.

[25] https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide

[26] https://www.aeaweb.org/journals/policies/sample-references

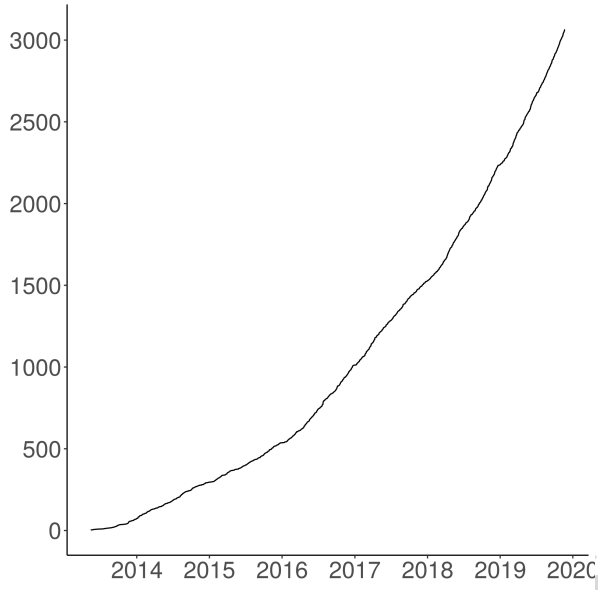

Figure 12. : Cumulative Count of Registrations at AEA RCT Registry

Since early 2020, monthly snapshots of the AEA RCT Registry data can be downloaded from the AEA RCT Registry Dataverse[27]. The first release was made on January 27, 2020 (AEA RCT Registry, 2020).

IV. Task 4: Working with the Economics Community to Enhance and Broaden Education on Replicable Science

From the preliminary verification work has emerged a clear indication that education and training is critical for the adoption of reproducible methods. We are expanding the support materials, and are adapting them to each discipline's idiosyncratic methods. Training materials will be made available by various organizations, with input from the AEA Data Editor. The AEA Data Editor's website[28] as well as the website of the Social Science Data Editors[29] are being regularly updated with additional guidance, and will hopefully allow authors to be responsive to various journals' DCAP.

V. Data and Code Availability Statement

All data and code used to generate figures and tables in this article can be found at Vilhuber, Wasser and Turitto (2020), with some data archived as Vilhuber (2020*a*) and Vilhuber (2020*b*). Data on the AEA RCT Registry was extracted by JT and KW from internal systems, but can now be downloaded directly (AEA RCT Registry, 2020). The data provided here as part of Vilhuber, Wasser and Turitto (2020) may differ.

[27] https://dataverse.harvard.edu/dataverse/aearegistry

[28] https://aeadataeditor.github.io/aea-de-guidance/

[29] https://social-science-data-editors.github.io/guidance/

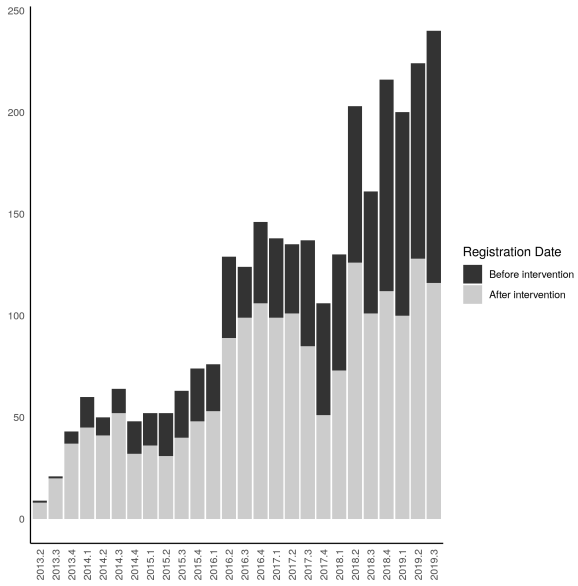

Figure 13. : Pre-Registration of Studies

NOTE: Number of studies, as of September 2019.

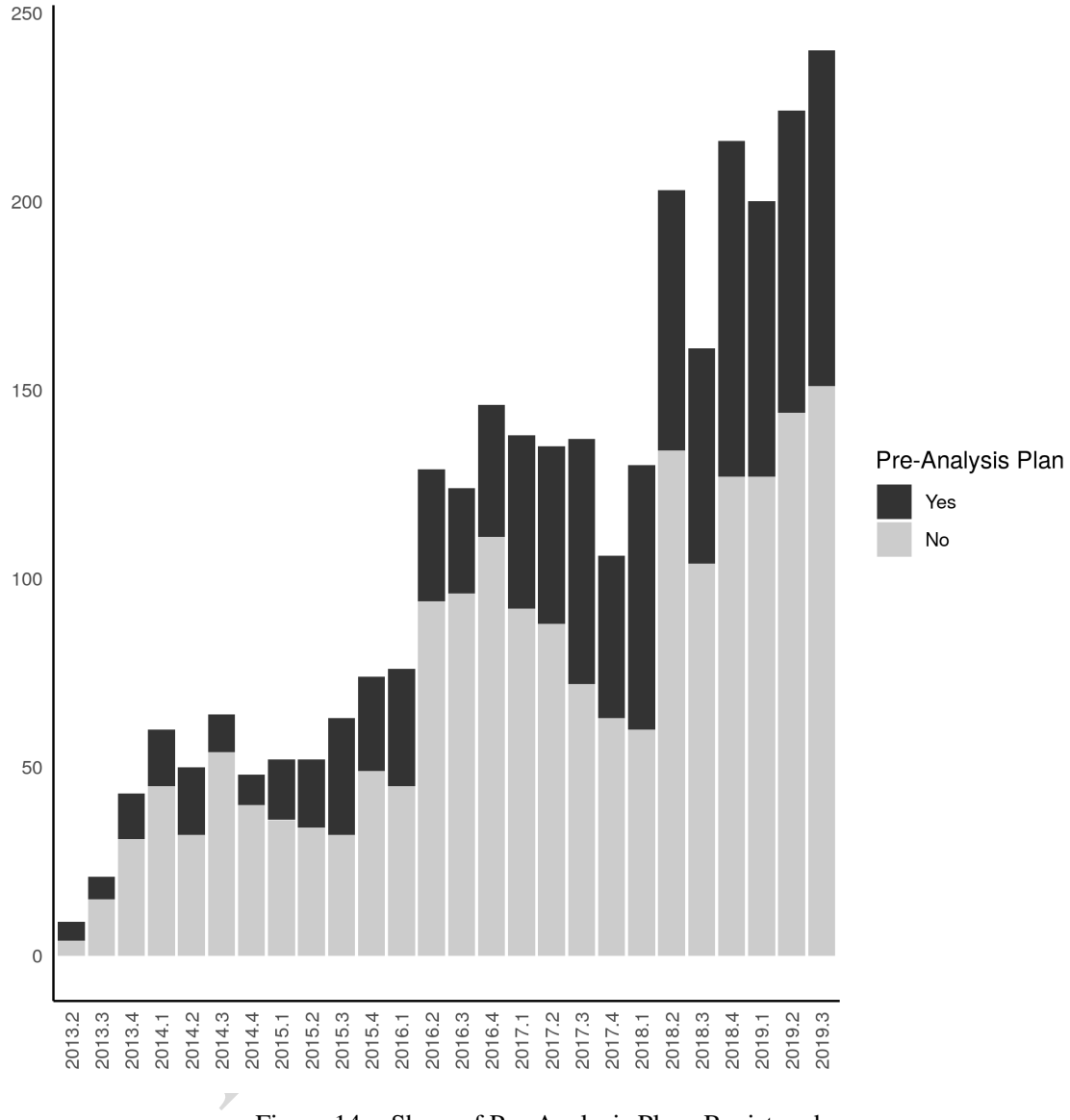

Figure 14. : Share of Pre-Analysis Plans Registered

NOTE: Number of studies, as of September 2019.

REFERENCES

AEA RCT Registry. 2020. "Registrations in the AEA RCT Registry (2013-04 to 2020-01)." Harvard Dataverse UNF:6:kWiM5wm1x75KKsxWAAqr4g==, https://doi.org/10.7910/DVN/DFMLIU.

Altman, Micah, and Mercè Crosas. 2013. "The Evolution of Data Citation: From Principles to Implementation." *IASSIST Quarterly*, 62–70.

American Economic Association. 2008. "Data Availability Policy." https://web.archive.org/web/20180927113622/https://www.aeaweb.org/journals/policies/data-availability-policy (accessed 2019-09-21).

American Economic Association. 2019. "AEA Member Announcements: Updated AEA Data and Code Availability Policy." https://web.archive.org/web/20191208160745/https://www.aeaweb.org/news/member-announcements-july-16-2019 (accessed 2019-09-21).

American Economic Association. 2020. "Data and Code Availability Policy." *AEA Papers and Proceedings*, 110: xxx. https://doi.org/10.1257/pandp.110.xxx.

Bollen, Kenneth, John T. Cacioppo, Robert M. Kaplan, Jon A. Krosnick, and James L. Olds. 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." National Science Foundation Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf (accessed 2018-05-20).

Campbell, Rebecca, Megan Greeson, Deborah Bybee, and Angie Kennedy. 2013. "Adolescent Sexual Assault Victims' Experiences with SANE-SARTs and the Criminal Justice System, 1998-2007." https://doi.org/10.3886/ICPSR29721.V1.

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." https://doi.org/10.17605/osf.io/cfdba.

Clemens, Michael. 2017. "Raw Scanned PDFs of Primary Sources for Workers, Wages, and Crops." https://doi.org/10.7910/DVN/DJHVHB.

CoreTrustSeal. 2017. "Data Repositories Requirements." https://www.coretrustseal.org/why-certification/requirements/ (accessed 2018-06-14).

Cousijn, Helena, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone, and Tim Clark. 2018. "A Data Citation Roadmap for Scientific Publishers." *Scientific Data*, 5: 180259. https://doi.org/10.1038/sdata.2018.259.

Creative Commons. 2017. "About The Licenses." https://web.archive.org/web/20181208161819/https://creativecommons.org/licenses/ (accessed 2018-12-08).

Data Citation Synthesis Group. 2014. "Joint Declaration of Data Citation Principles." , ed. Maryann Martone. Force11. https://doi.org/10.25490/A97F-EGYK.

DataONE. 2011. "Data Citation and Attribution." (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. https://doi.org/10.1257/pandp.108.745.

FORCE11. 2016. "THE FAIR DATA PRINCIPLES." https://www.force11.org/group/fairgroup/fairprinciples (accessed 2017-05-26).

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. 2011. "United States Newspaper Panel, 1869-2004: Version 6." https://doi.org/10.3886/ICPSR30261.V6.

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

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. https://doi.org/10.2218/ijdc.v12i1.531.

Jacoby, William G., Sophia Lafferty-Hess, and Thu-Mai Christian. 2017. "Should Journals Be Responsible for Reproducibility?" https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility (accessed 2018-07-22).

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science.* Washington, D.C.:National Academies Press. https://doi.org/10.17226/25303.

Nature Scientific Data. 2018. "Nature Scientific Data Recommended Repositories." figshare, https://doi.org/10.6084/m9.figshare.1434640.v12.

Open Source Initiative. 2018. "Open Source Licenses by Category." https://web.archive.org/web/20181208161736/https://opensource.org/licenses/category (accessed 2018-12-08).

Pesaran, Hashem. 2003. "Introducing a Replication Section." *Journal of Applied Econometrics*, 18(1): 111–111. https://doi.org/10.1002/jae.709.

Schmucker, Alexandra, Andreas Ganzer, Jens Stegmaier, and Stefanie Wolter. 2018*a*. "Betriebs-Historik-Panel 1975-2017." https://doi.org/10.5164/IAB.BHP7517.DE.EN.V1.

Schmucker, Alexandra, Johanna Eberle, Andreas Ganzer, Jens Stegmaier, and Matthias Umkehrer. 2018*b*. "Betriebs-Historik-Panel 1975-2016." https://doi.org/10.5164/IAB.FDZD.1801.DE.V1.

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*. https://doi.org/10.2139/ssrn.2108510.

Vilhuber, Lars. 2019. "Report by the AEA Data Editor." *AEA Papers and Proceedings*, 109: 718–29. https://doi.org/10.1257/pandp.109.718.

Vilhuber, Lars. 2020*a*. "Data files for AEA Repository migration." American Economic Association [publisher], https://doi.org/10.3886/E117873V1.

Vilhuber, Lars. 2020*b*. "Process data for the AEA Pre-publication Verification Service." American Economic Association [publisher], https://doi.org/10.3886/E117876V1.

Vilhuber, Lars, David Wasser, and James Turitto. 2020. "Code and Data for: Report for 2019 by the AEA Data Editor." American Economic Association [publisher], https://doi.org/10.3886/E117884V1.

CONTRIBUTIONS

LV wrote most of the text and conducted most of the analyses. JT and KW wrote the section on the AEA RCT Redistry and conducted the statistical analysis for that section. David Wasser prepared the analysis of the pre-publication verification system.

REPLICATION TEAM

The Replication Team in 2019 was ably lead by David Wasser. The following members of the Replication Lab provided valuable assistance to the Data Editor in one or more of his tasks:

Alexia Ge Anthony Peraza, Aviv Caspi, Craig Schulman, Elijah B. Ruiz, Gabriel Bond, Jason S. Katz, Jeong Hyun Lee, Jiayin Song, John Park, Joshua Passel, Kirubeal T. Wondimu, Linchen Zhang, Louis Liu, Luis Lopez Cabrera, Luke O'Leary, Mary-Jo Ajiduah, Naomi Li, Nicholas Swan, Nishat Peuly, Ryan Ali, Samuel Frey, Siyang (Elaine) Yu, Steve Yeh, Weilun Shi, William Hernandez, Yanyun Chen, Yuan-Hsuan (Sharon) Lin, Zebang Xu.

DATA AVAILABILITY

The document, data, and programs for this report are available at https://github.com/AEADataEditor/report-aea-data-editor-2019-interim/.

DEFINITIONS

In this article, we adopt definitions articulated by Bollen et al. (2015) and National Academies of Sciences, Engineering, and Medicine (2019), 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. Generally, this is equivalent to "*computational reproducibility*."
- *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).
- *Generalizability* refers to the extension of the scientific findings to other populations, contexts, and time frames, perhaps using different methods (Hamermesh, 2017, "scientific replication").

*

Corrigendum 2020-08-04

After publication, Alan Riley (Stata) pointed out a discrepancy between Table 1 and Figure 5, which should be depicting (on average) the same data. This was corrected in the code associated with the AEA Repository migration, and carried through to this document. See the commit implementing the code change.

Original text (pg. 767, 2nd column):

> Most supplements (48.37 percent) use Stata at least partially, followed by Matlab (37.47 percent) (Table 1 and Figure 4).

Corrected text:

> Most supplements (72.96 percent) use Stata at least partially, followed by Matlab (22.45 percent) (Table 1 and Figure 4).