Report for 2023 by the AEA Data Editor

By LARS VILHUBER[*] JACK CAVANAGH[†]

The American Economic Association (AEA) Data Editor's mission is to "design and oversee the AEA journals' strategy for archiving and curating research data and promoting reproducible research" (Duflo and Hoynes, 2018). The 2018 Report by the Data Editor (Vilhuber, 2019) articulates how to implement that mission. We conduct comprehensive pre-publication reproducibility checks for all regular AEA journals, develop and maintain guidance for authors, and work with peers at societies and groups in economics and elsewhere. We conduct basic checks on replication packages for Papers and Proceedings. General policy and various auxiliary policies are listed in Appendix A.

In order to achieve the greatest transparency and data availability, we engage with data creators and providers to discuss access to data for narrow reproducibility checks, and for broader data availability and re-use, including providing guidance on how to make data publication compliant with FAIR practices (FORCE11, 2016), and assist them in finding additional resources.

This report also discusses progress and developments at the AEA RCT Registry (see Section III.A), on behalf of the AEA Oversight Committee for Registry of Random Controlled Trials, in line with the transparency and reproducibility goals of the Data Editor.

I. Infrastructure for Verification of Reproducibility

The Data Editor manages the infrastructure needed to access data and code, conducts reproducibility checks, and archives and preserves replication packages. In general, the first two infrastructure pieces are provided by the replication team at Cornell University, the latter primarily by the AEA Data and Code Repository provided by openICPSR at the University of Michigan, with additional support from the AEA's inhouse IT staff. In 2023, the Data Editor continued to explore the use of several other infrastructures for conducting reproducibility checks and for the preservation of data for replication packages.

A. Pre-publication verification of computational reproducibility

THE PROCESS

Pre-publicaton verification is conducted by the Data Editor's team at Cornell University. Requests for assessment of reproducibility are received and assigned to a team member, who then assesses data availability and compliance with requirements. When some data are available, a full or limited reproducibility check is conducted. If we cannot obtain access to the data or computational resources in a timely fashion, we may reach out to third-parties who can, and request a reproducibility check from them. Once all computations have been completed, a process that can take anywhere from a few minutes to several weeks, a report is compiled, reviewed and approved by the Data Editor, and submitted back to journal editors, who handle most communications with the authors.

The report will have one of four possible recommendations (see Table 1), the role of which differs from the stage ("Conditional acceptance"

[*] Cornell University, lars.vilhuber@cornell.edu.

[†] J-PAL. Cavanagh provided analysis for registry-associated data and materials.

(CA) or "Revise and Resubmit" (R&R)). During the **CA stage**, "acceptance" means that no further changes are necessary, and both the manuscript (after copy-editing) and the replication package can be scheduled for publication.[1] However, to streamline processing, we may also recommend an "acceptance with modifications requested." In such cases, the remaining modifications are minor, and can be handled during copy-editing (for instance, a small number of tables need minimal changes) and prior to publication of the replication package (for instance, a fixable error in a program, or a clarification in the README, not affecting any important tables or figures). While we check that authors comply with the request for modifications, no further computational assessment is made. A recommendation of "Conditional acceptance" implies that the manuscript needs to go through another round of author revisions (stays in the CA stage), and a revised manuscript and replication package will need to be resubmitted to address any identified shortcomings. Finally, a recommendation of "revise and resubmit" is recorded when the Data Editor has serious concerns that might warrant that referees and the journal editor have another look at the manuscript. This has never been used, but on occasion, the Data Editor will consult with the journal editor about the right process.

For some journals, we may also receive a request prior to a conditional acceptance by the journal editor, i.e., during the "R&R" stage. This is regularly used by AEJ:Applied Economics, and may be used for comment articles, at the discretion of the journal editor. During this stage, only two recommendations are habitually used: "Accept", indicating that from the point of view of the Data Editor, the replication package and the manuscript can be given a conditional acceptance, and "Revise and resubmit", which indicates that there is still substantial work by the authors in order to bring their replication package in compliance. In all cases, the replication package will be reviewed again once the journal editor has given it a conditional acceptance.

Table 1—: Recommendations

	CA	R&R
Accept	7	24
Accept - with Changes	249	0
Conditional Accept	62	0
Revise and Resubmit	0	3

Assessments made

Between 2022-12-01 and 2023-11-30, the AEA Data Editor team received 646 requests, for 527 manuscripts.[2] Requests typically are channeled to the team by the AEA's journal submission and review system, but others were initiated by authors or editors directly, often while preparing the replication materials. Of these, 426 reports (345 manuscripts) were submitted back to editors,[3] and 296 were completed up to the point of publication of the data deposit, including any post-acceptance modifications. Table 1 shows the distribution of the last recommendation on record for manuscripts as of 2023-11-30. Table 2 breaks these numbers down by journal, showing the number of requests received ("rcvd") and reports completed ("cplt") in the left panel. The right panel shows the number of manuscripts for which one or more requests were received ("rcvd") and reports completed ("cplt"). The columns marked "ext." identify cases where we reached out to external

[1] Manuscript and replication package are generally published at the same time, though at the request of either editors or authors, the replication package can be published at any time after acceptance.

[2] This includes only requests submitted between those dates, and does not take into account in-progress requests on 2022-12-01.

[3] The balance are either in progress or are not coded in the adminstrative system as having been submitted to ScholarOne, such as replication packages for Papers and Proceedings.

replicators, which we discuss later. Finally, the last column identifies manuscripts for which the entire process has been completed, and which are "pending" publication.

Typical Issues

Incomplete data provenance and data availability: Articles provide imprecise or incorrect information regarding access to data that is not provided, or for data that is provided. In some cases, authors fail to provide data that should be provided, and in other cases, authors inadvertently provide data for which they do not have redistribution rights. Examples of such data are the World Values Survey, the Panel Study of Income Dynamics, the Socio-Economic Panel, the UN COMTRADE database, and even in one case a Compustat extract. In such cases, authors are required to remove the data, or provide evidence that they have received permission from the data owner to redistribute the data.

Specification of computational environment: Sufficiently precise descriptions of the required auxiliary packages or libraries, as "manifest"-like files in R, Python, and Julia, or as "`setup.do`"-like programs in Stata remain rare. Very few replication packages use containers ("Docker"). We continue to work with some users to leverage such environments when appropriate (see our discussion later under *Computational Infrastructure*).

Incomplete instructions and manual manipulation: We observe many packages that do not have a small number of control programs ("master.do" or similar). Between 28 and 50% of packages contains such files, the remainder using manual instructions to run multiple code files. Similarly, many authors still manually save figures and copy tables. These relatively simple coding practices detract from speedy and efficient reproduction by third parties, including the Data Editor and the authors themselves. We continue to accept packages that do not do this, but are developing guidance and references which should help authors make their packages more streamlined with little extra effort.

Delays

A recurring concern expressed by authors, editors, and staff members continue to be delays in processing by the Data Editor, due to the verification process. The median manuscript is reviewed once (Table 3 shows the breakout by journal), but the duration of an evaluation round remains too long, with median times by journal hovering around 90 days. We are working, in coordination with the editors, on making that significantly shorter.

Disseminating process information

We continue to improve our documentation, based on careful monitoring of the process, and where more information could be beneficial. Our documentation aims to (a) provide authors with the information as early as possible, when it is still easy to include reproducible practices in projects at relatively low cost and (b) provide authors with the best information, to reduce frictions and uncertainty. Authors are provided with an informational form upon submission, and a short form, provided upon conditional acceptance, collects salient information about the replication package, but also links to important guidance.[4] Authors are required to provide the information defined in the Social Science Data Editors' template "README" (Vilhuber et al., 2022*b*), though they are free to use their own documents, as long as all the information is available.

Computational Infrastructure

Most replication packages are computationally verified by replicators on the computers

[4] These forms can also be found at aeaweb.org/journals/data.

Table 2—: Processing Statistics

	Issues			Manuscripts			
Jour–l	(rcvd)	(cplt)	(exter–l)	(rcvd)	(cplt)	(ext.)	(pend.)
AEA P+P	109	–	–	108	–	–	105
AEJ:Applied Economics	121	87	3	79	66	3	23
AEJ:Economic Policy	75	58	6	65	52	6	18
AEJ:Macro	77	59	4	59	45	3	15
AEJ:Micro	40	30	2	32	23	2	9
AER	133	110	7	110	92	7	73
AER:Insights	35	28	2	28	23	2	13
JEL	15	15	–	9	8	–	7
JEP	41	39	1	37	36	1	33
Totals	646	426	25	527	345	24	296

Notes: Data for requests received by the AEA Data Editor between 2022-12-01 and 2023-11-30. See text for details.

Table 3—: Assessment rounds for completed manuscripts

			American Economic Journals			
Rounds	AER	AER Insights	Applied	Macro	Micro	Policy
1	59	16	39	22	14	36
2	12	4	5	4	0	3
3	1	1	0	1	3	0

Notes: Data for manuscripts first sent to the AEA Data Editor between 2022-12-01 and 2023-11-30, and for which all rounds have been completed. AEA P&P, JEP, and JEL are excluded from this table. See text for details. Numbers differ slightly between this table and Table 2 because they are extracted from two different administrative systems, with different timing cutoffs.

available to the Data Editor at the Cornell University Economics Department and the ILR School. The majority are handled on the Windows Server systems of the Cornell Center for Social Sciences, while some are run on the Linux-based Bioinformatics cluster. Occasionally, personal macOS laptops are used. Systems can handle memory requirements up to 1024 GB or up to 100 cores.

While these systems are fairly standard, they cannot cover all scenarios described in authors' computational requirements. Furthermore, these systems, much like the authors' own systems, are not shareable more broadly, and thus sometimes make it difficult to control for specific requirements, or to share error messages in the most reproducible way.

We continue to leverage additional computational environments. We have used CodeOcean,[5] (Clyburne-Sherin, Fei and Green, 2019) both to collaboratively work with some authors on solving the problems in partially successful reproduction efforts, and to publish reproducible "capsules." We also work with the team behind "WholeTale" (Brinckman et al., 2018). Both CodeOcean and WholeTale rely on containerization, often known under the commercial name "Docker," which can be independently used to precisely define and then share computational environments. We use containers through CodeOcean and or WholeTale, when appropriate, and on the Cornell-based resources, in particular when storage or compute resources are insufficient at the public providers. Sample code can be found at the AEA Data Editor's Github repository.[6] Pre-configured Stata and manuscript-specific Docker images can be found at Docker Hub.[7] A more expansive overview of containerization issues in economics can be found at AEA Data Editor (2021). Common guidance for economists is forthcoming in collaboration with other data editors.

B. Archive for Replication Packages

The default archive for replication packages accompanying articles in AEA journals is the AEA Data and Code Repository. Deposit instructions are provided on the Data Editor's website, and provided to authors upon conditional acceptance. However, it is not the only acceptable archive, as we discuss below.

Table 4 shows statistics for all currently published replication packages at the AEA Data and Code Repository. There are currently 4627 published replication packages. Between 2022-12-01 and 2023-11-28, 496 deposits were made, with over 55 thousand files and over 1 TB of data.

In the past year, the median package size was 35.28 MB, but a significant number of packages (17 percent) had packages larger than 2GB. 3 percent of deposits were larger than 20GB. Some packages have more than 1,000 files, hitting a technical constraint on openICPSR. Provision of opaque ZIP files are generally prohibited. Instructions on how to proceed when file numbers are large, while maintaining maximum visibility onto the file and package structure, are provided on our website. Authors with large packages, or packages with more than 1,000 files, should contact the AEA Data Editor. Depositing at other trusted repositories is one option, described in the next section.

Since the migration to the AEA Data and Code Repository, these replication packages have been viewed 3.6 million times and downloaded nearly 4 hundred thousand times.

C. Third-party repositories

The data and code availability policy (DCAP) allows for code and data to be deposited at other trusted repositories, as long as all other elements of the DCAP are complied with. In

[5] https://codeocean.com

[6] https://github.com/AEADataEditor/

[7] Generic images are at hub.docker.com/u/dataeditors, specific images at hub.docker.com/u/aeadataeditor.

Table 4—: Deposit statistics

Repository	Published	Downloads	Views	Uploads	Files	Size (GB)
ICPSR	4,627	412,790	3,580,827	495	54,864	1,044.74
Zenodo	11	8,208	1,504	3	6,165	373.09

Note: Unit of observation are deposits at the named repository. Columns 1-3 are for all currently published deposits as of 2023-11-28. Columns 4-6 are for deposits made between 2022-12-01 and 2023-11-28. The number of uploads may not correspond to the number of manuscripts processed by the Data Editor team. Not all uploads have been published yet.

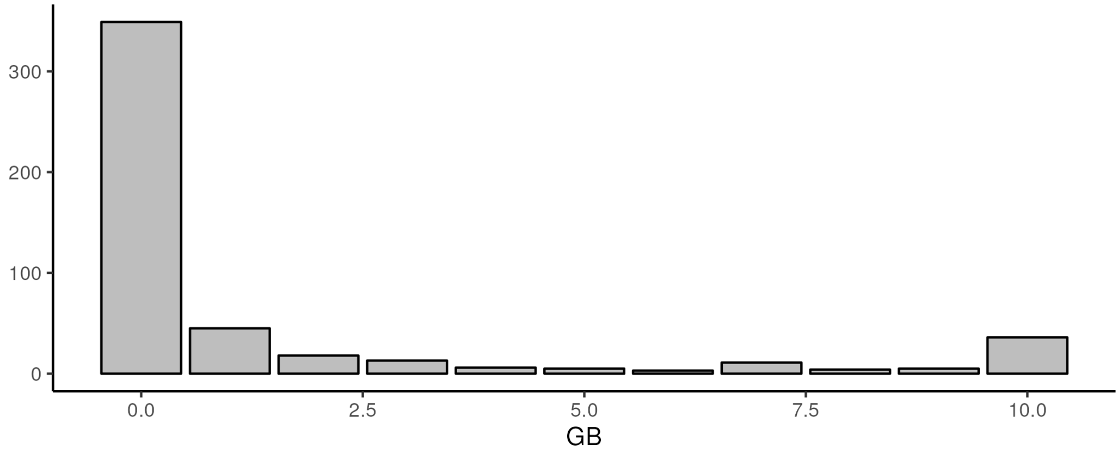

Figure 1. : Size distribution of replication packages deposited at openICPSR between 2022-12-01 to 2023-11-30, top-coded at 10GB.

fact, authors are *discouraged* from duplicating deposits they have made elsewhere. This is intended to allow authors to create replication packages prior to submitting at the AEA's journals, or any other journal, as a component of a reproducible workflow and possibly in compliance with funder data management policies. Examples of other general purpose repositories include the Harvard Dataverse[8] and Zenodo.[9]. Authors depositing on Zenodo can request inclusion in the "AEA community" at zenodo.org/communities/aeajournals/. For example, we work with authors to deposit partial data packages on Zenodo when the size of the data files surpasses 30 GB. Some of these are deposited on Zenodo, where the "AEA community" is available. Table 4 shows a small number of very large packages on Zenodo, with 373 GB of data deposited in only 11 packages.

More specific repositories that only allow specific authors to deposit (but anybody to obtain data from), such as the Swedish National Data

[8] https://dataverse.harvard.edu/
[9] https://zenodo.org/

Service.[10] or the new World Bank Reproducible Research Repository,[11] have been used more recently for partial or complete replication packages. In many of these cases, the Data Editor has actively assisted authors in preparing data archives, and shared tools that make such data publication easier (see also next section).[12] Third-party repositories are linked to the main AEA Data and Code Repository deposit, and are cited in the main article when appropriate. Authors wishing to deposit replication packages early in the research lifecycle are encouraged to consult the Social Science Data Editors website,[13] where links to trusted repositories are provided.

II. Providing Support for Compliance with AEA Data and Code Availability Policy

A. *Working with authors*

Since the introduction of the AEA's strengthened data and code availability policy in 2019 (American Economic Association, 2020), we have monitored how authors work to comply with the policy upon first submission of their packages. To help authors, we have published and continually update guidance available at aeadataeditor.github.io. The template README (Vilhuber et al., 2022*b*), which we published with several other economics data editors, helps authors compile all the information required for complete documentation of their data and code deposit.[14] Guidance on specific topics is published in the form of blog posts.[15]

We note that authors can be compliant with the policy without providing a copy of data used, as long as the reason for the inability to provide the data is acceptable, correct, and documented as part of the replication materials.

Compliance with the policy has been excellent. In some cases, we have requested data that was not initially provided, when such data could be legally and ethically provided; by the second round of assessments, compliance was generally achieved (see also our discussion of outreach to data providers).

Occasionally, a manuscript may be published while the replication package is still being updated. This leads to (temporary) non-compliance. Non-compliance may arise for technical reasons, such as when a file becomes corrupted in the upload process, preventing the deposit from being released. This year saw a week-long outage at ICPSR that lead to some temporary disruption in the processing of replication packages. In other instances, researchers have notified the Data Editor of a particular aspect of non-compliance - a file may be missing, a dataset may be able to be published that was not initially provided, etc. As of 2023-11-30, 4 packages were tagged as non-compliant. All cases are eventually resolved, either through pre-publication amendments, or post-publication updates.

At the time of publication, a manuscript is linked with one (or more) archived replication packages, constituting the *version of record* for the replication package. Occasionally, issues are brought to our attention by authors, readers, and data providers. Authors may have a better README, readers might have noticed a missing code or data file, or data providers might ask for a dataset to be removed because it infringes on terms of use agreed to by the author. The supplemental "*Policy on Revisions of Data and Code Deposits in the AEA Data and Code Reposi-*

[10] https://snd.gu.se/en

[11] https://reproducibility.worldbank.org/

[12] Code to support uploading large quantities of data to Zenodo via the Zenodo API, originally created by LDI Lab Member Vansh Gupta, can be found at github.com/AEADataEditor/Upload-to-Zenodo.

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

[14] The README is available at social-science-data-editors.github.io/template_README/.

[15] Blog posts by the AEA Data Editor can be found at aeadataeditor.github.io/year-archive/.

tory"[16] specifies which modifications constitute a minor edit to the version of record, and which modifications lead to a higher version number, without modification of the existing version of record. In particular, any change that potentially changes a computational result or adds (untested) code will lead to a new version of the deposit being created, without changing or removing the version of record, even if the modifications fixes an error. However, the presence of replication packages that are newer than the version of record is signalled to readers via a banner, and is recorded in the metadata.

We identified 17 actions regarding post-publication modifications in 2023. Table 5 identifies who initiated the updates. Several updates were initiated following a *Replication Game* (see Section III.D for details).

Table 5—: Updates

Origin	Manuscripts
Author	7
Data Editor	2
Faculty	2
Other	1
Researcher	4
Student	3

Note: Unit of observation are manuscripts assessed between 2022-12-01 and 2023-11-30. A total of 17 had updates; multiple origins of the information may have been identified.

B. Intellectual Property and Licenses

Authors retain copyright for any data and code deposited by them in the AEA Data and Code Repository, unless that copyright belongs to others and the authors have a license to republish it. The default license for all repositories based at openICPSR is the Creative Commons Attribution (CC-BY) (Creative Commons, 2017), but authors can choose their own license. All licenses are vetted by the Data Editor for compliance with the DCAP. We encourage authors to consult our licensing guidance.[17]

In some cases, authors may wish to publish data under more restrictive licenses or conditions, due mostly to ethical concerns, while ensuring that replication remains possible.[18] In the past, we have been able to leverage multiple mechanisms in place at the openICPSR repository hosting the AEA Data and Code Repository; however, some of those mechanisms are no longer available. Authors who wish to explore ways to make their data ethically accessible should contact the Data Editor early enough in the submission process.

The AEA replicators will sometimes access confidential or proprietary data for the purpose of verifying computational reproducibility (see Section I.A), as provided by the authors, or directly requested from the data providers via application or subscription services. Such data are not published as part of authors' replication packages. However, we do encourage authors to seek permission to share such data, where possible, and encourage data providers to allow for publication of extracts of their data, sufficient to support future reproducibility efforts. Table 6 shows the number and type of agreements we entered into for the 66 formal or informal agreements in 2023.

[16] See Appendix A for a list of links to all supplemental policies.

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

[18] Examples from past years include Deryugina, Shurchkov and Stearns (2021*b*) and Goncalves and Mello (2021*b*), which accompany Deryugina, Shurchkov and Stearns (2021*a*) and Goncalves and Mello (2021*a*), respectively.

Table 6—: NDAs and DUAs

Type	Manuscripts
Data Use Agreement	1
NDA (formal)	3
NDA (informal)	62

Note: Unit of observation are manuscripts assessed between 2022-12-01 and 2023-11-30.

We continue to assist authors in remaining compliant with data use agreements and copyright law, to the extent possible, but authors should be aware of their potential liability in the cases of infringements. Since the summer of 2022, all authors are required to attest, via the published README, that they have "*legitimate access to and permission to use the data used*" in the manuscript, and that they also have "*documented permission to redistribute [publish] the data*" (Vilhuber et al., 2022*b*, pg.1). Unintentional posting of data that authors do not have permission to publish is one of the causes for replication packages being withdrawn, and thus becoming non-compliant until remediation and publication of an update.

C. Direct outreach

In order to reach authors and researchers with immediate or prospective questions, the Data Editor has presented and given workshops at McMaster University, Banco de Portugal, at a Symposium on Open Science in Berlin, Université du Québec à Montréal, University of Tokyo, University of Osaka, and University of Virginia, as well as presentations or keynotes at the meetings of the Royal Economic Society, Midwest Economics Association, the Western Economic Association, the European Economic Association, and the Japanese Economic Association. He also met with students and faculty through informal in-person meetings and consultations at the ASSA meetings 2023, UC Berkeley, Stanford, University of Toronto, Humboldt University Berlin, and University of Tokyo.

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

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

A. AEA RCT Registry

The AEA RCT Registry (Registry) provides services to the economics (and social science) community at large. Managed at J-PAL and funded by the AEA, registration at the Registry is mandatory for field experiments published in American Economic Review (AER), American Economic Review: Insights (AERI), American Economic Journal: Applied Economics (AEJAPP), and American Economic Journal: Economic Policy (AEJPOL), but is also used more broadly in the economics discipline, with numerous publications in top economics journals as well as field journals identifying preregistration in the Registry. The Registry is available at www.socialscienceregistry.org.[19] In collaboration with members of the AEA's *Oversight Committee for Registry of Random Controlled Trials*, the Registry team at J-PAL have continued to work on improving the usability of the Registry, as well as ensuring the availability of Registry data.

Data for the Registry is being curated and preserved in the *AEA RCT Registry Dataverse* at dataverse.harvard.edu/dataverse/aearegistry, al-

[19] This section is based in part on information provided by Stuti Goyal, Jack Cavanagh, and Sarah Kopper. All errors are mine.

lowing for reproducible analysis of the universe of registrations by researchers.[20]

Usage of the registration continues to increase strongly. As Figures 2 and 3 show, the rate of registrations per year continues to rise, and the registry now has over 8100 registrations, which are being added at a rate of 1422 per year. More than 12400 unique researchers are associated with these registrations (left panel, Figure 3), 3040 of which are associated with registrations that have been active in 2023. The share of pre-registrations, favored by some, surpassed post-registrations for the first time in 2021, and remains higher (right panel, Figure 3).

As the registry has continued to grow, attract new users, and become a standard for experiments across economics and related social sciences, it is improving in its function as a central database of those experiments. In the last five years, an increasing number of articles use the publicly available registry metadata as either their main analysis data or as a supplement to it. These can be grouped into a few broad sets. One group of papers use the registry to study research transparency questions in the social sciences, either studying the registry as a transparency object in itself (Christensen and Miguel, 2018; Abrams, Libgober and List, 2020; Miguel, 2021), taking it as a means to glean information on other transparency behaviors (Ofosu and Posner, 2019; Laitin et al., 2021), or using it as an auditing tool, both for self-reported data (Christensen et al., 2019) and the implementation of transparency policies (Buckley et al., 2022). A second group use the registry data as a significant portion of a larger compiled dataset attempting to proxy for the universe of experiments and experimenters in a particular subset of social science research (Corduneanu-Huci, Dorsch and Maarek, 2021, 2022). A last group of studies use the data to answer meta-scientific questions about the studies on the registry (Leight, Asri and Imai, 2022; Murtagh-White, Wall and O'Sullivan, 2023).

B. *Data Providers*

We regularly meet and communicate with academic, governmental, and commercial data providers, on behalf of specific authors or because we have identified a data provider as a frequently used resource. Discussion topics include making data citations easier, clarifying licenses, requesting blanket or streamlined data redistribution or access authorizations, or suggesting improved data curation practices to avoid repeatedly copying data from uncurated websites to the curated AEA Data and Code Repository.

In the course of the past year, we have talked about some or all of these topics with IPUMS, the World Bank (see also below), staff at the U.S. Census Bureau, the Internal Revenue Service (IRS), the Bureau of Labor Statistics (BLS), the (German) Institute for Employment Research (IAB), the Banco de Portugal Microdata Research Laboratory (BPLIM), and the French *Centre d'accès sécurisé de données* (CASD). We have also talked to various research groups on how to improve data curation, visibility, and citability of the data created by their efforts.

C. *Economics Journals*

We continue to coordinate with other data editors conducting similar activities at other journals. An informal mailing list managed by the AEA Data Editor is used to interact with others.[21] Mailing list members who wish to be more actively involved can join the **monthly video call**, and participate in the development of the website of the Social Science Data Editors.[22] In addition to the Data Editor of the

[20] Data shown in this report are derived from AEA RCT Registry (2024), and show data for calendar year 2023.

[21] Journal editors are encouraged to join the mailing list by contacting the AEA Data Editor.

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

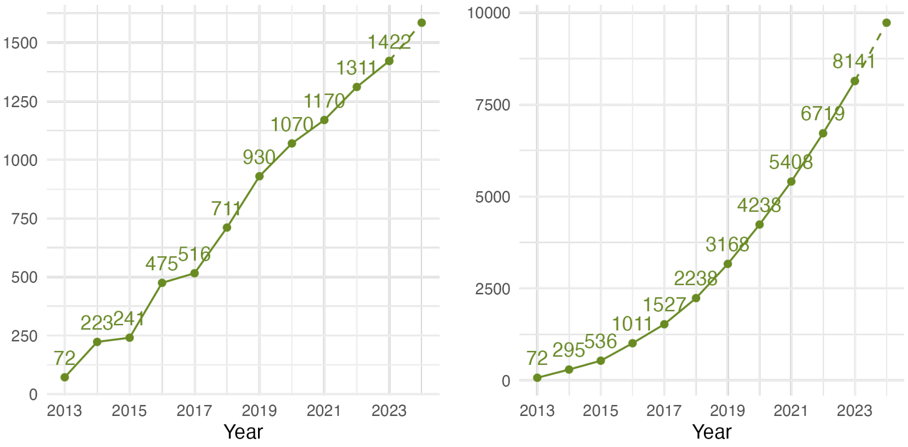

Figure 2. : Annual (left) and Cumulative Registrations (right)

Source: AEA RCT Registry (2024). Dashed lines are extrapolated increases for 2024.

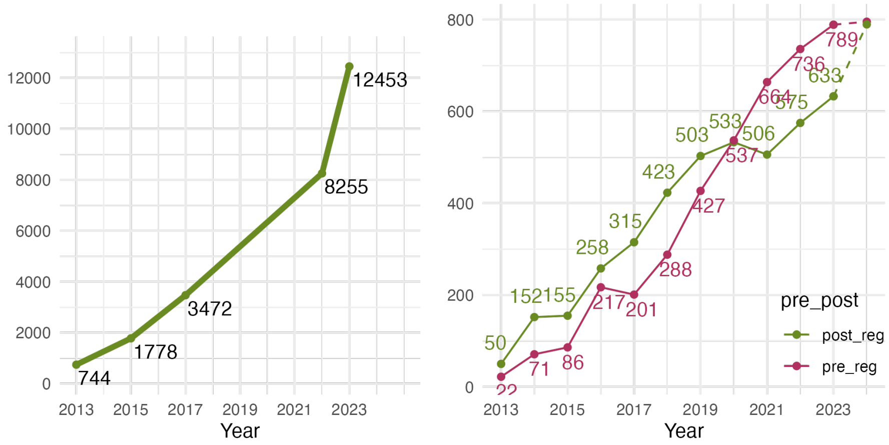

Figure 3. : Unique registered investigators (left), Post vs Pre-registrations (right)

Source: AEA RCT Registry (2024). Dashed lines are extrapolated increases for 2024.

AEA, the current group includes Miklós Koren (Data Editor, Review of Economic Studies), Florian Oswald (Data Editor, Economic Journal and Econometric Journal), Joan Llull (Data Editor, Econometrica), Marie Connolly (Data Editor, Canadian Journal of Economics), Anna Dreber Almenberg (Editor, JPE Microeconomics), and Maia Güell (Data Editor, Journal of the European Economic Association). The website contains guidance on data citations and data availability statements, best practices for coding and data preparation, and links to various tools useful to replicators. In particular, the group coordinates the *Template README* (Vilhuber et al., 2022*b*), which was last updated in November 2022, and which is referenced as a *common* reference implementation for providing required information across a number of journals. Furthermore, the group has developed a common standard for replication packages (Data and Code Availability Standard (DCAS)), which journals can use to signal that their requirements are similar to those commonly required by all endorsers of the standard (Koren et al., 2022).[23] Joint presentations and tutorials by the group involving the AEA Data Editor include the aforementioned presentations and workshops in Montréal (with Connolly), at the RES (with Connolly) and the EEA (with Koren). The AEA Data Editor also coordinates with the Editor and Data Editor of Economic Inquiry, and has had discussions with other editors and data editors seeking in put on their practices and policies. He furthermore participates in the Steering Committee of a group of data repository leaders at Data-PASS organized as the Journal Editors Discussion Interface (JEDI).[24]

D. *Third-party verification services*

We continue to rely on and have discussions with third-party verification services. As noted earlier, 25 reports were provided by external replicators or replication services (see Table 2 for statistics by journal, and Appendix VI.A for a list of third-party replicators). Of note, the World Bank has introduced a reproducibility verification service,[25] which "[verifies packages] to ensure that they are complete and fully functional before they are published to the repository." We will accept such third-party verifications, both at the time of submission as well as upon request, as long as the process by which they are created is documented and consistent with the approach that the economics journals are using. We are working on a mechanism to robustly "certify" such processes.[26]

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

Outreach through presentations and publicly available tools is a key component of an effective data and code availability policy. Recordings of presentations (when available) and presentation materials are listed at the Data Editor's website.[27] In addition to presentations, the Data Editor occasionally is an observer at Replication Games,[28] where teams of researchers and students attempt to reproduce and extend papers published in leading economics journals. Such post-publication verifications are a necessary and useful check on published replication packages. Replicators may find issues that were not discovered, or discoverable, by prepublication replicators such as the AEA Data Editor team. The Data Editor facilitates and

[23] After the 2024 Annual Meetings, but before final copyediting for this report, the AEA's data and code availability policy (DCAP) was updated to conform to the order, format, and content of the DCAS, see aeaweb.org/journals/data/data-code-policy. The update does not change any requirements. The previous version of the policy can be found at aeaweb.org/journals/data/archive/2020-2024.

[24] https://dpjedi.org/

[25] https://reproducibility.worldbank.org/index.php/about

[26] https://transparency-certified.github.io/

[27] https://aeadataeditor.github.io/talks/

[28] https://i4replication.org/description.html

monitors that any corrections or suggested improvements are conveyed to authors, and are reflected on replication packages via the AEA's *Policy on Revisions of Data and Code Deposits* (see Appendix A). Several such (minor) corrections have lead to updates (see Table 5).

A. *Resources*

The AEA Data Editor maintains public resources available to the economics community. These are made available through a dedicated website at aeadataeditor.github.io/ and code and project templates provided at github.com/aeadataeditor. In particular:

- Step-by-step guidance on how to prepare a replication package is provided at aeadataeditor.github.io/aea-de-guidance/, including video tutorials and a description of the process.
- The template README (Vilhuber et al., 2022*b*) is referenced as part of the guidance, and separately accessible at social-science-data-editors.github.io/template_README/.
- Various blog posts on topics relating to computational reproducibility are posted at aeadataeditor.github.io/year-archive/, and may be summarized on Twitter under the Data Editor's handle @AEAData, on Mastodon under @aea-data@mstdn.social, and on BlueSky under @aeadata.bsky.social
- Instructions to replicators for assessing authors' replication packages are provided at github.com/AEADataEditor/replication-template.
- Template code for using containers for Stata, R, Julia, and Gurobi can be found by searching for "docker" on the Github site.

V. Replication team at Cornell University

A. *Replicators*

The following **45** students have provided excellent assistance in reproducing the results from the 527 articles processed by the Replication Lab: Adam J. Faridi, Akshay Yadava, Alice Wei, Amie Li, Ananya Bakshi, Andrew Phiri, Andrew Wallace, Anjini Khanna, Anurag Tiwari, Arnaav Sareen, Bianca Jimenez, Caitlin Song, Crystal Lim, Elian Gomez, Ethan Carlson, Gary Wu, Hawi Tolera, Ilona Khimey, Jade Yang, Jaeyoung Shim, Jason Lan, Jessica Rizzo, Joshua Wallace, Kareena Stowers, Kate Chanpong, Kayla Yang, Kirin Eicher, Kristine Li, Leslie Geng, Lincy Chen, Luke Trautwein, Manvir Chahal, Melanie Brown, Micere Mugweru, Miranda Zhou, Nguyen Vo, Olivia Liu, Phalguni Miraj, Sherry Li, Siddhi Malvankar, Sohit Gurung, Talia Boehm, Tommy Wang, Vidya Balaji, Yuchang Tian. Graduate students Leonel Borja Plaza and Linda Wang and Research Aide Sofia Encarnación (all Cornell University) have been invaluable assistants in training and coordinating the work as well as developing the methods and procedures which we have made public. Linda Wang contributed programming to this report. Sofia Encarnación contributed to all parts of the process. A description and evaluation of the training of replicators for the AEA's Data Editor team was published as Vilhuber et al. (2022*a*). The training and reference manual can be found at online.[29]

B. *Computing support*

We thank the Economics Department and the ILR School for providing us with computing resources at the Cornell Center for Social Sciences and the Bioinformatics cluster.

[29] https://labordynamicsinstitute.github.io/ldilab-manual/

VI. Third-party contributors

A. Replicators

We are grateful to the third-party replicators who assisted us with verifications when we were unable to access data or, in some cases, computing resources. Graduate research assistants at Stanford, Penn State University, and at Cornell helped us. We thank Paulo Guimarães and staff at BPLIM who contributed their time to run code on confidential data and provide us with detailed knowledge about the data being used. We in particular want to again thank Olivier Akmansoy, Christophe Hurlin (Université d'Orléans), and Christophe Pérignon (HEC Paris), all of cascad, a certification agency for scientific code and data, who have been generous of their time and resources, and have provided us with multiple reports during this time.

We do not name the authors with whom we signed non-disclosure agreements, or who otherwise provided us with access to data that could not be published. We are grateful for their flexibility and patience.

B. Computing resources

We are grateful to Codeocean, NBER, WholeTale, and Harvard Business School, who all provided us with access to computing resources at no cost, and technical assistance when necessary. We use free academic resources on Github and Bitbucket. WholeTale is free to use for any academic user.

VII. Disclosures

We received a generous compute and storage quota from Codeocean, a free license to use Stata 17/18 for one year in cloud applications from Stata, and a subaward on NSF grant 1541450 "CC*DNI DIBBS: Merging Science and Cyberinfrastructure Pathways: The Whole Tale" from the University of Illinois to evaluate the WholeTale platform for the purpose of reproducibility verification. None of the sponsors have reviewed this preliminary assessment, or have had influence on any of the conclusions of this document. Codeocean currently offers academic users a certain number of monthly free compute hours. WholeTale is free to use.

VIII. Data and Code Availability Statement

All publicly available data and code used to generate figures and tables in this article are available (Vilhuber, 2024*a*,*b*). Some detailed data from the editorial system, used for Table 3, are considered confidential and cannot be made available in a way that preserves the privacy of the editorial process at this time.

LARS VILHUBER, *Data Editor*
JACK CAVANAGH, *Manager, Research Transparency, J-PAL/MIT*

REFERENCES

Abrams, Eliot, Jonathan Libgober, and John List. 2020. "Research Registries: Facts, Myths, and Possible Improvements." National Bureau of Economic Research w27250, https://doi.org/10.3386/w27250.

AEA Data Editor. 2021. "Use of Docker for Reproducibility in Economics." https://aeadataeditor.github.io/posts/2021-11-16-docker (accessed 2022-01-03).

AEA RCT Registry. 2024. "Registrations in the AEA RCT Registry (2013-05-15 through 2024-02-01)." Harvard Dataverse [data], https://doi.org/10.7910/DVN/2RZF2X.

American Economic Association. 2020. "Data and Code Availability Policy." *AEA Papers and Proceedings*, 110: 776–78. https://doi.org/10.1257/pandp.110.776.

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*. https://doi.org/10.1016/j.future.2017.12.029.

Buckley, Pamela R., Charles R. Ebersole, Christine M. Steeger, Laura E. Michaelson, Karl G. Hill, and Frances Gardner. 2022. "The Role of Clearinghouses in Promoting Transparent Research: A Methodological Study of Transparency Practices for Preventive Interventions." *Prevention Science*, 23(5): 787–798. https://doi.org/10.1007/s11121-021-01252-5.

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature*, 56(3): 920–980. https://doi.org/10.1257/jel.20171350.

Christensen, Garret, Zenan Wang, Elizabeth Levy Paluck, Nicholas Swanson, David J. Birke, Edward Miguel, and Rebecca Littman. 2019. "Open Science Practices are on the Rise: The State of Social Science (3S) Survey." https://doi.org/10.31222/osf.io/5rksu.

Clyburne-Sherin, April, Xu Fei, and Seth Ariel Green. 2019. "Computational Reproducibility via Containers in Psychology." *Meta-Psychology*, 3. https://doi.org/10.15626/MP.2018.892.

Corduneanu-Huci, Cristina, Michael T. Dorsch, and Paul Maarek. 2021. "The politics of experimentation: Political competition and randomized controlled trials." *Journal of Comparative Economics*, 49(1): 1–21. https://doi.org/10.1016/j.jce.2020.09.002.

Corduneanu-Huci, Cristina, Michael T. Dorsch, and Paul Maarek. 2022. "What, Where, Who, and Why? An Empirical Investigation of Positionality in Political Science Field Experiments." *PS: Political Science & Politics*, 55(4): 741–748. https://doi.org/10.1017/S104909652200066X.

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

Deryugina, Tatyana, Olga Shurchkov, and Jenna Stearns. 2021*a*. "COVID-19 Disruptions Disproportionately Affect Female Academics." *AEA Papers and Proceedings*, 111: 164–168. https://doi.org/10.1257/pandp.20211017.

Deryugina, Tatyana, Olga Shurchkov, and Jenna Stearns. 2021*b*. "Data for: COVID-19 Disruptions Disproportionately Affect Female Academics." American Economic Association [publisher] Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E139263V1.

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).

Goncalves, Felipe, and Steven Mello. 2021*a*. "A Few Bad Apples? Racial Bias in Policing." *American Economic Review*, 111(5): 1406–1441. https://doi.org/10.1257/aer.20181607.

Goncalves, Felipe, and Steven Mello. 2021*b*. "Supplementary Data for: A Few Bad Apples? Racial Bias in Policing." American Economic Association [publisher] Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E123921V1.

Koren, Miklós, Marie Connolly, Joan Lull, and Lars Vilhuber. 2022. "Data and Code Availability Standard." Zenodo v1, https://doi.org/10.5281/zenodo.7436134.

Laitin, David D., Edward Miguel, Ala' Alrababa'h, Aleksandar Bogdanoski, Sean Grant, Katherine Hoeberling, Cecilia Hyunjung Mo, Don A. Moore, Simine Vazire, Jeremy Weinstein, and Scott Williamson. 2021. "Reporting all results efficiently: A RARE proposal to open up the file drawer." *Proceedings of the National Academy of Sciences*, 118(52). https://doi.org/10.1073/pnas.2106178118.

Leight, Jessica, Viola Asri, and Taisuke Imai. 2022. "Publication Bias in Randomized Controlled Trials: Evidence from the American Economic Association Registry." https://doi.org/10.17605/OSF.IO/WSXD9.

Miguel, Edward. 2021. "Evidence on Research Transparency in Economics." *Journal of Economic Perspectives*, 35(3): 193–214. https://doi.org/10.1257/jep.35.3.193.

Murtagh-White, Matt, P. J. Wall, and Declan O'Sullivan. 2023. "Learning from the Evidence: Impact Evaluations, Ontology and Policy." 104–106. https://doi.org/10.1109/ICSC56153.2023.00023.

Ofosu, George, and Daniel N. Posner. 2019. "Pre-analysis Plans: A Stocktaking." Berkeley Initiative for Transparency in the Social Sciences (BITSS) MetaArXiv Preprints e4pum. https://doi.org/10.31222/osf.io/e4pum (accessed 2020-06-15).

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. 2024*a*. "Code and Data for: Report for 2023 by the AEA Data Editor." American Economic Association [publisher], https://doi.org/10.3886/E198444V1.

Vilhuber, Lars. 2024*b*. "Process data for the AEA Pre-publication Verification Service." American Economic Association [publisher] V5, https://doi.org/10.3886/E117876V5.

Vilhuber, Lars, Hyuk Harry Son, Meredith Welch, David N. Wasser, and Michael Darisse. 2022*a*. "Teaching for large-scale Reproducibility Verification." *Journal of Statistics and Data Science Education*, 30(3): 274–281. https://doi.org/10.1080/26939169.2022.2074582.

Vilhuber, Lars, Marie Connolly, Miklós Koren, Joan Llull, and Peter Morrow. 2022*b*. "A template README for social science replication packages." https://doi.org/10.5281/zenodo.7293838.

LIST OF POLICIES

All policies are listed at https://www.aeaweb.org/journals/data.
Main policy: https://www.aeaweb.org/journals/data/data-code-policy
Supplemental policies:

- The *Policy for Papers Conducting Experiments and Collecting Primary Data* (aeaweb.org/journals/data/policy-experimental) applies to laboratory experiments, field experiments, and any surveys, and specifies what should be provided in addition to the main policy.
- The *Policy and Protocol for Third-Party Verifications* (aeaweb.org/journals/data/policy-third-party specifies how we send out or accept verifications conducted by parties other than the Data Editor team.
- The *Policy on Revisions of Data and Code Deposits in the AEA Data and Code Repository* (aeaweb.org/journals/data/policy-revisions) applies when it becomes necessary to revise published replication packages. This may happen when readers and researchers contact the Data Editor, when authors identify a problem, correction, or improvement themselves, or when it becomes necessary for whatever reason to modify the replication package. The policy also defines under what conditions the version of record is modified, or whether the updated deposit simply becomes connected to, but does not replace the version of record.