2023 Report by the Data Editor

This repository contains the code, data, and manuscript files for the 2023 report by the AEA Data Editor. If you are reading this on openICPSR, then only code and data are present.

Statement about rights

Raw data come from the JIRA system used by the AEA Data Editor and cannot be made available outside of the organization, as it contains names of replicators, manuscript numbers, and verbatim email correspondence. Anonymized data is publicly available at Vilhuber (2024).

- ☑ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☑ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permissions are documented in the `LICENSE.txt` file.

Locations

The repository at https://github.com/AEADataEditor/report-aea-data-editor-2023 contains text, code, data, and output from running the code.

The deposit at http://doi.org/10.3886/E198444V1 contains code and data, as well as output.

Citing the report

> Vilhuber, Lars. 2024. "Report by the AEA Data Editor." AEA Papers and Proceedings.

```
@article{ReportDE2024,
Author = {Vilhuber, Lars},
Title = {Report by the {AEA} Data Editor},
Journal = {AEA Papers and Proceedings},
Volume = {},
Year = {2024,
Month = {},
Pages = {},
DOI = {},
URL = {}}
```

Citing the code and data

> Vilhuber, Lars. and Linda Wang. 2024. “Code and Data for: Report for 2023 by the AEA Data Editor.” American Economic Association [publisher], http://doi.org/10.3886/E189602V1

Data

Summary of Availability

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

Data for pre-production verification

Anonymized files from the internal production system are provided in this repository, sourced from Vilhuber (2024).

```
data/jira/anon/jira.anon.RDS
data/jira/anon/README.md
```

Data on lab members' names is directly downloaded from the Github repository associated with Vilhuber (2023), see `programs/config.R`.

openICPSR data on deposits (ICPSR, 2023a)

The data are obtained on demand from the internal systems underlying the AEA Data and Code repository. The internal systems are accessible only to ICPSR staff, and were provided to the AEA Data Editor upon request. They are not accessible to others. The data were lightly hand-edited to account for formatting errors (double double-quotes and other issues related to the conversion from internal database representation to CSV).

For those with access to the system:

> Go to https://www.openicpsr.org/openicpsr/tenant/openicpsr/module/aea/reports and download the CSV file. Save it with a date-stamp added.

Data were extracted on all published replication packages. The raw data are not provided.

```
data/icpsr/utilizationReport-2023-12-06.csv
```

The clean data files are provided in the folder `data/icpsr`. For additional use of the data, see the processing code.

```
data/icpsr/anonUtilizationReport.Rds
data/icpsr/anonUtilizationReport.csv
```

openICPSR data on deposit sizes (ICPSR, 2023b)

These data are provided by ICPSR staff upon request. They are usually not accessible to others, but could be scraped. Only aggregated statistics are computed from these files. Data file is provided.

```
data/icpsr/AEA-2023-Jan-1-through-Nov-28-2023.xlsx
```

Data on processing time (AEA, 2023)

The data on processing times were extracted from the ScholarOne manuscript management system used by the AEA. Microdata are not available (even to the author), only summary statistics are provided as Excel sheets. These were simply reformatted for the report.

```
data/scholarone/dataEditorReport_20221111-20231110.xlsx
```

Registry data (AEA RCT Registry, 2023)

The data on the J-PAL registry are from publicly available registry archives (AEA RCT Registry, 2023). Processing is via code available in `data/registry/`, which is a copy of an unpublished repository (with permission).

Zenodo data (Zenodo, 2023)

Data on AEA-related deposits in Zenodo are obtained through a call to the Zenodo API (see `01_zenodo_pull.py`) for the `aeajournals` community. Because data accessed via the API change over time, the data are provided in this repository. The scripts pulls both microdata (per deposit, `zenodo_data_YEAR.csv`) and computes summary statistics (`zenodo_data_YEAR_summary.csv`).

```
data/zenodo/zenodo_data_2023.csv
data/zenodo/zenodo_data_2023_summary.csv
```

Computational requirements

Software Requirements

The code was run with the following software versions, though others are likely to also work:

- R 4.2.3
  - Package versions set to as-of **2023-11-01**, using the Rstudio Package Manager, except for Github installed versions
  - dplyr
  - here
  - tidyr
  - tibble

    - stringr
    - readr
    - splitstackshape
    - digest
    - remotes
    - readxl
    - writexl
    - ggplot2
    - ggthemes
    - janitor
    - dataverse
    - xtable
    - github("markwestcott34/stargazer-booktabs") (overrides standard stargazer!)
- Python 3.10.12
    - requests==2.31.0
    - requests-oauthlib==1.3.1
    - requests-toolbelt==1.0.0

Packages are installed by `global-libraries.R` or defined in `requirements.txt`, and are sourced in the Dockerfile. For manual installation, the following may work (not tested).

```
R CMD BATCH global-libraries.R
pip install -r requirements.txt
```

A container was built, using the following files in this deposit:

```
Dockerfile
global-libraries.R
requirements.txt
build.sh
.myconfig.sh
```

and can be used by running `start_rstudio.sh` (for development) or `run.sh` (to simply produce all figures and tables not related to the registry). These scripts are known to work on multiple Linux workstations, and on Intel Macs. They have not been tested in a Windows/Docker environment.

All results in the report were created by running the R and Python code within the container. Running in other environments is untested.

The registry code was run in an uncontrolled environment with R, but should be runable in any R environment support `tidyverse 1.3.2` and its component packages.

Hardware Requirements

Code was last run on the following environment:

- OS: "openSUSE Leap 15.5"
- Processor: AMD Ryzen 9 3900X 12-Core Processor, 24 cores
- Memory available: 31GB memory
- Docker version 24.0.7-ce, build 311b9ff0aa93
- Docker image `aeadataeditor/report-aea-data-editor-2023:2023-12-06` built from `rocker/verse:4.2.3`

Memory requirements are minimal, and the code should run on any modern computer.

Programs

All programs, except those processing the Registry data, are in the `programs` subdirectory:

```
programs/01_lab_members.R
programs/02_zenodo_pull.py
programs/03_jira_dataprep.R
programs/04_prepare_icpsr.R
programs/05_prepare_icpsr2.R
programs/11_table1_compliance.R
programs/12_table2_stats.R
programs/13_table3_stats.R
programs/14_table4.R
programs/15_table5_webstats.R
programs/21_figure1_filesize.R
programs/99_write_nums.R
programs/config.R
programs/README.md
programs/run_all.sh
```

The Registry data was provided by J-PAL, code can be found in `data/registry/Scripts`:

```
00_functions.R
99_write_nums.R
AEA Annual Report_reproducible.Rmd
```

Running code

Each R file can be run independently (separate R sessions), in numerical order, e.g., `R CMD BATCH 02_lab_members.R`.

The Python file `01_zenodo_pull.py` can be run as `python3 01_zenodo_pull.py`.

The script `run_all.sh` is used within a (Linux) shell to implement the above run order, but is optional.

To run the registry code, `knit` the Rmd file.

Mapping tables and figures to article

Table and figure numbers in the paper do not map to program names, due to editorial decisions. The table below maps files, figures/tables, and the programs used to generate them. Some tables contain minor manual formatting edits, indicated by the suffix `_mod`.

Name of file	Figure/ Table in article	Program to create
jira_response_options_mod.tex	Table 1	13_table3_stats.R
n_journal_numbers_mod.tex	Table 2	12_table2_stats.R
n_rounds.tex	Table 3	14_table4.R
n_webstats.tex	Table 4	15_table5_webstats.R
plot_filesize_dist.png	Figure 1	21_figure1_filesize.R
n_compliance_manuscript_mod.tex	Table 5	11_table1_compliance.R
n_ndas_manuscript_mod.tex	Table 6	11_table1_compliance.R
n_updates_manuscript_mod.tex	Table 6	04_table1_compliance.R

Registry-related figures are in `data/registry/Output/`:

| Name of file | Figure/ Table in article | Program to create | | reg_pre_year_2023.png | Figure 2a | AEA Annual Report_reproducible.Rmd | | reg_cumulative_2023.png | Figure 2b | AEA Annual Report_reproducible.Rmd | | registered_users_2023.png| Figure 3a | AEA Annual Report_reproducible.Rmd | | post_pre_reg_2023.png | Figure 3b | AEA Annual Report_reproducible.Rmd |

In-text numbers are collected throughout all programs, and written out in `programs/99_write_nums.R` to `tables/latexnums.tex`.

License

See LICENSE.txt for data and code license.

References

- American Economic Association. 2023. "Aggregated processing times by journal from ScholarOne". Received by email in December 2023.
- AEA RCT Registry. 2024. "Registrations in the AEA RCT Registry (2013-05-15 through 2024-02-01)", https://doi.org/10.7910/DVN/2RZF2X, Harvard Dataverse, V1.
- ICPSR. 2023a. "Utilization Report for the AEA Data and Code Repository." ICPSR [publisher]. Accessed December 2023.
- ICPSR. 2023b. "Deposit sizes for the AEA Data and Code Repository." ICPSR [publisher]. Received by email in December 2023.

- Vilhuber, Lars. 2024. “Process data for the AEA Pre-publication Verification Service.” American Economic Association [publisher], https://doi.org/10.3886/E117876V5
- Zenodo. 2023. "Metadata on deposits in community 'aeajournals'", accessed via Zenodo API on December 6, 2023.