Fireside Chat with AEA Data Editor

Demystifying Reproducibility

Lars Vilhuber
Cornell University

The opinions expressed in this talk are solely the authors, and do not represent the views of the U.S. Census Bureau, the American Economic Association, or any of the funding agencies.

You may, however, find these opinions quite useful.

Data and Code Availability Policy

AMERICAN ECONOMIC ASSOCIATION

American Economic Review

The *American Economic Review* is a general-interest economics journal. Established in 1911, the *AER* is among the nation's oldest and most respected scholarly journals in economics.

American Economic Review: Insights

AER: Insights is designed to be a top-tier, general-interest economics journal publishing papers of the same quality and importance as those in the *AER*, but devoted to publishing papers with important insights that can be conveyed succinctly.

Journal of Economic Literature

The *Journal of Economic Literature (JEL)*, first published in 1969, is designed to help economists keep abreast of and synthesize the vast flow of literature.

Journal of Economic Perspectives

The *Journal of Economic Perspectives (JEP)* fills the gap between the general interest press and academic economics journals.

American Economic Journal: Applied Economics

American Economic Journal: Applied Economics publishes papers covering a range of topics in applied economics, with a focus on empirical microeconomic issues.

American Economic Journal: Economic Policy

American Economic Journal: Economic Policy publishes papers covering a range of topics, the common theme being the role of economic policy in economic outcomes.

American Economic Journal: Macroeconomics

American Economic Journal: Macroeconomics focuses on studies of aggregate fluctuations and growth, and the role of policy in that context.

American Economic Journal: Microeconomics

American Economic Journal: Microeconomics publishes papers focusing on microeconomic theory; industrial organization; and the microeconomic aspects of international trade, political economy, and finance.

AEA Data & Code Availability Policy (2019)

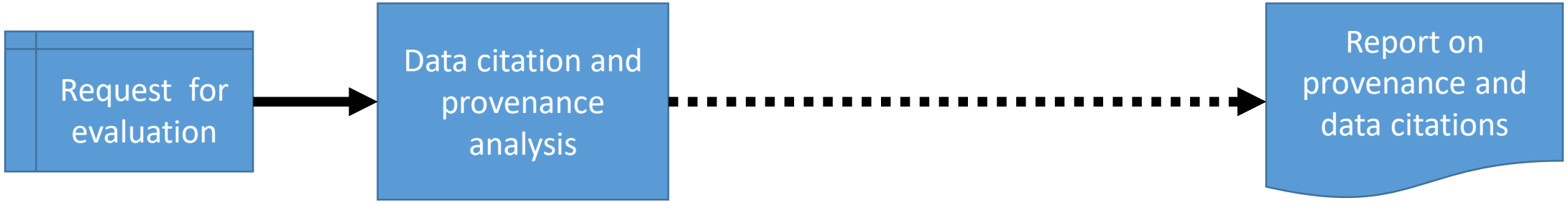
- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are **clearly and precisely** documented and **access to the data and code is clearly and precisely documented and is non-exclusive to the authors.**
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide**, **prior to acceptance**, the data, programs, and other details of the computations **sufficient to permit replication**, as well as **information about access to data and programs.**

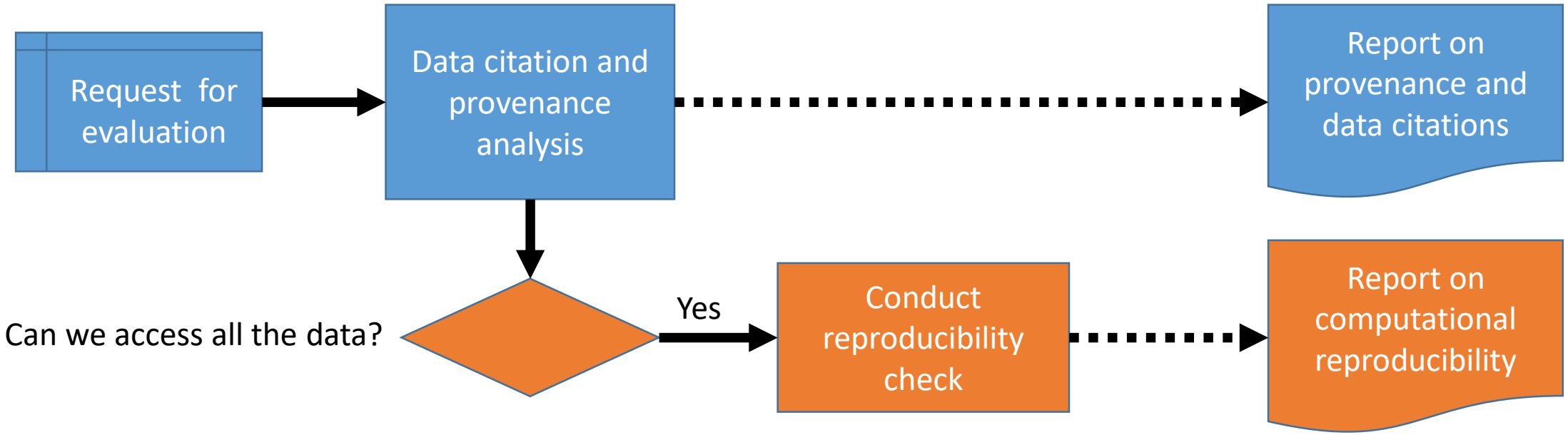
https://www.aeaweb.org/journals/data/data-code-policy

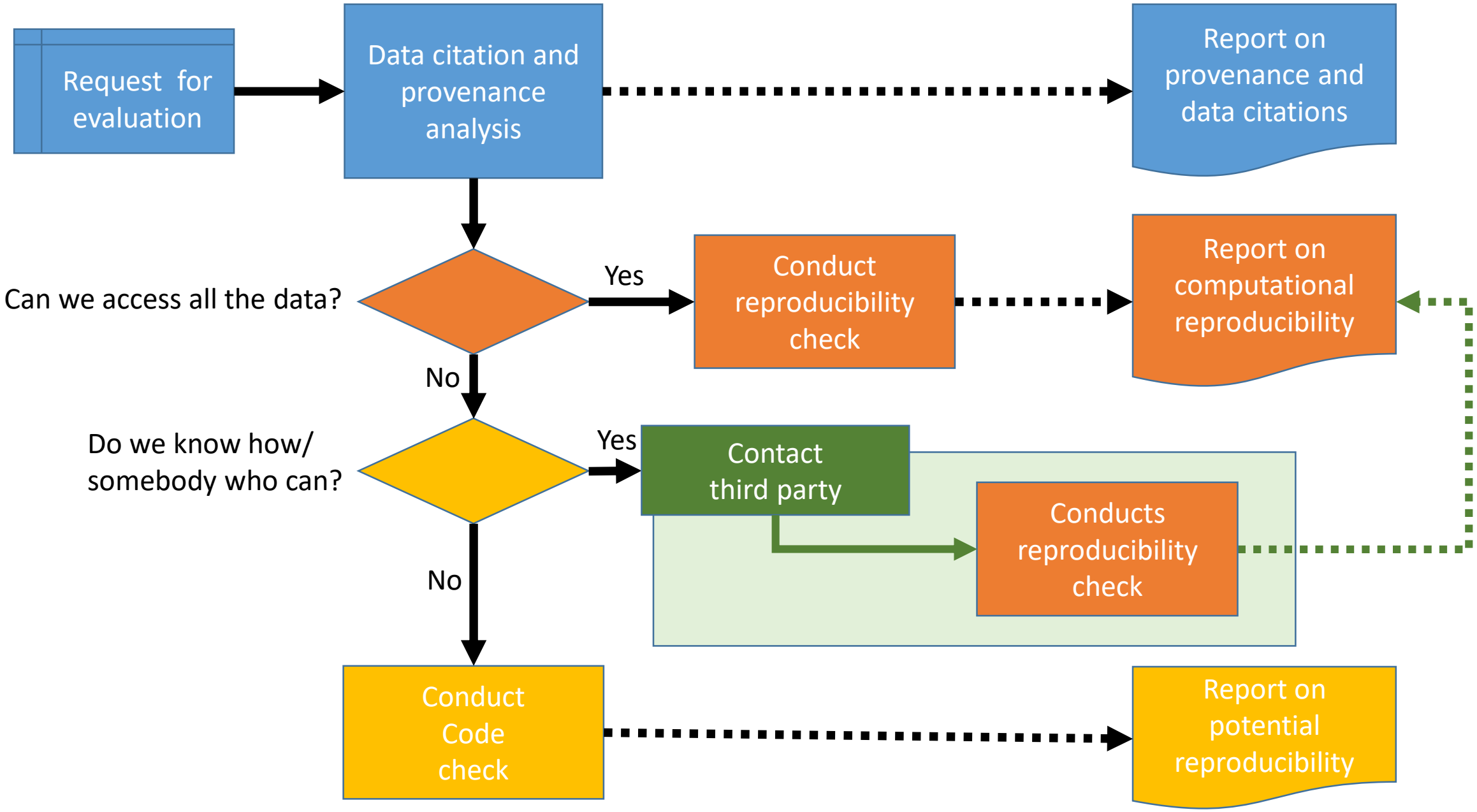
Current efforts at the AEA

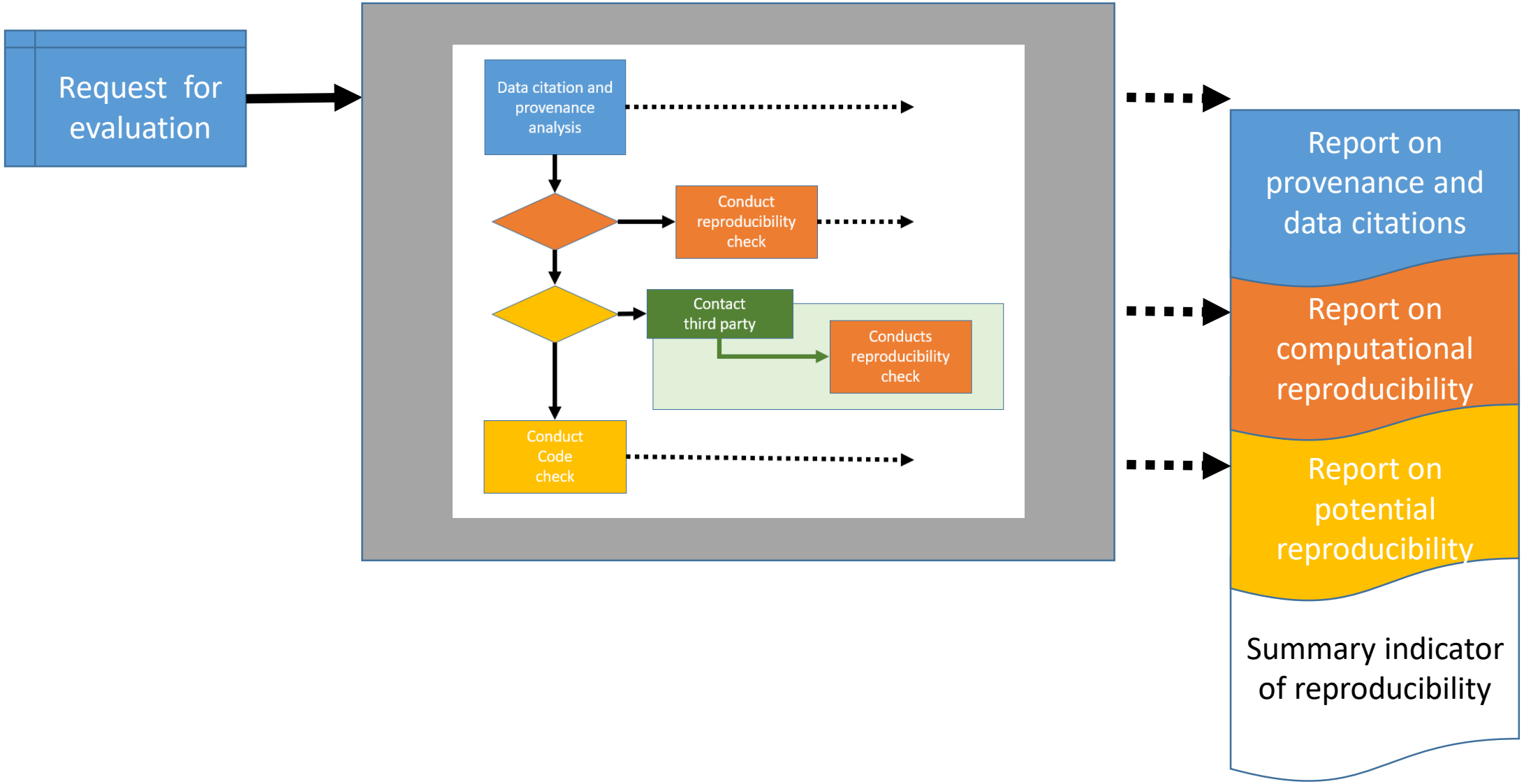
- **Pre-emptively improve code archives**
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot
- **Better archives**
  - Greater transparency of the code and data archives
- **Better provenance tracking**
    - Leave code where it is when appropriate
    - Leave data where it is almost always
    - Display that information

Reproduciblity checks: The Process

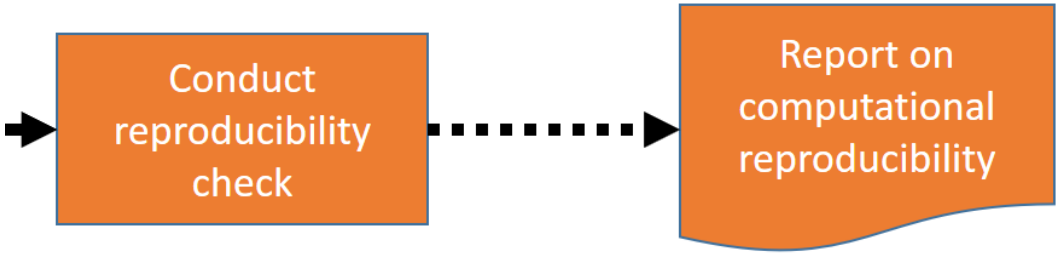

Request for evaluation
Data citation and provenance analysis
Report on provenance and data citations

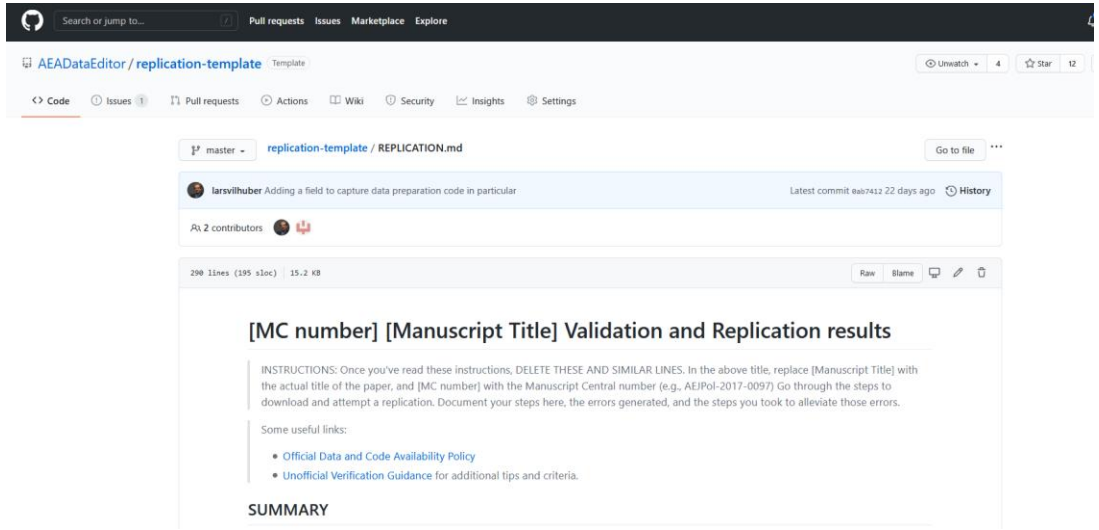

Request for evaluation
Data citation and provenance analysis
Report on provenance and data citations
Can we access all the data?
Yes
Conduct reproducibility check
Report on computational reproducibility


Request for evaluation
Data citation and provenance analysis
Report on provenance and data citations
Can we access all the data?
Yes
Conduct reproducibility check
Report on computational reproducibility
No
Do we know how/ somebody who can?
Yes
Contact third party
Conducts reproducibility check
No
Conduct Code check
Report on potential reproducibility


Request for evaluation
Data citation and provenance analysis
Conduct reproducibility check
Contact third party
Conducts reproducibility check
Conduct Code check
Report on provenance and data citations
Report on computational reproducibility
Report on potential reproducibility
Summary indicator of reproducibility

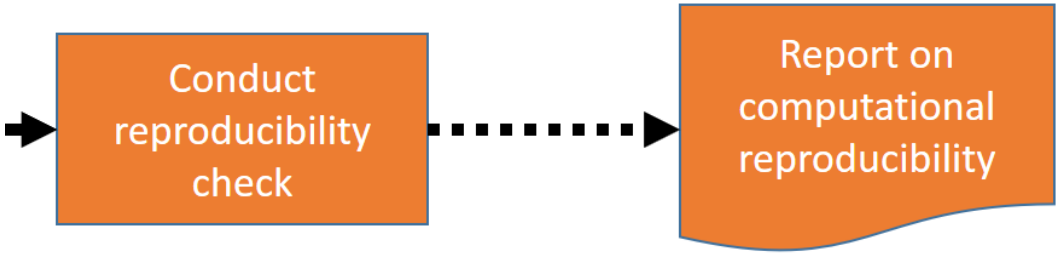
What is the reproducibility check?





Template report available at github.com/AEADataEditor/replication-template/

What is the reproducibility check?

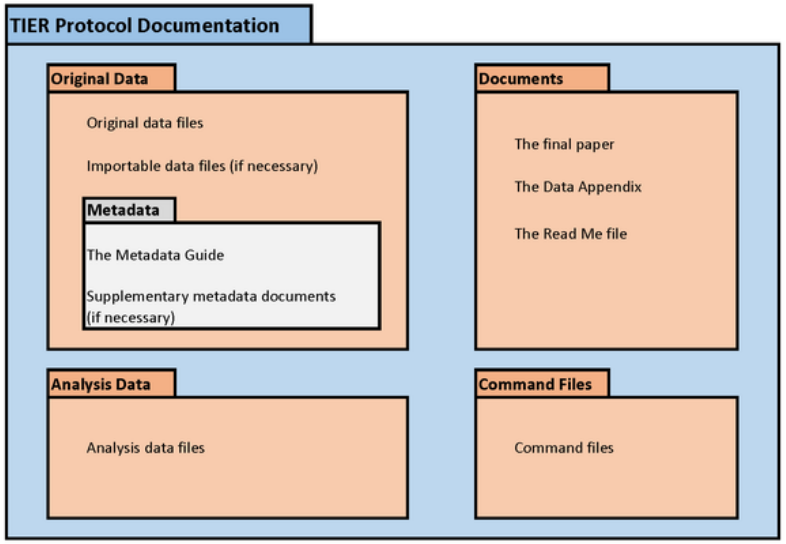


- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

Where/how to start

Basic project setup

- Structure your project
  - Data inputs
  - Data outputs
  - Code
  - Paper/text/etc.
- Version your project (git)
- Track metadata
  - Cite articles you reference
  - Cite data sources you use



https://www.projecttier.org/tier-protocol/specifications-3-0/

Computational empathy

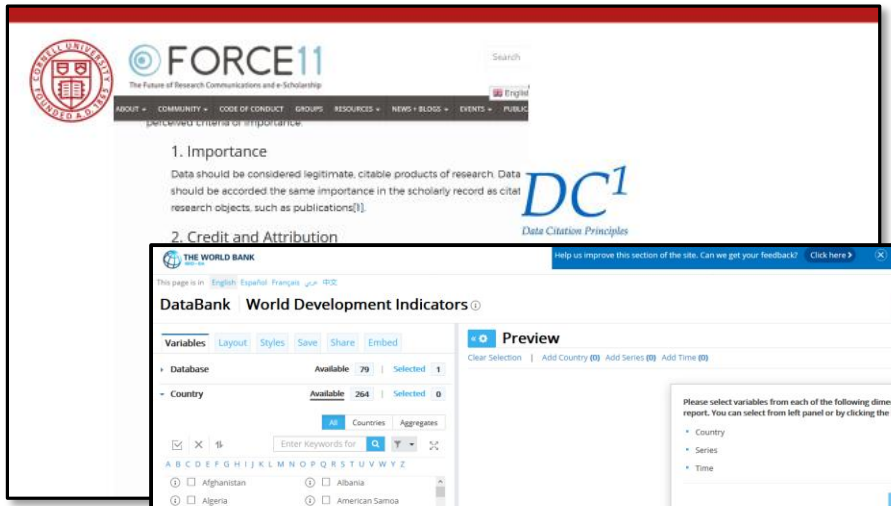
Computational empathy

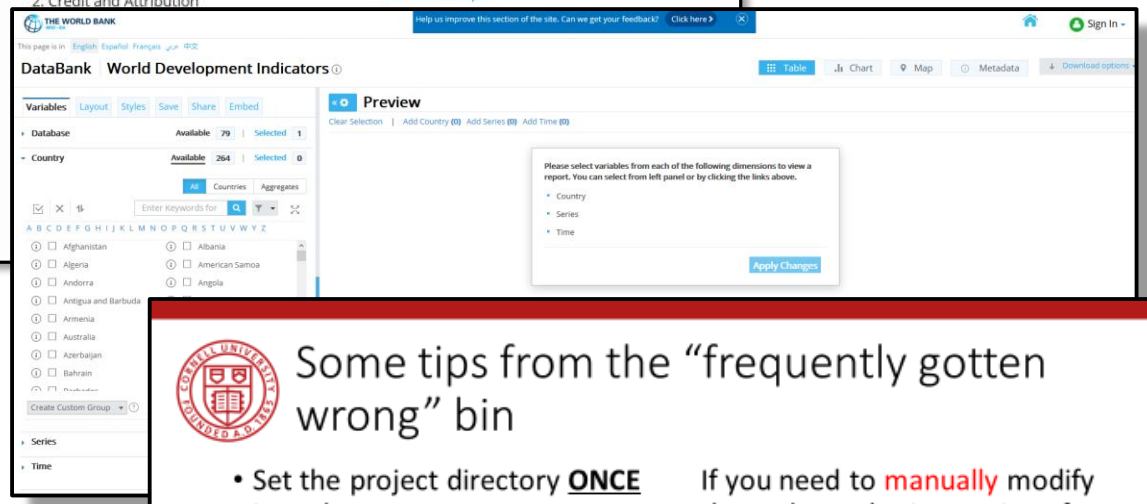
- Consider how the next person will (be able) to compute
  - You don't know what they don't know
  - Assume some frequent characteristics
    - Empirical background (2-3 yrs undergrad?)
    - Likely to know about frequently used software, but not very specific software
    - Have **none** of your add-on packages/ libraries/ etc. pre-installed
- Don't force them to do tedious things

More

Details

- Data provenance
- Data citations
- Good coding practices





Some tips from the "frequently gotten wrong" bin

- Set the project directory **ONCE** in code, or **NEVER** (Stata, R, Python)
- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)
- **Write out all tables, figures**, and in-text numbers into separate files

If you need to manually modify the code to obtain a series of tables/figures/columns, you're doing something wrong:

- Use functions, ado files, programs, macros, subroutines
- Use **loops, parameters, parameter files** to call those subroutines

Restricted-access data

Current efforts at the AEA

- **Pre-emptively improve code archives**
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot
- **Better archives**
  - Greater transparency of the code and data archives
- **Better provenance tracking**
  - Leave code where it is when appropriate
  - Leave data where it is almost always
  - Display that information

Restricted-access data pose a challenge

How do you check code when the data access is complex?

How do you improve archives when you do not control data management?

How do you document data provenance when you cannot provide the data?

AEA Data & Code Availability Policy (2019)

- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are **clearly and precisely** documented and **access to the data and code is clearly and precisely documented and is non-exclusive to the authors.**
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide**, **prior to acceptance**, the data, programs, and other details of the computations **sufficient to permit replication**, as well as **information about access to data and programs.**

Example: FSRDC

- Access can be **clearly and precisely documented**
- Is **non-exclusive to the authors**
- Intermediate files preserved

(example taken from Fort, Restud 2016)

- NOTE: for AEA, you are required to provide all programs, but a copy may/should be available within the FSRDC as well.

To reproduce the tables and figures in the paper:

1. *All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata, follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: https://www.census.gov/ces/rdcresearch/howtoapply.html.*
2. *You must request the following datasets in your proposal:*
   - *Longitudinal Business Database (LBD), 2002 and 2007*
   - *Foreign Trade Database – Import (IMP), 2002 and 2007*
   - *Annual Survey of Manufactures (ASM), including the Computer Network Use Supplement (CNUS), 1999*
   - *[...]*
   - *Annual Survey of Magical Inputs (ASMI), 2002 and 2007*
3. *Reference "Technology and Production Fragmentation: Domestic [redacted] proposal. This will give you access to the programs and input datasets required to reproduce the results. Requesting a search of archives with the articles DOI ("10.1093/restud/rdw057") should yield the same results.*

NOTE: Project-related files are available for 10 years as of 2015.

https://social-science-data-editors.github.io/guidance/DCAS_Restricted_data.html#us-census-bureau-and-fsrdc

Example: Danish administrative data

- Access can be **clearly and precisely documented**
- Is **non-exclusive to the authors**

(example taken from Fadlon and Nielsen, AEJ:Applied 2021)

The information used in the analysis combines several Danish administrative registers (as described in the paper). The data use is subject to the European Union's General Data Protection Regulation(GDPR) per new Danish regulations from May 2018. The data are physically stored on computers at Statistics Denmark and, due to security considerations, the data may not be transferred to computers outside Statistics Denmark. Researchers interested in obtaining access to the register data employed in this paper are required to submit a written application to gain approval from Statistics Denmark. The application must include a detailed description of the proposed project, its purpose, and its social contribution, as well as a description of the required datasets, variables, and analysis population. Applications can be submitted by researchers who are affiliated with Danish institutions accepted by Statistics Denmark, or by researchers outside of Denmark who collaborate with researchers affiliated with these institutions.

Health Data. To identify fatal and severe non-fatal health events we use two complementary datasets. Our first dataset is the *Death Registry* (Statistics Denmark 2020b), which includes deceased individuals' date of death. Our second dataset is the *National Patient Registry*

ization r

Statistics Denmark (2020a). Befolkningen (BEF, Population Demographics, 1985-2011 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

Statistics Denmark (2020b). Dœde i Danmark (DOD, Deaths in Denmark, 1980-2013 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

Statistics Denmark (2020c). Hustande og familier (FAIN, Households and Families, 1980-2007 [database]. Danmarks Statistiks Forskningsservice, accessed 2014.

https://social-science-data-editors.github.io/guidance/Requested_information_dcas.html#example-for-government-registers

http://www.dst.dk/extranet/forskningvariabellister/Oversigt%20over%20registre.html

Details on the Reproducibilty Check

What is the reproducibility check?

- **Data checks**
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS: When data are present, run checks:

- **Can data be read** (using software indicated by author)?
- Is data in **archive-ready formats** (CSV, TXT) or in custom formats (DTA, SAS7BDAT, Rdata)?
- Does the dataset **have variable labels?**
- Run **check for PII**. Apply judgement.

What is the reproducibility check?

- Data checks
- **Code description**
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS:

- **Review the code** (but do not run it yet).
- Identify programs that create "analysis files" ("**data preparation code**").
- Identify **programs that create tables and figures**. Not every deposit will have separate programs for this.
  - Identify all Figure, Table, and any in-text numbers.

What is the reproducibility check?

- Data checks
- Code description
- **Requirements**
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers
- Software Requirements
  - Version of software (Stata 15, Matlab R2019b, etc.)
  - Complete list and version of packages!
- Computational Requirements
  - Type, vintage, memory size, speed of computer
  - Disk space!
- Time Requirements
  - Minutes, hours, days, weeks, months?

What is the reproducibility check?

- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- **Verbose description** of steps to replicate
- Findings
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS

- Provide details about your process of accessing the code and data.
- DO describe actions that you did as per instructions
- DO describe any other actions you needed to do ("I had to make changes in multiple programs”)
- Findings come later

What is the reproducibility check?

- Data checks
- Code description
- Requirements
  - As stated by author
  - As encountered by replicator
- Verbose description of steps to replicate
- **Findings**
  - Compare tables
  - Compare figures
  - Compare in-text numbers

INSTRUCTIONS:

- Describe your findings both positive and negative in some detail, for each **Data Preparation Code, Figure, Table, and any in-text numbers**.
- When errors happen, be as precise as possible.
  - For differences in figures, provide screenshot of manuscript figure, as well as the figure produced by the code you ran.
  - For differences in numbers, provide both the number as reported in the manuscript, as well as the number replicated.

Coding for Reproducibility

Streamlining replication packages

- Master script preferred
  - Least amount of manual effort
- No manual manipulation
  - "Change the parameter to 0.2, then run the code again"
- No manual copying of results
  - Write out/save tables and figures using packages
  - Compute all numbers in package
- No manual install of packages
  - Use a script to create all directories, install all necessary packages/requirements/etc.
- Clear instructions!

Some tips from the "frequently gotten wrong" bin

- Set the project directory **ONCE** in code, or **NEVER** (Stata, R, Python)
- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)
- **Write out all tables, figures**, and in-text numbers into separate files

If you need to manually modify the code to obtain a series of tables/figures/columns, you're doing something wrong:

- Use **functions**, **ado files, programs**, **macros**, **subroutines**
- Use **loops, parameters, parameter files** to call those subroutines

Some tips from the "frequently gotten wrong" bin

Cleanly **separate**

- Confidential data and public use data
  - You are going to have to provide copies of the public use data without compromising confidentiality
- Confidential parameters and the rest of the code
  - Reduces need to redact programs

- Use **placeholders** (globals, libnames, etc.) for common locations ($CONFDATA, $TABLES, $CODE) (Stata, R, Python, SAS)

Some tips from the "frequently gotten wrong" bin

Have "**computational empathy**"

- Consider cross-platform programmng practices
- Consider that the replicator can learn from the process
  - They probably don't have the same knowledge
- Consider that the replicator might not have the same modules/packages/etc.

- Path and filenames:
  - Stata: always use forward slashes, even on Windows
    ```
    use "$data/path/data.dta"
    ```
  - R: use "`file.path()`"
    ```
    x <-
    read(file.path(data,"data.dta"))
    ```
  - SAS: use `filename` and `libname` to abstract
    ```
    data DATALIB.step1;
    set CONFLIB.slid_1996;
    ```

Extreme examples

- Matlab-based simulation
- Real example, 10 figures, 4 panels each

- For Figure 5a, comment line 52, uncomment line 151, run the code, then copy the figure into your document.
- For Figure 5b, comment line 151 again, leave line 52 commented, and change the parameter on line 75 to “3”
- ….

Extreme examples

- Stata-based estimation
- 4 variants

- Run the data creation programs, then copy the data to Folder A
- Copy programs “b.do” and “c.do” from Folder A to Folder B, but modify “c.do” on line 20
- Once done, convert the output from “d.do” to a Matlab file, and run the simulation in Folder B/C
- ….

Ideal setup

- 1 program to prepare the setup
  - Installs all packages
  - Creates all directories
- 1 program (or a very small number) that creates the rest
  - Possibly with macros/ ado files/ subroutines
  - Possibly with parameter files that might differ per directory
- All tables and figures are output programmatically

- Setting up can be done in all languages
  - Matlab, Stata, R, Python, Fortran
- Subroutines exist in all languages
  - You might need to learn how!
- Ability to output figures and tables (Excel, LaTeX) exist in all languages

Preparing Replication Package

Follow the steps

AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

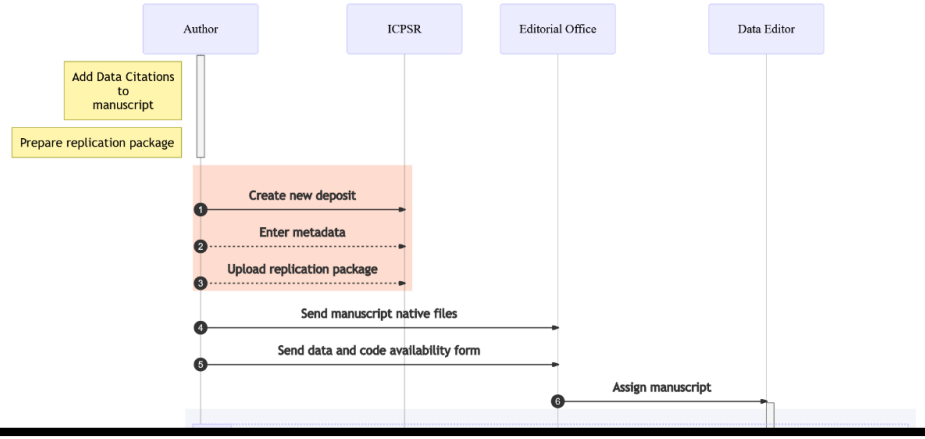
Guidance for authors wishing to create data and code supplements, and for replicators.

Step by step guidance

On this page:

The following steps outline what you should expect after conditional acceptance of your manuscript, in compliance with the AEA Data and Code Availability Policy:

1. Prepare your data and code replication package (including data citations and provenance information)
2. Provide metadata and upload the replication package, for verification and subsequently publication.
3. Submit the Data and Code Availability Form together with your manuscript native files as instructed, and as per guidelines at your journal (for example, AER guidelines).
4. The editorial office assigns the manuscript to the AEA Data Editor.
5. The AEA Data Editor team downloads materials, conducts reproducibility checks , writes report.
6. The report is communicated to the editorial office and the Editor of the journal.
   - If accepted, the manuscript is copy-edited, and published together with the data deposit as provided by the author.
   - If changes need to be made, the report is communicated to the authors, who make changes, until the replication package is accepted.



https://aeadataeditor.github.io/aea-de-guidance/step-by-step.html

How to test the replication package

- README
- Full package of all programs, data you intend to provide
- ZIP it up

- Now ask an RA/ colleague/ friend/ grandma/ daughter not previously involved to
  - Download the package
  - Follow the instructions in the README without talking to you!
  - Compare the results to the paper

How to prepare the replication package

- README
- Full pac
  data yo
- ZIP it u

- Now ask an RA/ colleague/
  nter not

in the
g to you!
the paper

Policy and Protocol on Third-Party Verifications

- Preliminaries
- What or Who Is a Third-Party Replicator
- Steps for the Third-Party Replicator

This protocol describes how third parties can, at the request of the AEA Data Editor, conduct a reproducibility check.

Alternate protocols are possible, but should be verified with the AEA Data Editor prior to engaging any resources.

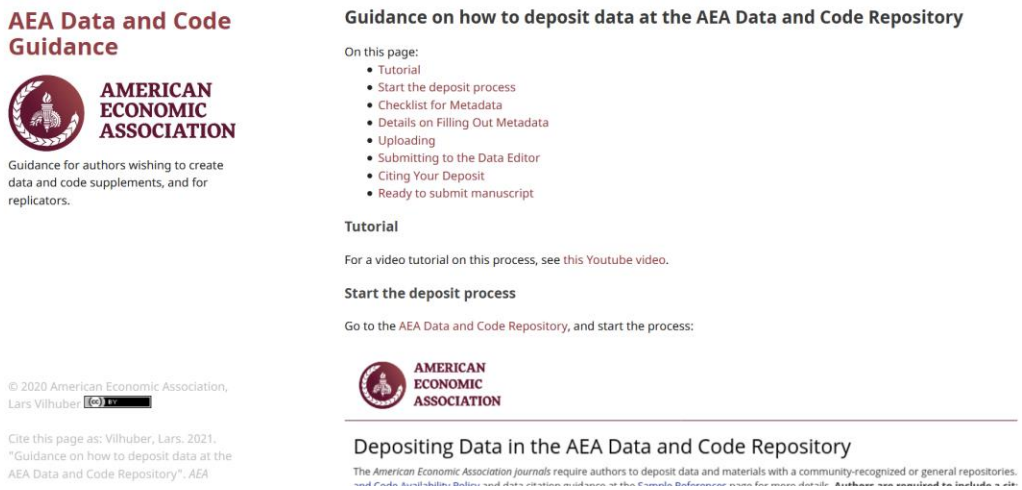
Preliminaries

- The author(s) should provide a complete and exhaustive archive, ready for publication, to the AEA Data Editor.
  - The archive does not need to be public at this stage, as long as it can be shared privately.
  - The archive does not have to contain the data necessary for the reproducibility check if data is confidential or proprietary. However, the archive must contain a publishable description of how an independent researcher can

https://www.aeaweb.org/journals/data/policy-third-party

Then follow the steps to upload

- No ZIP file *upload* - use the "Import from Zip."
- When revising, do not create new deposit – re-use existing project.



AEA Data and Code Guidance

AMERICAN ECONOMIC ASSOCIATION

Guidance for authors wishing to create data and code supplements, and for replicators.

© 2020 American Economic Association, Lars Vilhuber

Cite this page as: Vilhuber, Lars. 2021. "Guidance on how to deposit data at the AEA Data and Code Repository". *AEA*

Guidance on how to deposit data at the AEA Data and Code Repository

On this page:

- Tutorial
- Start the deposit process
- Checklist for Metadata
- Details on Filling Out Metadata
- Uploading
- Submitting to the Data Editor
- Citing Your Deposit
- Ready to submit manuscript

Tutorial

For a video tutorial on this process, see this Youtube video.

Start the deposit process

Go to the AEA Data and Code Repository, and start the process:

AMERICAN ECONOMIC ASSOCIATION

Depositing Data in the AEA Data and Code Repository

The *American Economic Association journals* require authors to deposit data and materials with a community-recognized or general repositories.

https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea.html

The role for journals

Goal: Transportability

Any standards, tools, methods: must be transportable across journals (no custom solutions)

Social science “guild”

Social Science Data Editors

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

Some resources

Some resources

- https://social-science-data-editors.github.io/guidance/
  - General guidance
  - Template README
  - data citation guidance
  - discussion of licensing
- https://aeadataeditor.github.io/aea-de-guidance/
  - Step-by-step guidance