Introduction

In the Fall of 2014, the Cornell Institute for Social and Economic Research (CISER) launched its Data Curation and Reproduction of Results Service. This service was instituted for the following reasons:

- to encourage researchers to share their code and data by eliminating one of the obstacles to sharing – fear that others will find inconsistencies between the authors’ published results and the results of running the code and data linked to their publication. The service only focuses on the reproduction of results and not on the correctness of the methodology, analysis, or interpretation of the results.
- for transparency and to provide additional material for re-users to understand exactly what was done. No more reverse engineering and wasting of time and financial resources figuring out how the authors came up with the results.
- Studies with links to code and data get cited more.
- Study gets scrutinized more, and once it passes scrutiny the research community validates the results and provides legitimacy to the work done by authors.
- It furthers science and speed up the process of knowledge transfer. Others can build on the authors’ work which can lead to more citations.
- Most prominent journals now require replication materials.

Golden Rule of Reproduction Service:

Output produced by running code against the data should be identical to those printed on the article up to the last decimal place. Slight deviation is not acceptable and must be investigated.

Step 1

Researcher provides CISER Staff copy of article, code(s), and data.

Requirements for article submission to CISER
- Client highlights all figures, tables, graphs, and sentences which have numbers derived from running the code against the data so CISER staff will know which ones to reproduce
- Preferably revise-and-resubmit stage, but CISER accepts papers that are ready for submission to a journal for the first time

Requirements for code submission to CISER
- If codes consist of multiple files, specify the sequence of execution (preferably write Step number on the filename)
- Put comments in code especially on sections that produced the numbers used in the article. Comments should describe what the section of the code produces or is doing.

Requirements for data submission
- Must be base data from which data manipulations and final datasets used in models emanate
- Clean data
- All variables and values labeled (preferably)

Step 2

CISER Staff checks submission for completeness

Common problems:
- Incomplete data and labels (values and variables)
- Filenames of codes have no indication of sequence

Proposed solution: Data Curation and Management Training

Step 3

CISER Staff prints the paper, runs the code against their data and compares the results with those cited (yellow-highlighted) on the paper

Staff has to read and understand the flow of the program before running it. Researchers should put extra effort to help ease the understanding of the code.

Common problems:
- very long
- complex
- output of codes do not match article (often figures on the article were derived from earlier versions of the code and were not updated)
- unnecessary sections of codes whose output are not found on the paper (this delays reproduction because the Staff has to go through entire code)
- Codes point to subdirectories for retrieving or saving data, thus Staff has to recreate the directory structure for the code to run correctly
- Some codes are not efficient, but this will not be modified by the Staff. The Staff, however, will suggest ways to make it efficient.
- Often in multiple files and no indication of sequence. Staff has to determine which to run first especially if codes build on top of the other.
- Not all figures printed on the article are produced by the code. Some involved other software packages such as Excel.

Proposed Solution:
- Training in code writing and organization
- Training in macro programming to make code more efficient (and short)
- Ideally researcher should sit down next to the Staff while reproduction is underway so questions are answered right there and then.

Step 4

Create repository for code, data, and documentation

Once research is reproduced, the codes, data, documentation, and study pre-print are hosted on the CISER Data Archive and other sites for maximum exposure (e.g., Cornell’s E-commons, Github, Bitbucket, or Dataverse).

Some journals prefer a specific type of repository, such as Git. If the researchers do not know how to create and use a Git repository, CISER will provide training.

Infrastructure and Staffing Requirements and Costs

Infrastructure needed
- At the very least, a supercomputer with various type of statistical software packages (including conversion software packages) to accommodate the size and memory requirement, and format of the code and datasets being deposited.
- Data Archive

Required Staffing Skills
- Statistical software programmer with knowledge of advanced statistics
- Data Librarian with expertise in digital curation and archiving

Costs
- Time of Researcher or RA in preparing code, data and documentation
- Time of programmer and data librarian
- Printing of article and tables
- Archiving of datasets, codes, documentations, and pre-print