R²: CISER's Data Quality Review and Reproduction of Results Service

Florio O. Arguillas, Jr. William C. Block
Cornell Institute for Social and Economic Research (CISER)

Corresponding author e-mail:
foa2@cornell.edu

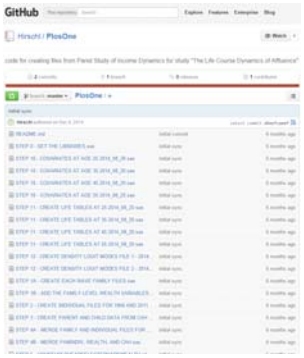
Abstract

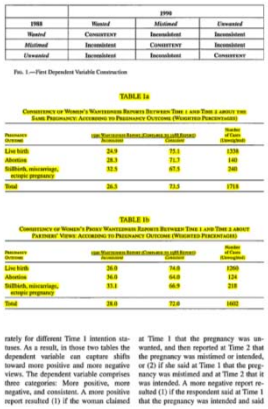
A year and a half after its implementation, CISER's Research Data Quality Review and Reproduction of Results Service (or R², for short), a service developed to encourage sharing of high quality data, code, documentation, and metadata associated with a study for the purpose of reproducible research, continues to evolve and improve, and has become even more cost-effective, while at the same time encouraging and enhancing researcher skills in data and file management, data quality review, code writing, and version control. This poster discusses: a) the service at its current state; b) the improvements made to the service including cost-reduction strategies (such as trainings that pertain to the aforementioned skills) and buy-in strategies to encourage researchers to use the service; c) pre- and post-reproduction services to improve data, code, documentation, and metadata quality; d) the application of the Comprehensive Extensible Data Documentation and Access Repository (CED2AR) software for assessing and generating complete metadata in DDI format; and e) the utilization of the CISER Data Archive as the free and permanent home for the study and its associated files, along with persistent identifiers, download tracking metrics; reuse and citations monitoring; and commitment to support the collection through changing technologies, new media, and data formats.

Golden Rule of CISER's Replication Service:

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

Common problems:
- Very long, complex codes
- Unnecessary/excess sections of codes whose outputs are not found in the paper (this delays replication because the Staff has to go through the entire code and its output, and figure out where they are on the paper)
- Code points 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 will not be modified by the Staff. The Staff, however, will suggest ways to make it efficient.
- Codes are often multiple files with no indication of sequence. Reproducer has to determine which to run first especially if codes build on top of the other.





Researcher provides CISER Staff copy of article, code(s), and data; highlights the sections on the article with figures derived from running code against data; and put comments on the codes that describe what the section of the code will produce or is doing. CISER uses CED2AR to check for completeness and create DDI metadata

Common problems: No variable and value labels

Cost-reduction strategies
- Data curation and management training
- Code writing and organization training
- Code efficiency training e.g., macro programming, SQL programming
- Version control software training e.g., Github

Common problems:
- Some results do not match article
- Not all figures printed on the article are produced by the code. Some involved other software packages such as Excel.
- Order of variables in the model in the printed table do not match the order of the variables in the output table for that model, which slows down the verification process.

International Association for Social Science Information Services & Technology