Guidelines for data publication

February 26, 2015

Introduction

Journals, funders and research organizations are increasingly moving towards greater research transparency, including sharing materials such as data and code. Sharing research materials beyond published articles is valuable both because it allows others to examine the fuller body of evidence for reported results, and also because it facilitates fuller re-use of collected data.

Innovations for Poverty Action has joined the movement towards greater transparency, by adopting a data-sharing requirement and creating a public repository for sharing data. We invite researchers to submit their materials to our data repository team, which provides assistance in preparing the data and code and uploading it for publication.

This guide is for IPA affiliates, and includes our guidelines for data-sharing as well as an explanation of IPA's data curation process. For question and comments, please contact Stephanie Wykstra at researchsupport@poverty-action.org.

Contents

Why share data and code?

IPA has the objective of advocating for research transparency, including pre-registration of studies and data-sharing. This objective falls under our commitment to high quality research. By sharing data, we allow others to re-analyze results and perform robustness checks using the original materials, providing a fuller body of evidence for research results.

In moving towards greater transparency, IPA joins many other groups that have adopted data access policies. Increasingly, funders require researchers to share their data, including the Gates Foundation, the National Science Foundation and the National Institutes of Health. In addition, many journals are adopting data-sharing requirements which apply to the data/code underlying published results.[1]

IPA's data-sharing requirement:

IPA requires that researchers share data from IPA-implemented and IPA-funded research projects in our public repository. The timeframe for sharing is **three years** following the completion of final data collection. The timeframe is intended to allow time for publication; where publications are still pending, we will work with researchers on an extended timeframe. (We also welcome materials earlier, when it is feasible.)

IPA's Research Methods and Knowledge Management team **offers assistance** in preparing datasets for public sharing. We outline the data and other materials that we will ask for in the table below. We are happy to offer assistance with code checks and other tasks *prior to publication* upon request, as this may often by a helpful service to researchers asked to share data/code with journals.

Note that the middle row below illustrates **what we request as a part of IPA's requirement**. The "less than recommended" and "exceptional" levels are presented for a contrast.

What to share?

Level of sharing	Materials included	Allows for	Recommended file formats[2]
Less than IPA-recommended	• Data and code underlying published results • Readme file explaining relation between the files • Minimal study-level metadata (i.e., information about the study).	• Checking that the tables in the published article can be produced by running the code • limited re-use	**Data:** .dta (preferred: we will then share both .dta format and non-proprietary .tab format). .csv, plus codebook.

1 Stodden, Guo and Ma, 2013. "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals." http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067111

2 In this section on recommended file formats, we draw on a resource created by the UK Data Archive: http://ukdataservice.ac.uk/manage-data/format/recommended-formats.aspx

IPA-recommended	*First row, plus…* • Full set of collected and other (e.g., admin) variables, excluding personally identifiable information (PII). • Data documentation including additional useful information about the study (context, notes on data-cleaning process, etc.) • Survey instrument(s)	*First row, plus…* • Fuller potential for re-use and secondary analysis • Better understanding of study context – useful for systematic reviews and application to policy	**Code:** .do, .txt **ReadMe files:** .pdf, .doc, .docx **Survey instruments:** .pdf, .doc, .docx
Exceptional	*First two rows, plus…* • Cleaning and variable construction code	*First two rows, plus…* • Start-to-finish reproducibility	

How to prepare research materials to share?

Instructions for data sharing:

Step 1 - Prepare and send materials to data repository team:

Contact Stephanie Wykstra (swykstra@poverty-action.org) indicating that you have data to share. The data repository team at IPA will follow up with a link for you to securely share your data.

Step 2 – The data repository team checks and curates the materials:

The data repository team will conduct data curation steps – see below for details.

Step 3 - Public release of data into the repository:

We will check and confirm any additions or other suggestions before releasing research materials into the public IPA data repository.

Detailed steps for preparing data and code to share with the data repository:

1. **Remove personally identifiable information (PII):** Check thoroughly for PII, and make sure to remove before sharing with the data repository team.

- This is essential to protect study participants' personal information. Anyone receiving PII must be on the Institutional Review Board (IRB); typically, IPA's data repository staff members are not listed on the original IRB, nor added through an amendment.
- All direct identifiers such as unique IDs (social security numbers, bank account numbers, and so on) should be removed before sharing with the data repository team. Indirectly identifying data such as combinations of variables which could uniquely identify participants should also be considered carefully before publicly sharing data. Please contact us (research-support@poverty-action.org) with questions.

2. **Include clear variable labels and value code labels:**
    - Make sure that variables are clearly labeled.
    - If it is a variable collected directly from the questionnaires, indicate this with a question number. If it is constructed, either include the construction in the name or label, or if complex/lengthy, include additional information in notes.
    - Ensure that value code labels are provided, as they are needed for interpreting the data.
3. **Include code file(s) with headers/comments:**
    - **Headers:** Include header with name of person who last wrote/edited the code, date, and software used (package and version).
    - **Comments:** Use comments in the code to indicate which tables are produced.
4. **Prepare Readme files:**
    - Please indicate: 1) which files are included in what is shared; and, 2) how data and code files relate (i.e., what code runs on which data, to produce which outputs). We have a template for readme files that we are happy to share upon request.
5. **Include survey instruments:**
    - Ensure that you are sharing the final version used to collect the data.
6. **Fill in study-level metadata template (i.e. information about the study such as researchers, abstract, location, method of data collection, and so on):**
    - When you submit materials, we will ask you to fill out study-level metadata describing your study. *There are ~15 required fields; we will contribute staff time to helping with this step as needed.*

Data curation steps that data repository staff will complete:

As the data repository team works on the dataset submitted, we will conduct the following three steps to ensure the quality of the materials that we share in our repository.

1. **Confirming there is no personally identifiable information (PII) shared in data or code files**
    - ➢ It is the responsibility of the original researcher (s) to ensure that PII is removed, and IRB protocols do not permit sharing PII with the data repository team. However, the DR Unit will double-check that PII is removed before sharing, because of the high level of importance of maintaining confidentiality of research participant's information.

2. **Examining data and code for usability:**

- The data repository team will examine variable names and labels, value codes, and the statistical code. As a part of sharing high-quality data, we will attempt to fill in variable labels and/or notes in the dataset where we are able to glean further information from published tables or communication with researchers. Where there are a large number of unclear variables, we may ask the researcher(s) to improve the dataset before publishing.
- We will run the statistical code to ensure that it produces the published tables.

3. **Checking and sharing related materials:**
   - **Supplementary readme file:** As we conduct our data curation steps, we will track and share information that will help site users understand the steps that we took, and what we found. We will confirm with the original researcher before sharing this file along with the data.
   - **Study-level metadata:** we have created a custom template with fields that we will fill in from all studies.

FAQ:

- **Who can I contact with questions or for assistance with data-sharing?**

Please contact Stephanie Wykstra (Research Manager, Data Publication) at swykstra@poverty-action.org.

- **Where does IPA share data?**

We are sharing data using our Dataverse ("Randomized controlled trials in the social sciences Dataverse"). In addition to including IPA data, we are linking in RCT data from other groups such as J-PAL and Center for Effective Global Action (CEGA). Dataverse is commonly used for sharing data in the social sciences, and is a stable, well-resourced platform. While it is run out of IQSS at Harvard, it is open to all researchers and research groups (e.g., journals and research groups such as IPA) for data-sharing.

- **What if I want to share my data on my website or a different repository?**

IPA is investing in an effort to centralize data within one central data repository to ensure ease of use and access. We will contribute staff time to upload materials to the repository. If you choose to share the data in another location in addition, we'd ask that you let us know where else the data are shared, so that we can make a note of this in our repository.

- **What if a funder of my project has a different timeframe from IPA's, in terms of when they require data-sharing?**

We will approach these cases individually. Where there is a requirement with a shorter timeframe than IPA's, the shorter timeframe will apply.

- **What does IPA recommend in order to prepare materials earlier on in the research lifecycle, so that data-sharing is easier later on?**

IPA and J-PAL offers three global staff trainings per year, for new research staff. During these trainings, we offer training on best practices for managing data and code for research staff. In addition, we have a manual [forthcoming link] on these best practices. We also welcome requests to work with IPA and JPAL research staff on an individual basis, consulting and providing targeted training on best practices for managing code and data.