J-PAL Guidelines for Data Publication[1]

June 2015

I. Introduction

In order to promote greater transparency of research, a growing community of journals, funders, and research organizations have adopted policies that (a) require research projects be pre-registered in a public registry, and (b) that associated materials such as data and code be shared publically (*Innovations for Poverty Action (IPA) Researcher Guidelines*, 2014; *Transparency and Openness Promotion (TOP) Guidelines*, 2015*)*. The former is meant to combat publication bias.[2] The latter—publishing associated materials from research projects beyond its journal articles—is valuable because it allows others to examine the robustness of reported results, and facilitates fuller re-use of collected data in general.

The Abdul Latif Jameel Poverty Action Lab (J-PAL) has joined these efforts by promoting the development and use of the AEA trial registry, and now, **adopting organization-wide guidelines for data publication**. Both of these policies—registration and data publication—are consistent with J-PAL's mission to promote high quality, rigorous evidence.

This document contains J-PAL's policy on data publication:

- which projects fall under this policy;
- where materials should be published;
- by when should the materials be published (how long after data collection);
- what data and supporting materials should be published (variables, code, measurement instruments); and,
- in the appendix, how the data publication process works.

J-PAL Global's research team will provide affiliates assistance in publishing their data on the J-PAL Dataverse, such as preparing the data and code, replicating results, tables, and uploading it for publication. For help requests, or other questions and comments, please contact James Turitto (Research Manager, J-PAL) at data@povertyactionlab.org.

II. Data-Sharing Requirements

Which projects fall under this policy?

J-PAL's data publication requirement applies to all research projects that have received or will receive funding from J-PAL research initiatives. However, we *strongly encourage* affiliated researchers to share data and code from other research projects as well. As more of our affiliates publish datasets, others will feel compelled to do so as well.

---

[1] J-PAL and IPA try to synchronize research policies and defaults to avoid confusion for PIs who work with both organizations. This document, however, reflects J-PAL policy and not necessarily IPA policy.

[2] Pre-registration should not be confused with a pre-analysis plan (PAPs). Pre-registration includes a brief description of the research design including the primary outcomes, and ultimately the primary results. Authors who conduct meta-analyses can therefore be aware of the study and its results, even if no paper is published. Pre-analysis plans, on the other hand, describe regression specifications, subgroup analyses and are intended to prevent data mining.

Where materials should be published

All published data should be stored in **trusted digital repository**, such as the Harvard Institute for Quantitative Social Sciences (IQSS) Dataverse[3], the Interuniversity Consortium for Political and Social Research (ICPSR), the Yale Institution for Social and Policy Studies (ISPS) Data Archive, etc. A trusted digital repository has the following characteristics:

- Commits to, and is financially capable of, maintaining the repository in perpetuity;
- Ensures minimal to zero data loss or "data rot";
- Allows for version control; and
- Assigns each dataset and its associated program code with a unique identifier (i.e. the "Digital Object Identifier" (DOI)) to make cataloguing, citation, and search more official.

While posting data on personal websites (even if university-hosted) technically makes them "public," these locations are not sufficiently secure, and are more difficult to search. Instead, we encourage researchers to provide links to repository entries on their personal websites.

When should the materials be published

Data should be made available **18 months** following the completion of final data collection. Data will be held by J-PAL under an embargo agreement. Each year thereafter, J-PAL will ask researchers whether the dataset has been made available. If there is no response, J-PAL will keep the embargo. On the **fifth year** following data collection, the presumption is that J-PAL will share the data. J-PAL will again ask the researcher if the dataset can be made public. If there is no response, the dataset will be made public. In cases where the researcher requests a further extension, they will be asked to submit this request to the Initiative Co-chairs for their approval.

What data and supporting materials should be published

J-PAL requests the following data materials to be shared based on the type of data used to produce results:

- Authors reusing data available from public repositories must provide program code, scripts for statistical packages, and other documentation sufficient to allow an informed researcher to precisely reproduce all published results.

- Authors using original data must:
    - make the data available at a trusted digital repository (Note: If all data required to reproduce the reported analyses appears in the article text, tables, and figures then it does not also need to be posted to a repository.);
    - include all variables, treatment conditions, and observations collected from implemented survey instruments (Researchers will be requested to requested to fill out a form listing all datasets and identifying datasets to be made completely public, partially public, or kept private.);
    - provide a full account of the procedures used to collect, preprocess, clean, or generate the data;

- o provide program code, scripts, codebooks, and other documentation sufficient to precisely reproduce all published results; and
- o provide research materials and description of procedures necessary to conduct an independent replication of the research.

- In rare cases, despite authors' best efforts, some or all data or materials cannot be shared for legal or ethical reasons. In such cases, authors must inform J-PAL. Authors are encouraged to anticipate data and material sharing at the beginning of their projects to provide for these circumstances. It is understood that in some cases access will be provided under restrictions to protect confidential or proprietary information. J-PAL may grant exceptions to data and material access requirements provided authors:
  - o explain the restrictions on the dataset or materials and how they preclude public access;
  - o provide a public description of the steps others should follow to request access to the data or materials;
  - o describe software and other documentation that will precisely reproduce all published results; and
  - o provide access to all data and materials for which the constraints do not apply.

III. Appendix

How to prepare research materials to share?

When preparing a data files for sharing, researchers should follow the guidelines in ICPSR's Guide to Social Science Data Preparation and Archiving.

Instructions for Sharing Data on the Dataverse

1. **Step 1 - Prepare and send materials to data repository team:**
   Contact James Turitto (data@povertyactionlab.org) indicating that you have data to share. The data repository team will follow up you to securely share your data.

2. **Step 2 – The data repository team checks and curates the materials:**
   The data repository team will conduct data curation steps – see below for details.

3. **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 or J-PAL data repository.

Checklist for Preparing Data to be Shared

- ☐ All personally identifiable information (PII) has been removed, anonymized or suppressed.
- ☐ Variable labels and value code labels are clear.
- ☐ Code files have headers (including the name of the person who last wrote/edited the code, the date, and the software used) and code has commenting.

- ☐ Readme files have been created to indicating what files are included and how the data/code files relate.
- ☐ The final versions of all survey instruments used to collect data are included.
- ☐ A copy of the manuscript (published or working paper) that was written based on the specific data and code, if applicable.

Data Curation Steps to be completed by the J-PAL Research Team

If you are receiving assistance from the J-PAL Research Team to prepare the dataset for publication, we will conduct the following quality control steps prior to publication 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 J-PAL Research Team 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 replicability and usability:**
   - ➤ The J-PAL Research 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 replicates the published tables.

3. **Checking and sharing related materials:**
   - ➤ **Supplementary readme file:** As the J-PAL Research Team conducts its 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 will use a custom template developed by Innovations for Poverty Action (IPA) with fields that we will fill in from all studies.