Data and code sharing policy

Many CGD publications contain original statistical analysis. CGD's policy is that the full details of these analyses should be publicly shared. The original data and all computer code needed to prepare and perform an analysis should be posted on cgdev.org along with the write-up. Like most CGD policies, this one is soft. Exceptions will be necessary, as discussed below.

Transparency defined

Data analysis usually has three steps:

1. *collection,* such as through original survey work or downloading of public data sets;
2. *preparation*, of data for statistical analysis, which may involve filtering and aggregation;
3. *analysis*.

To be transparent, authors should document the first step and post all computer files that another person would need in order to rerun the second and third steps. Data and code should be organized and documented so that it is *easy* for others to rerun the data preparation and analysis and match results from the output with results displayed in the publication.

Benefits

CGD analyses should be acts of social science. By some definitions, a *sine qua non* of science is *replicability*. The responsibility for replicability is especially great for research that aims to influence policy and ultimately affect the lives of the poor. Bruce McCullough and Ross McKitrick put it well in their report, *Check the Numbers: The Case for Due Diligence in Policy Formation*:

> When a piece of academic research takes on a public role, such as becoming the basis for public policy decisions, practices that obstruct independent replication, such as refusal to disclose data, or the concealment of details about computational methods, prevent the proper functioning of the scientific process and can lead to poor public decision making.

In fact, transparency has many benefits:

- It makes analysis more credible.
- It makes CGD more credible when it calls on other organizations, such as aid agencies, to be transparent.
- Data and code are additional content, appreciated by certain audiences.
- It increases citation of CGD publications—by people using associated data sets.
- It curates, saving work that otherwise tends to get lost as the staff turns over.
- Preparing code and data for public sharing improves the quality of research: researchers find bugs.
- In short term, CGD's leadership in transparency will differentiate it from its peers. In the long term (one hopes), CGD's leadership will raise standards elsewhere.

Means

Disclosure can take different forms; the choice will depend in part on how an analysis was done. Disclosure can be as simple as a zip file of Stata data and (annotated) command files, linked to from a publication landing page. Or files can be posted individually, with links organized into a table of contents on the landing page (example: working paper 174). Data set of sufficient value can be posted as official CGD data sets, which are a publication type.

Usable file formats include, but are not limited to, Excel, Stata, Access, and SQL Server. To maximize accessibility, older formats should be considered, such as Excel's .xls format and Stata's version 9 data format; these are usually accessible through the Save As… command. Posting data sets in multiple formats can also be worthwhile. Excel files are more universally readable than Stata files.

Documentation should make obvious how to use the files and how they correspond to figures, tables, and statistics in text.

Ideally, program files will allow others to replicate a publication's entire analysis with a few keystrokes. This discipline should force all processing steps into command files. E.g., when performing the analysis for the first time, an author might type commands at the Stata prompt—to make a particular figure or build a particular variable. But once polished, that command should be stored in an executable file.

Exceptions

Several factors may impede data sharing:

- Confidentiality of source data.
- Commercial/proprietary nature of source data.
- High professional cost of obtaining data and high professional value of not sharing it. Data can be costly to obtain, whether because of extensive survey work in rural Sri Lanka or delicate relationship building with the Kenyan ministry of education. Mandatory sharing of high-cost data *may* reduce its professional value to the collector by giving competing scholars early opportunities to analyze and publish from it. And reducing the incentive for data collection will reduce data collection.

When such issues arise, authors should state in writing, to the research manager, their reasons for not being transparent. And they should strive for compromise—perhaps sharing not the raw data, but that used directly in analysis, along with code. When possible, authors should disclose more over time. For example, after an author has mined a dataset for most of its publishing potential, the professional cost of sharing it falls while the benefits remain. Data sets for works already published on cgdev.org can then be posted.