Replicating the Data Analysis in WDR 2018: Learning to Realize Education's Promise
Technical Note

Bradley Larson
World Bank Group

The World Development Report (WDR) 2018 data replication package comprises all source data and Stata replication files (do files) necessary to create the charts and graphs that appear in *WDR 2018: Learning to Realize the Promise of Education*.[1] These files have been compressed and posted on the WDR website, here (with individual files here).[2] After downloading and unzipping the files, users can replicate figures individually or—using the "master" do file—all together.[3] Stata 14 is required to run the replication files, along with user-written programs that can be downloaded from the Boston College Statistical Software Components (SSC) archive (or other websites as noted) and a handful of settings files and programs written specifically for the WDR 2018, which are described below.

Data files are saved in Stata format (DTA). Country names have been cleaned to meet the World Bank standard for fiscal year (FY) 2017, available here.[4] Variable and data labels have also been added to facilitate analysis. Otherwise, these source data files provide original data, as pulled from online databases, statistical annexes, published tables or figures, and unpublished datasets shared by scholars in the education field.

In most cases, all observations in the original data set were retained, even if the final figure does not include them. For example, figure O.2 in the Overview chapter of the WDR 2018 shows the distribution of PISA scores for 20 selected countries, along with the OECD average, whereas the source data in the replication package includes data for 71 countries. Running the replication files will generate the figure with 20 countries and an average, but the user could easily expand or adapt the analysis using the data provided. The only instances in which observations were dropped during the data cleaning process relate to subnational data (learning exam scores for U.S. states, for example). The replication data includes only countries and economies recognized by the World Bank and aggregations of those values.

In contrast, variables that appear in original datasets, but were not used in analysis for the WDR 2018, have been culled from the source data sets available here. The most significant benefit was reduced file sizes, which allows the package to be shared online. In cases where country-level or aggregate data were taken from published tables or figures, only those values are provided in the source data presented here; no effort was made to expand those datasets.

The great majority of replication files included in the package—57 out the 59 Stata do files—generate one of the 60 figures in the WDR 2018. These replication files are named in accordance

[1] http://www.worldbank.org/en/publication/wdr2018
[2] The package can be downloaded here: http://pubdocs.worldbank.org/en/781891534452606153/WDR-2018-data-replication-package.zip. Individual files can be found here: http://www.worldbank.org/en/publication/wdr2018/brief/world-development-report-2018-data.
[3] Replication files for other statistical packages will not be released, but all data points depicted in charts and figures are available in CSV format, linked in the source note for each figure of the Report.
[4] http://databank.worldbank.org/data/download/site-content/OGHIST.xls

with their numbering in the WDR 2018, so the file named "Figure O_2.do" generates Figure O.2, "Box Figure B6_3_1.do" generates Box Figure 6.3.1, and so forth. Three of the figures in the Overview are repeated later in the report. Separate do files are not provided for each copy of the figure. Rather, a single do file—named for the first occurrence of the figure in the Overview chapter—generates two copies with different reference numbers. Thus, "Figure O_6.do" generates Figures O.6 and 2.5; "Figure O_8.do" generates Figures O.8 and 3.9; and "Figure O_10.do" generates Figures O.10 and 3.12.

Five different output files are produced for each figure. Three of the files are image files, in Encapsulated PostScript (EPS), Enhanced MetaFile (EMF), and Portable Network Graphics (PNG) formats, which allow the figures to be viewed in different applications and operating systems. The fourth file contains the figure's metadata in DTA format, including the reference number, title, subtitle, source, note, and other fields used to organize and generate the figure. The fifth file is the comma-separated values (CSV) file that provides the data points plotted in the figure, and which is linked in the figure's source note in the WDR 2018. These output files are saved in the "output" subfolder using the figure reference numbers described above and subscripts "data" for the CSV archive files and "meta" for the DTA metadata files.

Each of replication files for figures can be run independently. However, all files that replicate figures require three additional programs to run successfully. These files are provided in the root directory, wherever the replication package was unzipped, and should not be moved from there. The first ancillary file is the WDR settings file, called "wdrsettings.do", which sets universal options, creates folders if they are missing, sets paths using global macros, and provides the WDR color palette for charts and graphs. The second is the data archiving program, "wdrarchive.ado", which takes a snapshot of the data points used in the graph or in a panel of the graph to form the basis of the data archive files described above. The third is the export program, "wdrexport.ado", which compiles and saves the five output files described above.

The replication files for figures can also be run as a batch using the master file called "WDR 2018 MASTER.do". Running this program will generate each of the 60 figures sequentially (300 files in total). It will also compile a Microsoft Word document containing all of the figures using a program called "wdrcharts.ado", which in turn draws on several user-written programs to read in the EMF image files and metadata files described above using the Markdown language.[5] If desired, the document-generating program ("wdrcharts.ado") can also be run on its own, and will simply generate a document containing all of the figures saved in the output folder.

Finally, generating certain figures, like Figure 2.4, require special user-written programs. If the replication files are run individually, and the user does not already have the programs installed, the do file will crash. In that case, the user can simply download and install the appropriate program. (In this case, the user would do so by entering the command "ssc install triplot".)

[5] These user-written programs were developed by E. F. Haghish of the University of Freiburg, Germany, and are available in the "markdoc", "weaver", and "statax" packages available from SSC or Professor Haghish's GitHub site: https://github.com/haghish/. One of the programs in the markdoc package available on SSC, "img.ado", does not run successfully in Stata 14. The version of "img.ado" on Professor Haghish's GitHub site, does work however. It has also been included in the WDR 2018 replication package, so the user can simply copy it the appropriate ado path.

Alternatively, running the master do file will install all the necessary programs, and the user should not have to install them individually.[6]

[6] The remaining program, "splitstring.ado", was written for the WDR 2018 for the simple task of splitting string values into two lines of roughly equal length. It can be copied to the "plus" folder in the user's adopath or left in the root folder of the replication package, at the user's discretion.