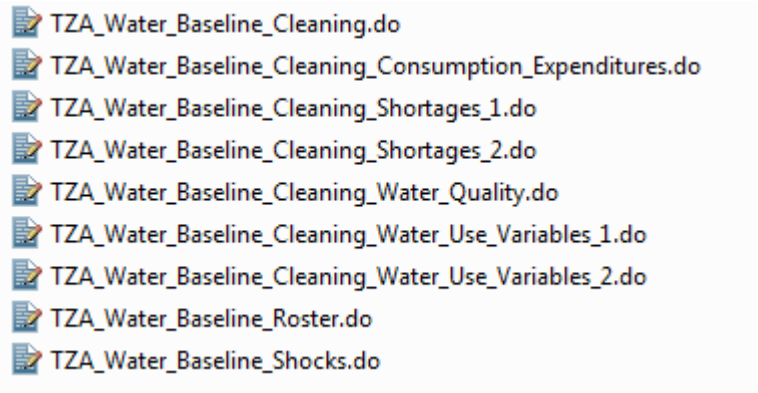
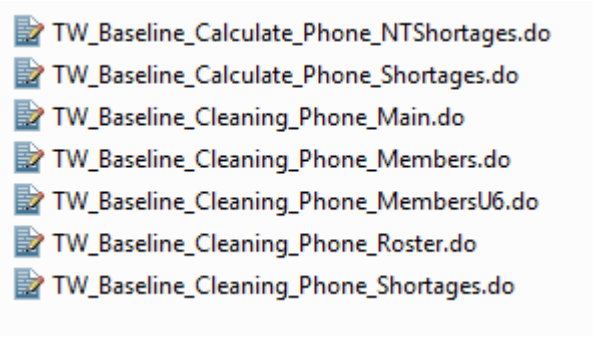
(1) DATA MANAGEMENT & DESCRIPTIVE STATISTICS:

For the main analysis files released for this evaluation, accompanying do files are provided to orient the secondary data user to any variables that were renamed from the original questionnaire(s), and also to follow the procedures used to calculate variables that were created during analysis.

The main **Read Me folder** contains the following files:



Within the read me folder, there is a sub-folder that contains do files that pertain to the phone surveys:



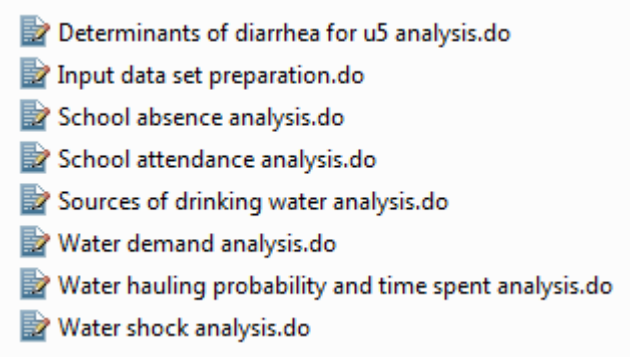
The file called **TZA_WATER_Baseline_Cleaning.do** is the parent file. Secondary users should start there. It will indicate when one of the other do files is needed for reference. The other do files have been included as supporting files, to make it easier to follow the process of renaming and calculation.

For variables for which renaming or calculation is relatively simple, and for which a separate do file is not necessary, the codes are provided within the parent file. *Note that codes contained in these files are not meant to be run in the existing datasets, which are considered final baseline versions – but rather to indicate the procedures and programs followed in order to achieve the datasets that are provided publicly for this evaluation. Note also that in many cases codes cannot be run because absolute raw data tables haven't been provided. But some codes can be replicated by applying them to the supplementary datasets – these cases will be apparent by following the structure and reading the explanations in the parent file.*

These do files should contain all needed reference codes to trace calculations and renaming for the main analysis files as well as the supplementary data files, along with the questionnaire reports, all provided on the MCC website. With any questions about programs used, contact: droumis@socialimpact.com

(2) STATISTICAL DATA ANALYSIS:

There is a second sub-folder called "Analysis files". These are programs that relate to the statistical models presented in the baseline report. There are many more models than we presented written in the codes, but the evaluation team has left them in, so that a secondary data user would get a sense of the other alternatives. In most cases, the first model that runs corresponds to the report, but someone consulting these do files would not have any problem going from the Stata output to the numbers in the report. **The data preparation program creates a few datasets that are used in the other programs; this should be run first. (Variables renamed or created within these files were not saved as part of the released datasets; to replicate them, the secondary user would simply run the code from within these files.)**. Everything else can follow in any random order. The files included in this sub-folder include:



(3) Note on application of sampling weights

The evaluation team implemented a two-stage stratified random cluster sample in Dar es Salaam and a two-stage random cluster sample in Morogoro. For this reason, sampling weights are necessary in order to obtain estimates that are representative of each city's population. Without the application of sampling weights, estimates only pertain to the sample of households in the dataset. In the baseline report, all estimates pertaining to each city's population have thus been adjusted with sampling weights. In the public release datasets, sampling weight variables are only contained in the HH datasets. For city-representative estimates using data from any of the other datasets (roster, supplementary, etc.) it will be necessary to merge in the sampling weights to those datasets. Code describing sampling weight calculation and application is included in the do file titled "TZA_Water_Baseline_Cleaning.do". Lines 79 through 107 describe the calculation and variables used in the baseline analysis. Lines 108 through 141 to show users how sampling weights can be merged into other datasets and applied to other datasets to calculate sampling weight-adjusted estimates of any indicators or outcomes.