GUIDE TO USING HOUSEHOLD DATASET

The questionnaire used for collecting this dataset is included with the data. The file is in MS-word format and is titled “questionnaire\_pit\_emptying\_smallholders\_140429\_v3.docx”).

The raw data is provided in excel format, and is spread across three excel files titled:

* “brac\_hh\_q1-q34”: contains data for Q1-Q34 of the questionnaire.
* “brac\_hh\_q35-q66”: contains data for Q35-Q66 of the questionnaire.
* “brac\_hh\_q67-q98”: contains data for Q67-Q98 of the questionnaire.

The variables in the dataset are named corresponding to the question numbers in the questionnaire. This allows anyone to use the data.

* For example, variable Q18 corresponds to Q18 in the questionnaire.
* As another example, Q22a corresponds to Q22 (how many).
* Q3221 corresponds to the answer in row 2, column 1 of Q32 (which is in a table format).

The questionnaire also displays the codes corresponding to the answers to all questions. The same codes are used in the excel sheets. For example, if the response of a household to Q26 is “1” in the excel sheet, this corresponds to option 1 in Q26 in the questionnaire (“we hired a sweeper from the same village”).

This dataset has been stripped of all identifiers.

* The names of the toilet owner (Q3 in the questionnaire); the name of the toilet owner’s husband/father (Q4); the name of the respondent (Q11) ; and the GPS readings of the house (Q9 and Q10) have been removed .
* The toilet serial number assigned by the NGO that financed the toilets has also been removed from the dataset (Q2).
* Instead, the research team created a variable called “slno” (Q1) that can be used to track data in the dataset, but not identify the households.
* The names of the enumerator, date of interview and the time of the interview (Q12-Q16) have also been removed from the dataset.

Demographic information: To ensure further privacy, names of all family members have been removed. This corresponds to Q1712, Q1722, Q1732, Q1742, Q1752, Q1762, Q1772, and Q1782 (This is column 2 in Q17 in the questionnaire).

A processed form of the dataset is provided in STATA format. This file is titled “households\_full\_dataset\_analysis\_160327”. This processed data is also provided in CSV format, with the file called “households\_full\_dataset.csv”.

The codes used for analyzing the data are provided in STATA do-file format (“households\_full\_dataset\_analysis\_160327.do”); they have also been provided in MS-Word format (the file name is “codes\_written\_for\_data\_analysis”).

The study design details are provided in the paper.

* Clustering unit: In the raw dataset, the variable “q05” in the excel file “brac\_hh\_q1-q34” is the clustering unit. In the processed dataset, “village\_name” is the variable for identifying clusters.
* Sampling weights: in the processed dataset, “s\_w” is the variable for sampling weights. We also provide the sampling weights in raw format as a separate CSV file (“sampling\_weights\_villages\_mymensingh\_150701.csv”).