Datasets

The main datasets for the project ("all_HH_nopii_bl.dta" and "all_HH_nopii_fl1.dta") include merged and appended data from 3 surveys: LHH, Targeting, and MHHH. The variable *survey* indicates which survey was administered to the household (LHH or Targeting) and variable *survey_m* indicates where an MHHH survey was also administered. Variables from the MHHH survey are merged to the main dataset using the Household ID (*hhid*) and variables names include the suffix "_m" to distinguish MHHH data from LHH/Targeting data.

#	Dataset	Description
1	all_HH_nopii_bl.dta all_HH_nopii_fl1.dta	Main project dataset (LHH, Targeting, MHHH surveys), excluding "roster" variables. Follow-up 1 cortisol results from saliva testing are included in this dataset (*cortisol_result*, *cortisol_result_m*).
2	b_hhmemcurrent_bl.dta br_hhmemcurrent_fl1.dta	Roster of current HH members with individual-level characteristics from module B (demographics) and R (information on TUP recipient members in fl1).
3	b_hhmemleftsincebl_fl1.dta	Roster of HH members that left the HH since Baseline survey.
4	c_a_foodconsumption_bl.dta c_a_foodconsumption_fl1.dta	HH food consumption, including type of item, number of days item was consumed, quantity consumed, and main source of item (weekly).
5	c_c_nonfoodconsumption_bl.dta c_c_nonfoodconsumption_fl1.dta	HH expenditures on non-food items (monthly and annually).
6	d_a_transfergiven_bl.dta d_a_transfergiven_fl1.dta	Transfers given by the HH to non-HH members, including type of item transferred, value of item, and item recipient.
7	d_b_transferrecd_bl.dta d_b_transferrecd_fl1.dta	Transfers received by the HH from non-HH members, including type of item received, frequency of transfer, value of item transferred, and source of transfer.
8	g_shocks_bl.dta g_shocks_fl1.dta	Shocks experienced by the HH (past 12 months), including type of shock, HH coping strategy, and monetary value of the shock.
9	k_a_livestock_bl.dta k_a_livestock_fl1.dta	Livestock ownership, value of livestock owned, expenditures on livestock purchases and rearing inputs, by livestock type.
10	k_b_animalproduct_bl.dta k_b_animalproduct_fl1.dta	Animal products production, including volume produced, volume sold, and earnings from sales of animal products. All animals owned by the HH, including TUP and non-TUP livestock.
11	k_c_crops_bl.dta k_c_crops_fl1.dta	Cultivation of crops.
12	k_d_nonagbiz_bl.dta k_d_nonagbiz_fl1.dta	Non-agricultural businesses.

#	Dataset	Description
13	k_e_timeuse_bl.dta k_e_timeuse_fl1.dta	Labor and time use of HH members.[1]
14	l_hhassets_bl.dta l_hhassets_fl1.dta	HH asset ownership, including asset type, number of assets or size of land (if HH owns land), ownership status within the HH, value of asset.
15	m_b_loans_bl.dta m_b_loans_fl1.dta	Number of outstanding loans, source of loan, HH member who took out the loan, intended purpose of the loan, and loan amount.
16	r_b_animalproduct_tup_fl1.dta	Animal products production, including volume produced, volume sold, and earnings from sales of animal products. Includes production from TUP livestock only.
17	v_mobile_fl1.dta	Phone ownership, including primary line owner (HH member), network carrier, length of line ownership.
18	market_clean_bl.dta market_clean_fl1.dta	Market survey.
19	village_clean_bl.dta	Village survey.

To facilitate cleaning and analysis, variables with unit of observation more granular than a household (i.e., where data were collected as a roster, such as characteristics of individual household members, consumption data on specific food or no-food items, information on individual businesses and loans, etc.) have been organized in separate "roster" datasets. These variables are not duplicated in "all_HH_nopii_fl1.dta". Hence, to use these variables, one should turn to the respective dataset as outlined in the table above as well as the Codebooks.

Roster datasets are named using the following convention: *module_section_description_wave.dta.* For example, the dataset "d_a_transfergiven_fl1.dta" contains data from module D, section A, with information on transfers given by the household, from follow-up 1. Sometimes it is not possible to neatly identify a section, especially when a roster dataset combines variables from several modules. For example, "br_hhmemcurrent_fl1.dta" contains data from modules B and R.

De-identification

All personally identifiable information (PII), including GPS coordinates, names, and phone numbers has been removed. District (a5) and Village (a6) variables are anonymized by dropping associated value labels. Textual variables in Dari/Pashto for "Other, please specify" responses and variable *q7* are also removed as they potentially contain PII.

Missing values

[1] Note that in Baseline, *k1_1- k4_1* are asked of both male and female household heads and are included in the roster without the "_m" suffix. In Follow-up 1, *k1_1- k4_1* are only asked of the *male* household head and included in the roster with a "_m" suffix.

Missing value codes used throughout the surveys (-98 = Don't know, -99 = Refused, - 97 = Other) are re-coded using Stata's extended missing values as detailed in the project's Github Wiki. Additionally, extended missing value ".n" is used to recode "-97".

A number of variables with ".99" and ".98" mistake entries (that should have been "-99" and "-98" missing value codes) are recoded to missing values.

Outliers

Published data is cleaned of gross outliers attributed to data entry errors (i.e., when the units have been misinterpreted, or where zeros have been added). We identify these observations typically by graphing the data and by using other variables to run consistency checks. This step entails a certain level of judgment calls. Only a small number of observations are impacted. Identified outliers are then put to missing by replacing their value with extended missing value ".o".

For a detailed description of our treatment of outliers in the cleaning and analysis phases, please see the project's Github Wiki.

Other data processing notes

Baseline

- In Baseline, MHHH survey responses are used for male-headed HHs in productive activities Module K, sections a) Livestock, b) Animal products, c) Crops, d) Non-agricultural businesses, e) Respondent time use (*k1_1 - k4_1)* and J (Entrepreneurship), since these questions were administered in the MHHH survey only (if the head of the household was male). These variables are identified in the Codebook (see column "MHHH response used") and are not duplicated in the data with an "_m" suffix since they simply replace the LHH variables (without the suffix).[2] Variable *respondent_productive_act* in the data also indicates the respondent of the relevant sections (1 = MHHH; 2 = LHH).
- MHHH survey responses are prioritized over (i.e., replace) LHH responses in Module M (Credit and savings) when both MHHH and LHH responses are available, and are not duplicated in the data with an "_m" suffix. LHH responses in Module M are preserved in variables with an "_f" suffix.
- In Respondent time use section of Module K of the Baseline LHH survey, variables *k1_1 - k4_1* correspond to the household head (member 1 in the HH roster). *k30, k30_1, k30_2* variables in the LHH survey were collected only for HH members other than the household head, and information from *k1_1 , k3_1, k4_1* has been replaced for the HH head in *k30, k30_1, k30_2* variables in the time use roster ("k_e_timeuse_bl.dta").
- *k30_2* is combined with *k30_1*, such there is only one monetary value in the roster ("k_e_timeuse_bl.dta").
- The data includes a handful of SurveyCTO (data collection platform)-generated variables that are useful for analysis that do not correspond to an item in the questionnaire. These include

[2]240 UP HHs with a female HH head filled out both LHH and MHHH surveys in error, in which case response from the LHH survey is prioritized. Additionally, 46 UP HHs with a male HH that should have filled out a MHHH survey did not do so, hence 46 UP HHs are missing responses to these modules. These HHs are identified in the *MHHH_missing* variable.

respondent_gender (gender of the respondent), *hh_head_gender* (gender of the HH head), *status, age, position* (HH member position in roster), *schoolyear* (schooling level), and *literacy* of the Primary Man ("_pm" suffix) and Primary Female ("_pf" suffix), based on the roster information.

- The data also includes relevant variables from implementation data (e.g., PRA lottery village, wealth ranking, survey weights).

Follow-up 1

- We similarly use responses of the MHHH survey in productive activities Module K, sections a) Livestock, b) Animal products, c) Crops, and d) Non-agricultural businesses for households with a male head. Specific variables for which such replacement has been made are indicated in the Codebook (see column "MHHH response used"). Variable *respondent_productive_act* in the data indicates the respondent of the relevant sections (1 = MHHH; 2 = LHH).
- *k30_2* is combined with *k30_1*, such there is only one monetary value in the roster ("k_e_timeuse_fl1.dta").
- In a few instances where this time of saliva biospecimen collection (*a18, a18_m*) is inconsistent with the time of the survey, we replace it with the SCTO-generated time of biospecimen collection (*time_collection* and *time_collection_m*).
- The data includes a handful of SurveyCTO (data collection platform)-generated variables that are useful for analysis that do not correspond to an item in the questionnaire. These include *hhh_name2* (roster position of the Male HH head), *lady_hh2* (roster position of the Lady of the HH), *respondent_gender* (gender of the respondent), *hh_head_gender* (gender of the HH head), *time_collection* (time of saliva sample collection).
- The data also includes relevant variables from implementation data (e.g., PRA lottery village, wealth ranking, survey weights).