2018 Sierra Leone Integrated Household Survey
Survey Documentation

Sample Design

There is a two stage sample design, with census enumeration areas (from the 2015 census) selected in the first stage, and then 10 households within each EA selected in the second stage. The first stage was stratified using district (the 14 districts as of the 2015 census) and urban vs. rural yielding 27 strata (there are no rural areas in the Western Area Urban district). The number of EAs per strata was chosen to provide efficient estimates of poverty rates at the strata level. Within each strata, the EAs were selected using probability proportional size, based on the estimated number of households during the census mapping.

The second stage selection of households happened in two different ways: A subsample of the selected SLIHS EAs were used for the MICS survey as well. The MICS conducted fieldwork between May and August 2017. For these EAs, 10 target households and 5 replacements were randomly selected for the SLIHS from those actually interviewed by the MICS teams. For the other EAs, a household listing was done between October and December 2017, and from this, 10 target households and 5 replacements were selected. The complete planned sample size was achieved: in each EA, ten households were interviewed and completed the majority of the questionnaire.

The data was weighted based on the sampling design, and post-stratification to chiefdom-level household population levels (the probability weight variable *_pweight*) is included in each dataset. More information on weighting can be found in the methodology note.

Field Procedures and Quality Control

Nineteen teams worked over a 12-month period from January to December 2018. Each team consisted of three interviewers, a data entry clerk and a supervisor. The team covered three EAs per month, with one interviewer being responsible for one EA, and (in the rural areas) residing within the EA for the month. As each questionnaire book is completed, it was brought by the supervisor to the data entry clerk, who entered it using a CSPro data entry application on a laptop. The application did extensive error reporting and the supervisor reverted to the interviewer / household to correct any errors / inconsistencies identified. The teams spent approximately three weeks in the EA, and then one week in the district headquarters. Internet connectivity was provided via cellular modems with wide coverage throughout the country. Data was automatically uploaded / backed up whenever the data entry laptop was online. In areas without coverage, data was uploaded during the week the team is at the district headquarters. Data uploaded was reviewed by headquarters with additional feedback and instructions provided to teams.

Guidelines for Using the Data

The Dataset Organization section below explains how different sections of the questionnaire are mapped to different datasets. All datasets contain the variables *_cluster* and *_hhno* which can be used to merge them (except **slihs2018_cluster.dta** which is cluster-level data and has just *_cluster*.)

The data entry system imposed tight controls on what values could be entered, and flagged additional possible inconsistencies, resulting in fairly clean data coming in. A light approach to data cleaning was thus taken. For the most part, the only replacement of outliers was done by replacing values more than twice the 99th percentile with twice the 99th percentile. Note that this was not done when two questions (say amount paid and period for which it was paid) need to be combined to give a total value. The data from the food diaries (Section X) required more extensive cleaning, and there are still issues, particularly when giving the unit cost for items from own production or received for free. When using this data in particular, check for outliers and mistakes.

For some questions, common "other" responses were identified and codes added. These are (list possibly incomplete): *b2, b24, b25, d4a, d4b, d4c, d13a, d13b, d26a, f22, o8, o14, p4*

Dates are stored in Stata elapsed data format, months are stored in Stata elapsed month format.

There are some known issues with the data:

- Question *b9f* (class attended Sept 2018 to Dec 2018). This question should have been answered in households interviewed during the last 4 months of fieldwork. Somehow, this question was not consistently answered and entered for those 4 months, the data seems to reflect a large drop in gross enrollment which is not reflected by any other source (including following up with these households by phone later).
- Questions *e25* and *e27* (child height and weight). This was clearly not done correctly by Teams 1, 3, 15, 16 and 18.
- Question *r25* (measurement of plot in acres). Teams 10 and 17 clearly did not do this correctly. Teams 1 and 2 (measuring just a few likely very small plots in the capital city) also likely have errors.

(note that the variable *_team* can be merged in from **slihs2018_cluster.dta** using the variable *_cluster* to merge.)

Collaboration with the MICS

The 2018 SLIHS and the MICS-6 cooperated on the selection of the sample of EAs and households so that over 5000 households are covered as part of both surveys, and the data merge-able on the household and individual level. The MICS covered 507[1] of the 684 EAs sampled for the SLIHS (referred to as shared EAs). The SLIHS datasets are mergeable with the MICS datasets at the household and individual level as follows:

- **slihs2018_cluster.dta** contains the variable *_mics_cluster_no* (which can be merged into another SLIHS dataset using the variable *_cluster*) for all shared EAs (it is missing for SLIHS-only EAs). *_mics_cluster_no* corresponds to *HH1* in the MICS-6 datasets.
- Each SLIHS dataset contains the variable *_hhno* which corresponds to *HH2* in the MICS-6 datasets for EAs that were covered by both surveys. *HH2* is between 1 and 26. There are some SLIHS households in shared EAs with values of *_hhno* between 90 and 99. These are households that were interviewed by the SLIHS team in the shared EA, but do not actually seem to be the target household that was interviewed by the MICS team.
- **slihs2018_ind.dta** is all the individual level data for the SLIHS. It has a variable *a18* which corresponds to *LN* in the MICS-6 datasets.

The data has been carefully cleaned so that any household / individual linked by these IDs is very likely to be actually the same household / individual.

[1] It should have been 508, but there was one EA where the SLIHS team was unable to find any of the target households, either the MICS or the SLIHS team must have gone to the wrong EA.

Dataset Organization

The questionnaire for the 2018 is divided into 5 books:

Book 1: Household Member Characteristics

Section A: Household Roster
Section B: General Education
Section C: Alternative Education and ICT
Section D: General Health and Disability
Section E: Child Preventative Health
Section F: Women's Reproductive Health
Section G: Health Knowledge, Behaviors and Attitudes
Section H: Employment and Time Use - 7 Days
Section I: Employment and Time Use - 12 Months
Section J: Migration

Book 2: Household Characteristics

Section K: Housing
Section L: Durable Goods
Section M: Social Assistance and Subjective Wellbeing
Section N: Non-Food Consumption (Infrequently Purchased Items)
Section O: Financial Services
Section P: Non-Farm Enterprises

Book 3: Agriculture

Section R: Agricultural Assets
Section S: Annual Crops
Section T: Permanent Crops
Section U: Forestry Activities
Section V: Fishing
Section W: Livestock

Book 4: Consumption

Section X: Food Consumption
Section Y: Non-Food Consumption (Frequently Purchased Items)

[Book 4A: secs X and Y for the first 10 days and sec Z; Book 4B: X and Y for days 11 to 20]

Section Z: Bulk Purchases

The different sections / questions are mapped into the different datasets as shown on the next page.

All datasets contain:

- ID variables (_cluster, _hhno and a third (and fourth) id if needed such as _ind or _line).
- variables for survey data analysis: _cluster (psu), _stratum and _pweight
- serial numbers of the questionnaire books from which they take data.
- In general, variables whose names start with an underscore are administrative variables.
- Other variable name correspond directly to the section and question number on the questionnaire: b2 is section B, question 2. Multipart questions (first, second and third reasons etc) and usually postfixed a, b, c etc (for example, d4a, d4b, d4c). When the "other" option was answered for a question, the text given to specify is the variable with the postfix _other (for example, d4a_other).

Datasets can be merged using the variable *_cluster* and *_hhno* (or just *_cluster* if merging with **slihs2018_cluster.dta**)

Book	Section	Questions	Dataset
cover page and administrative variables			slihs2018_cluster.dta
consumption aggregates and poverty lines			slihs2018_consexp.dta
1	all		slihs2018_ind.dta
2	K	1-41, 48-53	slihs2018_hh.dta
2	K	42-47	slihs2018_k42.dta
2	L	all	slihs2018_l.dta
2	M	2-8	slihs2018_m2.dta
2	M	9-20	slihs2018_hh.dta
2	M	21-23	slihs2018_m21.dta
2	M	24-26	slihs2018_m24.dta
2	N	all	slihs2018_n.dta
2	O	1, 2, 11, 15, 23, 30, 40	slihs2018_hh.dta
2	O	3-10, 12-14, 16-22, 24-29, 31-39	slihs2018_o.dta *
2	O	41-48	slihs2018_o41.dta
2	O	49-54	slihs2018_o49.dta
2	P	all	slihs2018_p.dta
2	Q	all but Q1	slihs2018_q.dta
3	R	1-7	slish2018_r1.dta
3	R	9-24	slihs2018_r9.dta
3	S	1-42	slihs2018_s.dta
3	S	43-44	slihs2018_hh.dta
3	T	1-39	slihs2018_t.dta
3	T	40-41	slihs2018_hh.dta
3	U	all	slihs2018_u.dta
3	V	1-8, 29	slihs2018_hh.dta
3	V	9-17, 18-22, 23-26	slihs2018_v9.dta *
3	V	30-40	slihs2018_v30.dta
3	W	all but W0	slihs2018_w.dta
2-4	misc	P0, Q1, R8, V1, V0, W0, Z0	slihs2018_hh.dta
4A/B	X	all	slihs2018_x.dta
4A/B	Y	all	slihs2018_y.dta
4B	Z	all but Z0	slihs2018_z.dta

* Note that datasets **slihs2018_o.dta** and **slihs2018_v9.dta** have data from multiple parallel tables.