

## The World Bank Listening to LAC (L2L) Pilot Project Sample Design for Peru

### Background

The Listening to LAC pilot program is a research exercise aimed at testing the feasibility of the SMS technology as a data collection method for conducting quick turnaround, self-administered, longitudinal surveys among households in Latin America.

The purpose of this document is to provide a complete and comprehensive description of the procedure used to draw the sample for Listening to LAC project using the fourth trimester of ENAHO 2010 as a sampling frame.

### Sample Design Premises

The sample design was guided by the following criteria:

1. The sample should be nationally representative including both, urban and rural areas;
2. Households close to the poverty line should be oversampled. For the purposes of this project, “close to poverty line” is defined as 40 percent of income distribution that symmetrically band the national poverty line: 20% above and 20% below. In 2010 in 27% of Peruvian households monthly per capita consumption was below the moderate poverty line (ENAHO 2010)<sup>1</sup>. Consequently, the households, where monthly per capita consumption falls between 7 and 47 percent of national distribution, have been oversampled.

In order to track potential errors and ensure that the aforementioned requirements are satisfied, the sampling exercise consisted of drawing and validating 2 samples: with and without oversampling.

### ENAHO Sample Design

As the fourth trimester of ENAHO 2010 is used as a sampling frame, a brief description of ENAHO sampling design is in order.

ENAHO sample is selected in three stages in urban and in more densely populated rural areas and in two stages in less densely populated areas. In the first stage, selection of the Primary Sampling Units (PSUs) occurs. All PSUs are grouped in 8 strata, defined by the size (number of residents). Strata 1 through 5 correspond to urban areas, strata 6 through 8 correspond to rural areas. The selection occurs within department-specific strata, and the probability of selection of PSUs is proportional to the number of households in them.

---

<sup>1</sup> The present calculation is made based on INEI data available in April 2011, when INEI was still finalizing poverty calculation methodology for 2011. Consequently there may be a discrepancy between this calculation and official poverty number to be released in June 2011.

The second stage depends on which of the two large categories a PSU belongs to. The first category is comprised of urban PSUs and more populous rural PSUs. The second category consists of less populous rural PSUs. For simplicity, this document refers to the first category as “urban” and the second as “rural”.

In urban category, Secondary Sampling Units – *conglomerados* - are selected from each PSU with the probability proportional to their size, and with implicit stratification. Implicit stratification is based on a number of socio-economic variables. In rural category PSUs consist of one *conglomerado* (SSU) only. For these PSUs, the second stage is omitted.

At the third stage, households are randomly selected from SSUs. In rural areas 8 households are randomly drawn from each SSU; in urban areas 6 households per SSU are selected.

The ENAHO survey includes a panel component. The SSUs (conglomerados) from which panel households are drawn are randomly selected. In order to be able to compare L2L data to the future rounds of ENAHO at the SSU level, L2L sample will be drawn from panel conglomerados only.

## **L2L sampling procedure**

### **Notes on materials from INEI, creation of additional variables, etc.**

In order to carry out the sampling procedure, we took advantage of the document used by INEI to draw the 2010 ENAHO sample (hereafter referred to as INEI sampling table, or the sampling table). Each line of the sampling table corresponds to a Secondary Sampling Unit (SSU), or conglomerado. The table is limited to conglomerados included in ENAHO 2010 (3406 conglomerados). The table also includes the strata and Primary Sampling Unit identifiers (PSU), to which conglomerados belong, and the information necessary to derive household sampling weights: probability of selection of PSUs, SSUs and the household. The numbers of dwellings in the strata, PSUs and SSUs, necessary to derive these probabilities are also available for each conglomerado.

There is no PSU identifier in the original table: “upmfinal” is a unique PSU identifier within departamento, but not in the entire dataset. Therefore, a unique PSU identifier is created by concatenating departamento identifier and “upmfinal”.

### **Drawing L2L sample**

There are multiple ways to oversample households 20 percent above and 20 percent below. This document describes the implementation of an oversample of PSUs with higher proportions of households of interest. The number of households drawn from each SSU were the same as in ENAHO (6 households per urban SSU and 8 per rural SSU).

## Sampling exercise #2

(1) PSU selection. ENAHO 2010 4th trimester panel sample includes 3,782 households. 1,500 of L2L sample correspond to 40 percent of it. Additional stratification (based on the proportion of households from 7<sup>th</sup> to 47<sup>th</sup> income percentile). All PSUs were divided in two strata: PSUs where the proportion of households from 7<sup>th</sup> to 47<sup>th</sup> percentiles of the income distribution constitute the majority – stratum I, and PSUs where this proportion is less than 50 percent – stratum II. These two strata were treated as separate samples; subsequent steps (2) through (X) were applied to each stratum. There are 181 and 391 PSUs in stratum I and stratum II, respectively.

(2) PSU selection. 60 percent of the L2L sample (900 households) were drawn from stratum I and 40 percent (600 households) from stratum II. As the table below shows, there are approximately 1,346 households in stratum I and 2,436 households in stratum II. Consequently, 67%<sup>2</sup> of households from stratum I and 25% of households from stratum II were selected. PSUs were drawn with probability proportional to size from each of 8 ENAHO strata.

Table: L2L strata (based on fraction of hh in 7-47 income percentile band)

ENAHO strata	L2L stratum II		L2L stratum I	
	number of conglomerados	number of hh	number of conglomerados	number of hh
1	109	654	0	0
2	115	690	4	24
3	46	276	3	18
4	36	216	18	108
5	40	240	26	156
6	9	72	15	120
7	19	152	79	632
8	17	136	36	288
Total		2,436		1346

(3) SSU selection – urban areas. The number of SSUs (conglomerados) per PSU in urban strata varies highly: from 1 to 38. We drew 940 households<sup>3</sup>, treating all urban PSUs as a single stratum, according to probability proportional to the size of SSU. Before making the draws, we sorted all SSUs according to socio-economic index, as in the ENAHO sampling, to ensure implicit stratification along the socio-economic dimension.

(4) Household selection: 6 households were randomly selected from each urban SSU, and 8 households from each rural SSU.

<sup>2</sup>  $0.6 = 900/1346$ ;  $0.25 = 600/2,436$

<sup>3</sup>  $940 = 1500 - 560$ ;  $1,500 - \text{L2L sample size, } 616 - \text{number of rural households selected in (2)}$