HONDURAS

Beyond Connections

ANNEX 2. SAMPLING STRATEGY

SAMPLE SIZE CALCULATION PARAMETERS

The sample size proposed for the MTF countries is designed to get sufficiently precise estimates of each tier at national, urban, and rural levels. A much smaller sample size would have been adequate to produce precise estimates at the national level within those domains. This section discusses the factors to consider in determining sample size calculation and provides a justification for the proposed sample size for each country. Major issues in determining the appropriate sample size for a survey are the following:

- Precision of survey estimates (sampling error)
- Quality of data collected by the survey (nonsampling error)
- Cost in time and money of data collection, processing, and dissemination

Precision of survey estimates. The concept of the precision of a sample survey estimate is crucial in determining the sample size. By definition, a sample from a population is not a complete picture of the population. However, an appropriately drawn random sample of reasonable size can provide a clear picture of the characteristics of that population, certainly sufficient for policy implication or decision-making purposes. From a sample of households, one can collect data and generate a sample (or survey) estimate of a population parameter. The population parameter value of a characteristics of interest is generally unknown. Sampling errors (or margin of errors) depend very much on the size of the sample, and very little on the size of the population. To maximize the sample size and to reduce the sampling error, the prevalence rate in this calculation is 50%. The formula (B.1) to calculate the sample size is as follows:

$$n = \frac{z^2 r(1-r) fk}{e^2} = \frac{z^2 r(1-r)[1+\rho(m-1)]k}{e^2} \qquad \text{(B.1)}$$

where:

n = Sample size to be determined.

z = z-statistics corresponding to the level of confidence. The commonly used level of confidence is 95%, for which z is 1.96.

r = Estimate of the indicator of interest (50%).

f = Sample design effect. This represents how much larger the squared standard error of a two-stage sample is when compared with the squared standard error of a simple random sample of the same size. Its default value for infrastructure interventions is 2.0 or higher, which should be used unless there is supporting empirical data from similar surveys that suggest a different value. The sample design effect has been included in the sample size calculation formula (B.1) and is defined as: $f = 1 + \rho\ (m - 1)$.

ρ = Intracluster correlation coefficient. This is a number that measures the tendency of households within the same primary sampling unit (PSU) to behave alike regarding the variable of interest. ρ is almost always positive, normally ranging from 0 (no intracluster correlation) to 1 (when all households in the same PSU are exactly alike). For many variables of interest in Living Standards Measurement Study (LSMS) surveys, ρ ranges from 0.01 to 0.10, but it can be 0.5 or larger for infrastructure-related variables.

m = Average number of households selected per PSU.

k = Factor accounting for nonresponse. Households are not selected using replacement. Thus, the final number of households interviewed will be slightly less that the original sample size eligible for interviewing. The sample size should be calculated to reflect the experience from the country in question. For most developing countries, the nonresponse rate is typically 10% or less. Therefore, a value of 1.1 (= 1 + 10%) for k would be conservative.

e = Margin of error, or level of precision. We apply various levels of margin of error from 1% to 5.5% to the calculation.

Quality of data (nonsampling error). Beside sampling errors, data from a household survey are vulnerable to other inaccuracies from causes as diverse as refusals, respondent fatigue, measurement errors, interviewer errors, or the lack of an adequate sample frame. These are collectively known as nonsampling errors. Nonsampling errors are harder to predict and quantify than sampling errors, but it is well accepted that good planning, management, and supervision of field operations are the most effective ways to keep them under control. Moreover, it is likely that management and supervision will be more difficult for larger samples than for smaller ones (Grosh and Muñoz 1996, 56). Thus, one would expect nonsampling errors to increase with sample size, and we would like to limit the sample size to less than 5,000.

Cost of data collection, processing, and dissemination. The sample size can affect the cost of the survey implementation dramatically. It will also affect the time in which the data can be collected, processed, and made available for analysis. The availability of survey firm and cost for each country would affect the total cost of survey implementation, too. Thus, the cost of data collection, processing, and dissemination should be considered in determining the sample size for each country.

SAMPLE SIZE CALCULATION

The sample size for the country was calculated using the prevalence rate of 50% as the most conservative choice and to achieve the minimum margin of error. A nonresponse rate of 10% and a value of 1.1 for nonresponse rate were considered. The number of households selected per PSU was 12. Due to the characteristics of infrastructure variables and indicator, 0.45 for intracluster relation coefficient was selected, which determined a design effect equal to 6.

In the process of defining a strategy to calculate the sample size for the selected countries, the sample size was calculated using the distribution between urban and rural as two analytic domains. Then these two values were added to obtain the national sample size. Following this approach, a margin of error of 6.2% at urban and rural levels gives a national sample size of 3,324 households with an error of 4.4%, of which 1,656 households are urban and 1,668 households are rural.

Thus, the household survey sample selection is based on a three-stage stratification strategy, aimed at being representative of both urban and rural populations.[1] The survey was implemented in 16 out of 18 departments, following the same approach adopted by the National Institute of Statistics (INE) to produce the Master Sample. The MTF Sample Framework is based on the Master Sample based on the last Population and Housing Census, prepared by the INE in 2013 (INE 2015).

The departments of Gracias a Dios and Islas de la Bahía are excluded from the survey, and they are not even included in the master sample and official surveys of INE. The reasons include the following. First is the low population weight—both departments represent less than 1% of the total national census segments of INE. Second, there is a high relative cost of access and internal mobilization. The two departments are accessible only by plane: Islas de la Bahía is a tourist destination, and in Gracias a Dios, the cayucos and small boats are the only means of transport. Third, part of the population in Gracias a Dios communicate in a local dialect (Miskito), and in Islas de la Bahía, only a minor part of the population speaks English.

Finally, the electrification conditions in both departments are peculiar. In Islas de la Bahía, energy has a thermal source, produced by a private company that is in the process of becoming managed by the municipality. Since fuel and other inputs are carried from the mainland, production costs and, therefore, energy prices are particularly high. Gracias a Dios has only a few electrified communities, and, given the limited communication channels and the dispersion of households throughout the jungle area, the costs of bringing energy in this area are high. For these reasons, both departments are usually excluded from the household surveys in the country, including official surveys implemented by INE.

INE provided advice on the sampling strategy and supported the MTF team in identifying the electrification status of the EAs, identified as enumeration areas, sectors, or PSUs.[2] The team surveyed 3,324 households in 276 EAs, equally split between urban and rural areas, in 16 departments, following the stratification criteria: 50–50 ratio of electrified and nonelectrified households for the tier analysis and equal allocation between urban and rural areas.
Even though the original sample size was as planned (3,324 households, of which 1,668 are in rural and 1,656, urban), the final sample size was reduced considerably due to the high nonresponse rate. The actual sample size is 2,815 households, of which 1,574 are rural and 1,241, urban. The reduction of the sample size was due to (i) the exclusion of two departments from the survey in line with INE's strategy and (ii) the high nonresponse rate in urban areas because of safety issues, which may partially affect the reliability of the results and the standard errors. The outcomes of the MTF analysis, thus, have to be read considering these limitations. The sample of electrified and nonelectrified sectors was drawn in a particular way by the firm, given that most of the population in Honduras is connected to the grid and only a few segments are not electrified.

[1] The political division of Honduras is made up of 18 departments,298 municipalities, 3,714 villages, and 29,950 *caseríos* (409 urban and 29,541 rural).
[2] Within each sector the firm selected a segment.

In rural areas, to maintain the department structure of the Master Sample, the sample of sectors was distributed proportionally to the number of sectors by department (excluding Gracias a Dios and Islas de la Bahía). To have a larger number of sectors with no access to the national grid, the firm oversampled the substratum of sectors, with some segments with no access the national grid, as shown in Table B.1.

Table B.1 Rural area, distribution of the MTF sample of sectors for access to the national grid, by department

Department	Sectors with *some* segments with *no access* to national grid[a]	Sectors with *all* segments with *access* to national grid[a]	Total
1 Atlántida	2	3	5
2 Colón	3	3	6
3 Comayagua	4	5	9
4 Copán	4	5	9
5 Cortes	5	7	12
6 Choluteca	4	5	9
7 El Paraíso	5	6	11
8 Francisco Morazán	7	7	14
10 Intibucá	3	4	7
12 La Paz	2	3	5
13 Lempira	5	5	10
14 Ocotepeque	2	2	4
15 Olancho	6	6	12
16 Santa Bárbara	5	6	11
17 Valle	2	2	4
18 Yoro	5	5	10
Total	64	74	138

a. Segments that have 3% or more of households connected to the national grid are classified as having "access to the grid"; otherwise, they are classified as "no access to the grid."

To maintain the structure of the Master Sample for urban areas, the sample of sectors was distributed proportionally to the number of sectors by strata resulting from matching geographical domains and departments. The firm oversampled subsection of sectors with a segment that had a percentage lower than 97% of the households connected to the national grid, according to the Population and Housing Census of 2013 (INE 2015) (Table B.2.)

Table B.2 Urban area, distribution of the MTF sample of sectors by domain and access to the national grid, by department

Department	Central district (substratum)		San Pedro Sula (substratum)		Median cities (substratum)		Small cities (substratum)		Total
	1[a]	2[b]	1[a]	2[b]	1[a]	2[b]	1[a]	2[b]	
Atlántida	0	0	0	0	5	1	1	1	8
Colon	0	0	0	0	1	0	2	1	4
Comayagua	0	0	0	0	3	1	1	1	6
Copan	0	0	0	0	1	0	1	1	3
Cortes	0	0	27	3	9	1	8	1	49
Choluteca	0	0	0	0	1	1	1	1	4
El Paraíso	0	0	0	0	1	0	1	1	3
Francisco Morazán	30	4	0	0	0	0	3	1	38
Intibucá	0	0	0	0	0	0	1	1	2
La Paz	0	0	0	0	0	0	1	1	2
Lempira	0	0	0	0	0	0	1	0	1
Ocotepeque	0	0	0	0	0	0	1	0	1
Olancho	0	0	0	0	2	1	1	0	4
Santa Bárbara	0	0	0	0	0	0	3	1	4
Valle	0	0	0	0	1	0	1	0	2
Yoro	0	0	0	0	3	1	2	1	7
Total	30	4	27	3	27	6	29	12	138

a. Sectors with *some* segments with <97% of the households *connected* to the national grid.

b. Sectors with *all* the segments with >=97% of the households *connected* to the national grid.

SELECTION OF SECTORS (PSUs) AND SEGMENTS

Within each selection stratum, the number of sectors was selected systematically, with random start, subject to geographic ordering, to have an implicit substratification in line with the indicated strategy (Tables B.1 and B.2). Within each sector, a census segment was selected with simple random sampling, so that in the MTF sample, the number of census segments is equal to the number of sectors. Given that the segment was randomly selected within each sector, the distribution of the sample of segments by the substrata of access to the grid is random.

The segments in the MTF sample are classified according to the degree of grid access in 2013, when the Population and Housing Census was carried out (INE 2015). The distribution is presented in Table B.3 for rural areas and in Table B.4 for urban areas.

Table B.3 Rural area, distribution of segments in selected sample by degree of access to the national grid in 2013, by department

Department	Access rate of segment to the national grid				Total segments in rural stratum
	Segments with <3% connected households	Segments with 3% to <50% connected households	Segments with 50% to <97% connected households	Segments with >=97% connected households	
Atlántida	0	2	2	1	5
Colon	3	1	2	0	6
Comayagua	2	3	3	1	9
Copan	1	2	4	2	9
Cortes	1	1	7	3	12
Choluteca	2	3	4	0	9
El Paraíso	2	2	6	1	11
Francisco Morazán	3	4	7	0	14
Intibucá	2	1	3	1	7
La Paz	1	3	1	0	5
Lempira	3	1	5	1	10
Ocotepeque	2	0	2	0	4
Olancho	2	3	7	0	12
Santa Bárbara	2	4	4	1	11
Valle	2	1	1	0	4
Yoro	4	1	3	2	10
Total	32	32	61	13	138

Table B.4 Urban area, distribution of segments in selected sample by degree of access to the national grid in 2013, by department

Department	Access rate of segment to the national grid				Total segments in urban stratum
	Segments with <3% connected households	Segments with 3% to <50% connected households	Segments with 50% to <97% connected households[a]	Segments with >=97% connected households	
Atlántida	0	0	1	7	8
Colon	0	0	3	1	4
Comayagua	0	0	2	4	6
Copan	0	0	1	2	3
Cortes	1	1	15	32	49
Choluteca	0	0	1	3	4
El Paraíso	0	0	1	2	3
Francisco Morazán	0	1	17	20	38
Intibucá	0	0	0	2	2
La Paz	0	0	1	1	2
Lempira	0	0	0	1	1
Ocotepeque	0	0	0	1	1
Olancho	0	0	1	3	4
Santa Bárbara	0	0	2	2	4
Valle	0	0	2	0	2
Yoro	0	0	3	4	7
Total	1	2	50	85	138

a. Only 6 segments had <80% of the households connected to the national grid in 2013.

SELECTION OF HOUSEHOLDS

PROCEDURE IN RURAL AREAS

To capture a larger number of households not connected under the restrictions of the sample size calculated for the rural area (1,656 households) and the degree of accessibility that the segments of the MTF sample had to the grid in 2013, the firm oversampled the segments with the lowest percentage of households connected to the grid, in the following ways. (Table B.5 shows a summary of the number of dwellings interviewed in rural areas.)

Segments with less than 50% of the households connected to the network. In each segment of the sample, four starting points were selected, each one defining a subcluster of four households each, for a total of 16 households per segment. Bearing in mind the total number of households in the sample frame, four numbers (R1, R2, R3, R4) were randomly selected with random start in a systematic way.

These random numbers identified four subclusters of four households each. The first random number, R1, determined the starting point of the first subcluster. The starting point was the household in which R1 households accumulated. Following the numbering in the cartography, a total of four eligible households (i.e., occupied private dwelling) for the subcluster were selected. The same was done for the other three random numbers to identify the 16 eligible households in the segment, not necessarily physically adjacent to each other.

For segments with 50% to less than 97% of the households connected to the network. The firm selected two starting points, defining two subclusters of four households each, for a total of eight households per segment.

For segments that have 97% or more of the households connected to the network. Three starting points were selected, defining three subclusters of four households each, for a total of 12 households per segment.

Table B.5 Rural area, distribution of households in sample by degree of access to the national grid in 2013, by department

Department	Access rate of segment to the national grid				Total segments in rural stratum
	Segments with <3% connected households	Segments with 3% to <50% connected households	Segments with 50% to <97% connected households	Segments with >=97% connected households	
Atlántida	0	32	16	12	60
Colon	48	16	16	0	80
Comayagua	32	48	24	12	116
Copan	16	32	32	24	104
Cortes	16	16	56	36	124
Choluteca	32	48	32	0	112
El Paraíso	32	32	48	12	124
Francisco Morazán	48	64	56	0	168
Intibucá	32	16	24	12	84
La Paz	16	48	8	0	72
Lempira	48	16	40	12	116
Ocotepeque	32	0	16	0	48
Olancho	32	48	56	0	136
Santa Bárbara	32	64	32	12	140
Valle	32	16	8	0	56
Yoro	64	16	24	24	128
Total	512	512	488	156	1,668

PROCEDURE IN URBAN AREAS

The households in all the segments of the sample in urban area were selected following the same procedure, which consists of selecting three random starting points to define three subclusters of four households each, for a total of 12 households by segment. Table B.6 shows a summary of the number of households interviewed in urban area.

Table B.6 Urban area, distribution of households in sample by degree of access to the national grid in 2013, by department

Department	Access rate of segment to the national grid				Total segments in urban stratum
	Segments with <3% connected households	Segments with 3% to <50% connected households	Segments with 50% to <97% connected households	Segments with >=97% connected households	
Atlántida	0	0	12	84	96
Colon	0	0	36	12	48
Comayagua	0	0	24	48	72
Copan	0	0	12	24	36
Cortes	12	12	180	384	588
Choluteca	0	0	12	36	48
El Paraíso	0	0	12	24	36
Francisco Morazán	0	12	204	240	456
Intibucá	0	0	0	24	24
La Paz	0	0	12	12	24
Lempira	0	0	0	12	12
Ocotepeque	0	0	0	12	12
Olancho	0	0	12	36	48
Santa Bárbara	0	0	24	24	48
Valle	0	0	24	0	24
Yoro	0	0	36	48	84
Total	12	24	600	1020	1,656

SAMPLE WEIGHTING CALCULATIONS

To have valid estimates of the parameters of the target population and adequately estimated sample errors, the sample design should be considered in its calculation. It is necessary to apply weights to the sample's results to correct or reduce biases that may be introduced by selection with unequal probabilities or nonsampling errors (for example, refusals). The sample design weight is calculated as the inverse of the selection probability (B.2):

$$w_i = \frac{1}{p}, \qquad \text{(B.2)}$$

Where:

p is the probability of a unit to be included in the sample.

The sample of households of the MTF survey was selected from the Master Sample of sectors (PSU) in three stages: in the first stage a subsample of sectors was selected from the Master Sample; in a second stage a segment was selected in each sector of the MTF sample; in the third stage the households were selected within the segments of the MTF sample. The final probability of selecting the households was calculated by multiplying the probability of selecting the sectors of the Master Sample by the conditional probabilities of each stage and the basic expansion factor as the inverse of the probability of selection. The following paragraphs explain the sample weighting calculation in more detail.

SELECTION PROBABILITIES OF SECTORS (PSUS) IN MASTER SAMPLE

INE provided the firm with the selection probabilities of the sectors (PSUs) in the Master Sample. These sectors were selected in a single stage, except for the small cities in the urban area, which were selected in a previous stage within each department. The probability of selecting sector i of stratum h in the Master Sample (except for small cities) is as shown (B.3):

$$p_{hi} = \frac{n_h * M_{hi}}{M_h} \qquad \text{(B.3)}$$

Where:

n_h: number of sectors selected in stratum h.
M_h: total number of households within stratum h.
M_{hi}: total number of households in sector i within stratum h.

In the case of small cities, the selection probability was calculated as follows (B.4):

$$p_{hgi} = \frac{c_h * M_{hg}}{M_h} \frac{n_{hg} * M_{hgi}}{M_{hg}} \qquad \text{(B.4)}$$

Where:

c_h: number of small cities in the sample within stratum h.
M_{hg}: number of households in the small city g within stratum h.
M_h: number of households of all small cities within stratum h.
n_{hg}: number of sectors in the sample within the g^{th} small city of stratum h.
M_{hgi}: number of households in sector i, small city g of stratum h.

SELECTION PROBABILITIES AND EXPANSION FACTORS IN THE MTF SAMPLE

According to the sample design implemented, the probability of selecting a household in segment j of sector i in the stratum h' of the MTF sample is given as shown (B.5):

$$p_{h'ij} = p_{hi} \frac{n_{h'}}{N_{h'}} \frac{1}{L_{h'i}} \frac{m_{h'ij}}{M_{h'ij}} \qquad \text{(B.5)}$$

p_{hi}: probability of selecting sector i of stratum h in the Master Sample in case that small cities is p_{hgi} instead of p_{hi}.
n_h: number of sectors selected in stratum h' in the sample.
$N_{h'}$: number of sectors in stratum h' of the sample frame of the MTF sample.
$L_{h'i}$: number of segments in sector i of stratum h' of the sample frame of the MTF sample.
$M_{h'ij}$: current number of households in the sample segment j of sector i of stratum h'.
$m_{h'ij}$: number of households selected within the j^{th} sampling segment of sector i of stratum h'; it is equal to 12 in segments in urban area; 16 in segments in rural areas with less than 50% of their households connected to the energy grid in 2013; 8 in segments in rural areas between 50% and 97% of connected households; 12 segments in rural areas with 97% or more connected households.

The basic expansion factor for all households selected in the j[th] segment of the i[th] sampling sector of stratum h' is equal to the inverse of their probability of selection as shown (B.6):

$$W_{h'ij} = \frac{1}{p_{h'ij}} \qquad \text{(B.6)}$$

An expansion factor was calculated for each sample segment. This basic expansion factor was adjusted by the eventual nonperformance of some segments and by the nonresponse at the household level.

If $n_{h'}$ is the number of segments selected and $n'_{h'}$ is the number of segments made in the stratum h' of the MTF sample, the response rate at the segment level in that stratum is given as shown (B.7):

$$T_{sh'} = \frac{n'_{h'}}{n_{h'}} \qquad \text{(B.7)}$$

The response rate at the household level in stratum h' to adjust the expansion factors is calculated as follows (B.8):

$$T_{hh'} = \frac{\sum W_{h'ij} m'_{h'ij}}{\sum W_{h'ij} m_{h'ij}} \qquad \text{(B.8)}$$

Where the sum is over all segments in stratum h'.
$W_{h'ij}$: basic expansion factor for all households selected in the j[th] segment of the i[th] sampling sector of stratum h'.
$m_{h'ij}$: number of households selected within the j[th] sampling segment of sector i of stratum h'.
$m'_{h'ij}$: number of households interviewed within the j[th] sampling segment of sector i of stratum h'.

Finally, the expansion factor adjusted for all households in the j[th] sampling segment of the i[th] sample sector of stratum h' is shown as follows (B.9):

$$W'_{h'ij} = \frac{W_{h'ij}}{T_{sh'} T_{hh'}} \qquad \text{(B.9)}$$