Sampling Strategy for Liberia

The sample size proposed for the selected country is designed to get sufficiently precise estimates of each tier at the national level as well as the zonal (urban and rural) level. This section, at first, presents a discussion on the factors that should be taken into consideration in the determination of sample size calculation (1.1) and provides a justification for the proposed sample size for the selected countries (1.2). Then, it explains the criteria for stratification process (1.3).

1.1. *Issues determining a survey's sample size*

The major issues considered to determine the appropriate sample size for a survey are:

1. The precision of the survey estimates (sampling error);
2. The quality of the data collected by the survey (non-sampling error); and
3. The cost in time and money of data collection, processing, and dissemination.

The following sub-sections discuss each of these issues in turn.

1. *The precision of the 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 it. 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 characteristics of interest is generally unknown.

The formula to calculate the sample size is:

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

where:

- n = Sample size in terms of number of households to be selected.
- z = z-statistics corresponding to the level of confidence desired. The commonly used level of confidence is 95% for which z is 1.96.
- r = Estimate of the indicator of interest to be measured by the survey.
- 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 (1) and is defined as: $f = 1 + \rho (m - 1)$.
- ρ = Intra-cluster correlation coefficient. This is a number that measures the tendency of households within the same Primary Sampling Unit (PSU) to behave alike in regards to the variable of interest. ρ is almost always positive, normally ranging from 0 (no intra-cluster correlation) to 1 (when all households in the same PSU are exactly alike). For

many variables of interest in 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 non-response. Households are not selected using replacement. [1] Thus, the final number of household interviewed will be slightly less that the original sample size eligible for interviewing. For most developing countries, the non-response rate is typically 10% or less. So, a value of 1.1 (= 1 + 10%) for k would be conservative.
- e = Margin of error, sampling errors or level of precision. It depends very much on the size of the sample, and very little on the size of the population.

2. *The quality of the data (Non-sampling error)*

Besides 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 non-sampling errors. Non-sampling 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, M. E., & Muñoz, J., 1996, p56). Thus, one would expect non-sampling errors to increase with sample size.

3. *The 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 the firm conducting the survey and cost for each country would also affect the total cost of survey implementation. Thus, the cost of data collection, processing, and dissemination should be considered when determining the sample size for each country.

1.2. *Sample size calculation*

Sample surveys are appropriate for the collection of national and relatively large geographic domain level data on topics that need to be extensively explored. The main purpose of this survey is to identify and analyze the energy access tiers (Tier 0 to Tier 5) both at the national level and at the zonal (urban and rural) level. Equation (1) in the previous section indicates the formula to calculate the sample size. Given that the concept of the MTF has been recently introduced and the aim of this global survey is to establish the baseline of monitoring energy access globally, the indicator of interest (r) is unknown. Thus, the sample size for each selected country is calculated using the prevalence rate of 50% as the most conservative choice and to achieve the minimum margin of error (standard errors are inversely proportional to the square root of the sample size: $e=z*\sigma/\sqrt{n}$). Since the non-response rate is typically under 10% in developing countries (United Nations, 2011), a value of 1.1 for k (non-response rate) would be considered a conservative choice (United Nations, 2011, p42). The number of households

[1] The sample size should be calculated to reflect the experience from the country in question. Hence, we will introduce the possibility of replacement of certain households in particular countries, if needed. In this case, a different weight will be considered when preparing the estimates.

selected per PSU (m) is 12 (DHS normally visit 20-35 households per PSU, while socioeconomic surveys rely on 6-16 households per PSU); however, it can be modified depending on the level of homogeneity in a given PSU and community. Due to the characteristics of infrastructure variables/indicator, we select 0.45 for intra-cluster relation coefficient (ρ), consequently, the design effect (f) will be equal to 6 ($f = 1 + \rho (m - 1)$) (Grosh, M. E., & Muñoz, J., 1996, p59).
The number of analytic domains also has a large impact on the sample size and strategy. An analytic domain can be defined as the analytic sub-groups for which equally reliable data is required for the analysis. The sample size is increased by a factor equal to the number of domains desired, because it does not depend on the size of the population itself.
While defining a strategy to calculate the sample size for the selected countries, we have considered two approaches: one calculating, at first, the national sample size as one analytic domain and allocating the sample size proportional to urban and rural population; the other is calculating, at first, the sample size using the distribution between urban and rural as two analytic domains and adding these two values to obtain the national sample size. These two approaches have taken into consideration data on sample size by a margin of error, ranging from approximately 4% to 5.5% at the national level and from nearly 5% to 11% at the zonal level. Considering the results obtained, we have chosen to keep those of the second approach, which for a margin of error of 6% at urban and rural levels gives a national sample size of roughly 3,300 households with an error of 4.2%.
Within each cluster/state, PSUs are selected with probability proportional to its measure of size (PPS) and households are selected with equal probability within each PSU (the definition of this approach is reported in United Nations, 2011).

1.3. *Stratification*

Once the sample size is determined, we develop a stratification strategy, which is the process of dividing households into homogeneous smaller groups called strata and then sampling separately for each stratum following certain rules. Stratification often improves the sample's representativeness by reducing sampling error. Each stratum is treated as an independent population. Sampling weights need to be used to analyze the data reflecting the stratification strategy adopted. This section provides guidelines on stratification for the MTF Global Survey. The guidelines provided in this section are general, and ideally, this is what we aim to achieve in the stratification of the sample for the selected countries. However, these guidelines may not apply identically to all 16 countries where MTF surveys will be implemented as these countries may well vary in their geographical structure and population distribution within and across geographical units. That is, country-specific modification of the guidelines is likely, and such modification will be covered in the country-specific data collection reports.
It is useful to review the criteria that will guide the overall stratification strategy. This stratification is important for the tier analysis and capturing the diversity in different energy solutions and services. Such criteria are:

1. Equal allocation between urban and rural areas. This is established during sample size calculation. This will help conduct disaggregate and in-depth analysis for urban and rural areas, which are statistically sound.
2. While the parameters of interest for the MTF study are access to grid electricity and access to non-solid fuel, the prior will be used in the stratification. However, to make the

analysis representative of the underlying population, sampling weights will be applied to reflect the actual distribution of both grid users and non-solid fuel users in the population.

3. A sample will have 50-50 distribution of grid users and non-users. This will help us conduct in-depth analysis of both groups. As mentioned, sample weights will be used in the analysis to compensate for oversampling of either group.
4. Twelve households will be sampled from each village or urban block (PSU).

1.4. *Sampling frame*

A sampling frame is a complete list of all sampling units that entirely covers the target population. The sampling frame was the 2014 Household Income and Expenditure Survey (HIES) conducted by the Liberia Institute of Statistics and Geo-Information Services (LISGIS).

Table 1. Distribution of Enumeration Areas (EAs) and households by County and Locality

	URBAN		RURAL		TOTAL	
	EA	Households	EA	Households	EA	Households
Bomi	54	4,113	219	16,395	273	20508
Bong	256	20,729	671	49,081	927	69810
Gbarpolu	15	1,640	133	12,893	148	14533
Grand Bassa	129	12,214	339	35,226	468	47440
Grand Cape Mount	23	1,925	255	22,140	278	24065
Grand Gedeh	74	6,925	102	11,218	176	18143
Grand Kru	9	604	121	8,365	130	8969
Lofa	136	14,695	365	34,947	501	49642
Margibi	146	17,813	285	27,282	431	45095
Maryland	64	7,650	107	11,604	171	19254
Montserrado			182		182	0
Greater Monrovia	1967	100,626		-	1967	100625.5
Other Montserrado	101	12,847		18,804	101	31651
Nimba	173	19,300	608	61,434	781	80734
River Gee	27	2,552	81	7,270	108	9822
Rivercess	5	487	147	13,494	152	13981
Sinoe	23	2,594	195	13,235	218	15829
Total	7012	327,339	3810	343,388	10822	570101.5

Note: The sampling frame is the 2014 Household Income and Expenditure Survey (HIES) by the Liberia Institute of Statistics & Geo-Information Services (LISGIS);

1.5. Structure of the sample

The sample for the survey will be a stratified sample selected in two stages. The first stage will be selecting 292 enumeration areas (EA) in the sampling frame; the second stage is selecting 12 households in each sample EA. It is important to divide the sampling frame of EAs into strata in which households are homogeneous. Stratification can increase the efficiency of the sample for the survey. Stratification is achieved by separating in each county the urban and rural villages, and electrified and non-electrified villages; the urban and rural, and electrified and non-electrified villages in each province forms each a sampling stratum. In total, 32 sampling strata have been created. Samples were selected independently in each sampling stratum, by selection of one stage.

1.6. The sampling procedure

In the first stage, a random probability proportional to size (PPS) sample of EAs was selected from each county and urban-rural area. The sample was selected without replacement. At least two EAs were selected from each urban and rural area in each county. In the first stage, 292 EAs were selected using the PPS selection procedure for the entire country (Table 2).

In the second stage of sampling, a random sample of 12 households were selected from all EAs selected in the first stage of sampling. With a low grid electrification rate, particularly outside of the capital of Monrovia, the survey team adapted its sampling approach. The county of Montserrado, where the capital of Monrovia is, was divided into two regions – the Greater Monrovia area and the rest of the county, 'Other Montserrado. The EAs in each county (excluding Greater Monrovia) were divided based on their urban and rural classification. Then a random sample of enumeration areas was drawn based on a probability proportional to size (PPS) from each county and urban-rural area. Within each EA's geographic limits, NRECA International digitized all visible structures using existing GIS data from LISGIS and available satellite imagery. Each structure was digitized as points to represent a household. In each selected EA, once all the points were digitized, 12 structures were selected using a sampling design tool in GIS.

In the Greater Monrovia area, the sample selection used a two-stage stratification approach. In the first stage, 84 enumeration areas were selected in the Greater Monrovia (proportionate to the population). Each enumeration area was further divided into electrified and non-electrified households through a listing process. Enumerators carried out a listing of households to identify which were electrified and unelectrified.

In cases where the EA was too large to inspect, it was subdivided into subsections. One or more subsections were randomly selected and inspected. In cases where the team was unable to select six electrified and six unelectrified houses, for any of the selected EAs because there were fewer than six of either category, a sample of 12 homes of a single category (electrified or unelectrified) was randomly selected. Any shortfall was made up in other EAs.

At the second stage, in each selected EA, a fixed number of 12 residential households were selected among the residential households listed. A total of 3,504 households were selected to be enumerated, 1752 from rural and urban areas respectively.

TABLE 1 • Household survey distribution of samples across Liberia

	Urban		Rural		Total	
	Enumeration areas	Households	Enumeration areas	Households	Enumeration areas	Households
Bomi	2	24	7	84	9	108
Bong	8	96	20	240	28	336
Gbarpolu	2	24	6	72	8	96
Grand Bassa	5	60	15	180	20	240
Grand Cape Mount	2	24	9	108	11	132
Grand Gedeh	3	36	5	60	8	96
Grand Kru	2	24	4	48	6	72
Lofa	7	84	15	180	22	264
Margibi	8	96	12	144	20	240
Maryland	3	36	5	60	8	96
Montserrado			8	96	98	1,176
Greater Monrovia	84	1,008	-	-		
Other Montserrado	6	72	-	-		
Nimba	8	96	25	300	33	396
River Gee	2	24	3	36	5	60
Rivercess	2	24	6	72	8	96
Sinoe	2	24	6	72	8	96
Total	146	1,752	146	1,752	292	3,504