NIGERIA

NORTH-WEST GEO-POLITICAL ZONE

States of

Kaduna, Kano, Katsina, Kebbi, Jigawa, Sokoto, and Zamfara

Beyond Connections

**Energy Access Diagnostic Report Based on the Multi-Tier Framework**

**May 2020**

*Lucia Luzi, Nikunj Prakash Beria, Bryan Bonsuk Koo, Dana Rysankova, and Elisa Portale*





# *WBG-EnergyAndExtractives-Horizontal-RGB-high*ANNEX 2: SAMPLING STRATEGY

## Sample Size Calculation Parameters

The sample size proposed for 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 (non-sampling 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 (1) to calculate the sample size is as follows:

 (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 (1) and is defined as: *f = 1 + ρ (m – 1).*

= Intra-cluster 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 intra-cluster 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 non-response. Households are not selected using replacement. Thus, the final number of households interviewed will be slightly less than 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 non-response 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. The World Bank applies 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.

## Sampling Approach

For this household survey, the target sample was 3,696 households spread across 264 enumeration areas (EAs). The sample was split equally between urban and rural areas, which were treated as analytical domains (1,800 households in urban and rural areas, respectively). The sample size was based on the formula and assumptions as follows. The formula was used to determine the sample size per state:

n R,

where n is the sample size in terms of number of households to be selected and *z* is standardized z-score (normal variate) corresponding to a 95% confidence interval. Estimate of the indicator of interest to be measured by the survey is denoted by *r* and is taken to be 0.5, which yields the maximum sample size. The design effect, *f*, describes the loss of sampling efficiency due to using complex sampling design, and it is assumed to be equal 6. The factor accounting for the anticipated response rate, R, is calculated assuming 90% response. The margin of error, *e*, is assumed to be 6% at urban/rural level; this is equivalent to about 4% margin of error for the study area overall.

The sample was distributed across the seven study states according to their populations based on available data from the 2006 census. Urbanity split was not available at state level.

## Sample Design

The sample design adopted a stratified, cluster sampling approach to select the household survey sample. The sample was stratified by region and electrification status. The sample was selected using the following steps.

First, the firm selected primary sampling units (PSUs). The administrative unit used as PSUs was census enumerations areas (EAs). In rural areas this was equivalent to villages and in urban areas this was wards. EAs were then selected with probability proportional to population size within each state. A fixed number of households (14) was selected within each EA, meaning each household had the same probability of selection. Note that while this number was in principle fixed, a little flexibility was allowed in practice. Thus between 12 and 14 households was an acceptable number within each EA.

The National Population Commission (NPopC) provided population data at EA level and electrification status.

The sample of electrified and non-electrified EAs within each state was drawn from two separate lists of EAs reflecting the two electrification strata. At the EA level, villages or wards where 97% or more of households are connected to the grid was classed as electrified. Conversely, EAs where 3% or less of the households are electrified was treated as non-electrified.

The sample was evenly distributed between electrified and non-electrified areas. Given the different possible scenarios, EAs were selected as follows:

**The state has both villages with electricity and villages without electricity.** Where an uneven number of EAs was selected, the larger number was allocated to electrified EAs.

*Special case 1 (number of electrified PSUs in the state is less than the number of electrified EAs allocated to the state):* When this was the case, the firm selected all the electrified PSUs in the state and oversampled non-electrified EAs. To keep the ratio between on-grid and off-grid users to less than 1.1, the firm oversampled electrified EAs in other states.

*Special case 2 (number of non-electrified PSUs in the states is less than the number of non-electrified EAs allocated to the state):* The firm selected all the non-electrified EAs in the state and oversampled electrified EAs. If the ratio between on-grid and off-grid users was less than 1.1, there was need to oversample non-electrified EAs in other states.

*All the villages in the state have access to electricity (or only a few villages do not have access to electricity, for example, if less than 2% of villages do not have access to the grid in the state the firm adjusted the threshold in consultation with the World Bank team)*. This was a special case in which all the EAs were randomly selected from the list of the enumeration areas.

*No villages in the state have access to electricity*: In this case, all the EAs were randomly selected from the list of EAs. The firm then attempted to pair this state with another state where all sampled villages have electricity.

Within each EA, the firm aimed to interview seven electrified households and seven non-electrified households. Electrified households were defined as households who are connected to the grid while non-electrified households are those who are not connected to the grid. The number of EAs per state ranges between 23 and 67, giving a total of 258 EAs with an additional 6 spare EAs to take care of contingencies.

The EAs have approximately 200 households. Census 2006 block maps were used to identify the selected EAs and establish their boundaries. The firm obtained the block maps from the NPopC, and updated the maps using transect walk of each EA.

At the second stage of sampling, all the households in the area were listed. This listing identified institutional and residential buildings. The head of the household or his/her spouse was the point of contact with the listing team at this point. All the relevant household information was collected, including name of head of household, household size, and grid connection status (electrified and non-electrified). Next, the supervisor sent the household information collected to the administrative office where a fixed number of households was selected from all households within each EA. Systematic sampling was used, making use of a random start between 1 and the sampling interval (determined by sampling frame divided by sample size). Where empty households were encountered at the time of the listing, the team was instructed to ask about the household from neighbors.

Thereafter, the list of selected households was given to the field team, which went to the households to administer the survey questionnaire. This approach was adopted to reduce non-sampling error and ensure the sampling selection was free from any biasness. The main interview was conducted with the head of household or their spouse. The interviewer took the global positioning system (GPS) reading of the location both prior to and at the end of the interview for increased accuracy.

## Systematic Selection of Household

All households selected were listed during the listing exercise. A unique identification (ID) that identifies the EA, rural or urban stratum and connection status was given. In this survey, for a person to be considered a member of the household, he/she must be a member of the immediate family who normally lives in the household and has eats meals together with the other members for the past six months. Exceptions that were considered in the study were the following:

1. Newborn children who were members of the household, even if they were less than six months of age, were considered.
2. Women who had entered a marriage were considered as members of the household, even if they had not lived six months in their new household.
3. Students who had attended school during the school year were considered as members of the household in which they lived during the school year.

The selection of households from the sample frame was done in the following manner: The compiled household list was stratified by connection status and thereafter the selection of both categories of households was drawn.

Assuming N1 (electrified households) = 160 and N2 (non-electrified households) = 40, then the sampling gap for electrified and non-electrified households was 23 and 6 respectively as shown below:

Electrified households (N1 = 160, n = 7)

The firm then randomly selected a number from 1 to 23 as the starting point (random start) and every 23rd household on the list was chosen as an eligible household for the survey.

Non-electrified households (N2 = 40, n = 7)

The firm then randomly selected a number from 1 to 6 as the starting point (random start) and every 6th household on the list was chosen as an eligible participant for the survey.

## Weighting

To ensure that the household sample is representative of the target population weights were calculated. The process involves the steps described below. In terms of terminology, for this study PSU is equivalent to EA. The use of PSU below is therefore interchangeable with EA.

### Design weights calculation

The design weights will adjust for the differential sampling probabilities, reflecting the clustered sample:

*P1hi*: probability of selecting the *i*th PSU/cluster in stratum *h* in stage 1

*P2hi*: probability of selecting the household within the *i*th PSU/cluster in stage 2

Assuming that *nh* is the number of PSUs selected in stratum *h*; *Mhi* is the measure of size of the PSU used in the first stage’s selection—that means it is the number of households residing in the PSU according to the sampling frame (or census);  is the total measure of size in the stratum *h*. The probability *P1hi* of selecting the *i*th PSU in the sample is thus:





Assuming that  is the number of households selected in the EA *i* in stratum *h*, and is the number of households listed in the household listing operation in EA *i* in stratum *h*. The second stage selection probability *P2hi* for each household in the EA is thus:



.

Consequently, the overall selection probability of each household in PSU *i* of stratum *h* is the product of the selection probabilities of the two stages:

.

Finally, the firm calculated the design weight for each household in PSU *i* of stratum *h* as the inverse of its overall selection probability:

.

### Correction for non-response

To adjust for non-response among certain groups of the population, for example, the very wealthiest or poorest, non-response weights were created.

In general, correcting for unit non-response is required to calculate a response rate for each homogeneous response group; subsequently, the design weight must be divided by the response rate for each response group.

The firm first calculated the sampling weight by calculating the various response rates for unit non-response. For this study, only PSU and household levels response rates were considered.

*PSU/Cluster level response rate:*

Assuming that  is the number of PSUs selected in stratum h and  is the number of PSUs interviewed. The PSU level response rate in stratum h is:

.

*Household level response rate:*

Assuming that is the number of households found in PSU i of stratum h and is the number of households interviewed in the PSU. The household response rate in stratum *h* is:



where  is the design weight of PSU i in stratum h. The summation is over all PSUs in the stratum h.

The household sampling weight of PSU i in stratum h is obtained by dividing the household design weight (previously calculated) by the product of the response rate at PSU and at household levels, for each of the sampling stratum:

.

The household sampling weight above was then used to calculate any indicators at the household level. Given that a sampling weight is an inflation factor, the weighted sum of households interviewed is calculated as:

.

This is an unbiased estimate of the whole number of residential households of the country. The summation is over all PSUs and strata in the full sample.

### State-level population weights

The sample was drawn based on available population estimates from the NPopC. During the study, updated official household population projections for 2016 were released. While these population projects were largely in line with the data used for sampling, a state-level weight was created to reflect the latest population data.

The state level weights were calculated as follows:

State\_wt=1/(% HHs in state based on sample / % HHs in state based on 2016 projections).

Note that HH stands for *households*.

## Fieldwork

### Team formation

Teams were selected based on previous experience and involvement in similar household surveys of this nature. Educational qualification was also considered as a requirement for selection of field staff; a minimum qualification of ordinary national diploma was used as a benchmark. Females were given preference over their male counterparts because the culture in the north does not allow males into the households, except with the permission of the head of household.

The team composition during the household listing was three field staff per team (a mapper, lister, and team leader), while during the main household survey the team composition was five (4 enumerators and 1 team leader). In addition, the firm assigned one supervisor to each of the sample locations to monitor the fieldwork and approve/reject interviews on the data collection platform (survey solution). In total, the MTF base line survey employed 115 field workers (90 enumerators, 18 team leaders and 7 Supervisors).

### Field guidelines

**Substitution, callbacks, refusals:** The selected households were allowed to be substituted only after the interviewer made three additional unsuccessful visits over a two-day period and at different times of the day. After these visits, the supervisor gave the interviewer replacement households at the same point as the initial selection.

**Scheduling interviews and increasing the strike rate:** To increase the strike rate, we planned the interviews around the time that most members of the community were available once the EAs had been identified. Interviews were staggered over different days of the week and onto weekends for interviews in urban areas.



**Call log:** All records of successful calls, unsuccessful calls (due to different reasons such as closed doors, refusals, etc.), substitution, and callbacks were kept by the team.