Sampling Strategy-Zambia

A. Sample size calculation parameters

The sample size proposed for Zambia is designed to get sufficiently precise estimates of each tier at national as well as urban and rural level. 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 that should be taken into consideration in the determination of sample size calculation and provides a justification for the proposed sample size for each country. The major issues considered in determining 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 subsections 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 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 purpose. 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 to calculate the sample size is as below:

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

where:

n = Sample size 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: $f = 1 + \rho\,(m - 1)$.

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 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 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 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 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 or level of precision. We apply various level of margin of error from 1% to 5.5% to the calculation.

2. *The quality of the data (Non-sampling 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 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 and Muñoz 1996, p. 56). Thus, one would expect non-sampling error to increase with sample size and we would like to limit the sample size to less than 5,000.

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 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.

B. Sampling approach

In this study, stratified random sampling technique is used. The first stratification involves stratifying into urban and rural strata. The second stratification is based on electrification status of the enumeration areas (EAs) in the study population.

Urban and Rural stratification

The primary sampling units (PSUs) in this study are EAs, selected randomly from the list of EAs in Zambia obtained from CSO Zambia. The EAs were stratified into rural and urban strata. For each stratum, random numbers were allocated to each EA and these EAs were arranged in ascending order. The first EAs to satisfy the sample quota of each province were picked. The number of EAs picked in each province for either rural or urban stratum were dependent on the sample size of each province. A total of 14 households were sampled in each EA, so the sample size of each of the province was divided by 14 to get the total number of EAs to be sampled. An equal split of the sample between rural and urban stratum was done at the national level.

Electrified or non-electrified stratification

Listing was conducted only in the sampled EAs to determine whether to classify an EA into either electrified or non-electrified stratum. EAs with at least 3% of households that were connected to the national grid were classified as electrified while those with less than 3% of households connected to the national grid were classified as non-electrified. A 50-50 ratio of distribution of sample between grid and non-grid users was achieved.

Household selection

During the listing process, information on electricity connection (the number of households with or without electricity in a sampled EA) were collected. Random numbers were allocated to each household and arranged in ascending order for each stratum.

C. Sample size calculation

Sample size calculation is done using this formula:

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

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 using the MTF suggested prevalence rate so as to achieve minimum margin of error and the intra-cluster correlation coefficient $\rho = 0.45$ selected using knowledge of the characteristics of infrastructure. The number of households to be selected per EA, m, and 14 households are proposed. The factor accounting for non-response, k, is calculated to be 1.1 considering that in developing countries the non-response rate is typically 10% or less. The margin of error, e, is taken to be 0.044 (96% confidence). Using these values, the sample size was 3,658 households. Due to the fact that the sample quota allocated to some EAs was not divisible by 14, a slightly higher sample size of 3,668 was covered.

Listing was done for only sampled EAs in all provinces. The number of EAs listed in a given province was calculated as follows:

$$\text{Number of EAs} = \frac{\text{sample quota for both rural \& urban strata for the province}}{14}$$

All households selected were listed during the listing exercise. A unique identification (ID) that identifies the EA, rural/ 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 for the last 6 months. Exceptions that were considered in the study were:

(i) newborn children who were members of the household, even if they were less than six (6) months of age;
(ii) women who had entered a marriage were considered as members of the household, even if they had not lived six (6) months in their new household; and
(iii) students who had attended school during the school year were considered as members of the household in which they lived during the school year.

Of the original sample size of 3,668 targeted households in 262 EAs (130 EAs in urban and 132 EAs in rural areas) (table A2.1), 3,612 households in 260 EAs were contacted (table A2.2), and 3,537 in 260 EAs were effectively interviewed (table A2.3).[1] The response rate is thus 96%, which is the difference between the sample of household originally targeted and those finally interviewed. As explained in paragraph 4, the non-response was mainly due to movement out of the dwelling of respondents (43 households) and unwillingness to participate in the survey.

The following tables (tables A2.1 through A2.3) summarize the number of sampled EAs and household sample distribution. The sample is split into rural and urban strata and is further split between electrified and non-electrified strata.

[1] The sample of 3,537 was used to calculate the weight.

Table A2.1 Distribution of EAs and households in Zambia sampled for the Multi-Tier Framework survey – original sample (households targeted)

Province	Urban				Rural				Nationwide			
	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs
Central	70	42	112	8	19	191	210	15	89	233	322	23
Copperbelt	359	187	546	39	16	82	98	7	375	269	644	46
Eastern	35	35	70	5	6	316	322	23	41	351	392	28
Luapula	26	44	70	5	22	202	224	16	48	246	294	21
Lusaka	410	262	672	48	40	58	98	7	450	320	770	55
Muchinga	27	15	42	3	7	133	140	10	34	148	182	13
North Western	14	42	56	4	0	126	126	9	14	168	182	13
Northern	35	35	70	5	2	208	210	15	37	243	280	20
Southern	81	59	140	10	29	209	238	17	110	268	378	27
Western	21	21	42	3	0	182	182	13	21	203	224	16
Total	1,078	742	1,820	130	141	1,707	1,848	132	1,219	2,449	**3,668**	262

Table A2.2 Distribution of EAs and households in Zambia sampled for the Multi-Tier Framework survey – original sample (households contacted)

Province	Urban				Rural				Nationwide			
	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs
Central	69	42	111	8	18	192	210	15	87	234	321	23
Copperbelt	361	165	545	39	15	81	98	7	376	246	643	46
Eastern	37	32	70	5	7	301	308	22	44	333	378	27
Luapula	24	42	67	5	18	190	211	16	42	232	278	21
Lusaka	415	249	671	48	40	58	98	7	455	307	769	55
Muchinga	23	17	42	3	6	131	140	10	29	148	182	13
North Western	15	39	56	4	0	100	126	9	15	139	182	13
Northern	26	38	68	5	14	172	190	14	40	210	258	19
Southern	85	55	140	10	30	206	237	17	115	261	377	27
Western	21	21	42	3	0	182	182	13	21	203	224	16
Total	1,076	700	1,812	130	148	1,613	1,800	130	1,224	2,313	**3,612**	260

Table A2.3 Distribution of EAs and households in Zambia sampled for the Multi-Tier Framework survey – original sample (households interviewed)

Province	Urban	Rural	Nationwide

	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs	Electrified HHs	Non-electrified HHs	Total HHs	Total EAs
Central	69	42	111	8	18	192	210	15	87	234	321	23
Copperbelt	361	165	526	39	15	81	96	7	376	246	622	46
Eastern	37	32	69	5	7	301	308	22	44	333	377	27
Luapula	24	42	66	5	18	190	208	16	42	232	274	21
Lusaka	415	249	664	48	40	58	98	7	455	307	762	55
Muchinga	23	17	40	3	6	131	137	10	29	148	177	13
North Western	15	39	54	4	0	100	100	9	15	139	154	13
Northern	26	38	64	5	14	172	186	14	40	210	250	19
Southern	85	55	140	10	30	206	236	17	115	261	376	27
Western	21	21	42	3	0	182	182	13	21	203	224	16
Total	1,076	700	1,776	130	148	1,613	1761	130	1,224	2,313	**3,537**	260

C.1 Sample weighting calculations

Sample weights are important in analysing household survey data. Due to this fact sample weighting was executed to reduce bias due to imperfections in the sample. Since we used two-stage stratification, the sample design weight was calculated as $w_i = \frac{1}{p}$, where p is the probability of a unit to be included in the sample. The focus is on design weight, weight attributable to the compensation for non-coverage, and weight attributable to compensation for non-response. Calculation of the design weight was done as follows.

(i) First the probability of selecting a certain EA in rural and urban strata was established, which was the first stage calculated as the number of EAs selected in a stratum multiplied by the measure of size of the EA. The total number households in that stratum were then divided into the result. A 88-12% electrification ratio between urban and rural areas respectively was used to calculate the probability of electrification status of an EA. The 88-12% electrification status split was obtained from the CSO of Zambia.

(ii) The probability of selecting the household within the EA, which is stage 2, was then established. This was simply the number of households selected in the EA in a certain stratum divided by the total number of households listed in the EA in that stratum considering the electrification status.

(iii) We then calculated the overall selection probability of each household in an EA of a certain stratum as a product of values found in (i) and (ii) above.

(iv) We computed the design weight for each household in an EA of a certain stratum as the inverse of the overall selection probability.

Correction for non-response was done at EA and household levels. EA response rate was calculated as the number of EAs interviewed divided by the number of EAs selected in each stratum. Household level response rate was calculated as the design weight multiplied by the sum of households interviewed in a stratum divided by the design weight multiplied to the sum of households listed in a stratum.

D. Fieldwork challenges

The study was carried out successfully, although some challenges were met during the course of the fieldwork. Fieldwork challenges included:

- Inaccessible EAs: A total of 8 sampled EAs were in the wetlands and, thus, difficult to reach because of the rainy season. This delayed fieldwork, as enumerators used a primitive mode of transport. A total of 2 out of 8 EAs were totally inaccessible by any form of transport.
- Overall, about 4% of the sampled households were not interviewed because they were unwilling to participate; furthermore, 43 households moved out of the dwelling after listing.
- Electrification status discrepancies between listing and fieldwork: About 1% of the sampled households recorded as connected during listing were then identified as not connected to electricity during the fieldwork, and this problem was solved by recording the connection status during the fieldwork.
- Permission to interview facilities: The authorization letter from Ministry of Energy was received on time, while the letters from Ministries of Health and Education delayed to the end of the survey.
- Challenges in locating some households in the compound residential areas.