ANNEX 4

Methodology for selection of survey sites by PPS sampling[1]

Since it is usually not possible to randomly select households, a stratified method for household selection is used for population-based surveys. Such "cluster" surveys require identification of a sampling unit such as a village or ward as the "cluster" or site from which households are selected. In the selection of survey sites (or clusters), the basic goal is to select sites that will be representative of the area to be surveyed. Methods used for performing household-based and school-based surveys are described in this annex.

A4.1 Household-based surveys

For a standard, population-based "cluster" survey, the first step is to obtain the "best available" census data for all of the communities in the area of interest. This information is usually available from the central statistical office within the ministry that performs the census for the country.

From the census data, select the data for the area chosen for the survey. Make a list with four columns (see Table 11). The first column lists the name of each community. The second column contains the total population of each community. The third column contains the cumulative population – this is obtained by adding the population of each community to the combined population of all of the communities preceding it on the list. The list can be in any order: alphabetical; from smallest to largest population; or geographical.

The sampling interval (k) for the survey is obtained by dividing the total population size by the number of clusters to be surveyed. A random number (x) between 1 and the sampling interval (k) is chosen as the starting point using random number tables, and the sampling interval is added cumulatively. The communities to be surveyed are those with the (x+n)th person, the (x+2n)th, (x+3n)th, person and so on up to the (x+30n)th person.

The 30 clusters should be plotted on a map. Next, a logical sequence for the fieldwork should be developed for each of the survey teams.

[1] Adapted from: Sullivan KM et al. (*40*)

A4.1.1 An example of selecting communities in a cluster survey

In the fictitious area of El Saba, there are 50 communities (Table 11). In practice there would usually be many more than 50 communities, but this number is used for illustrative purposes to describe the method.

In Table 11 on the opposite page, the first column contains the names of the communities, the second column the population of each community, and the third column the cumulative population. A fourth column is used for identifying which communities will have one or more clusters selected.

Follow four steps to select communities to be included in the survey:

1. Calculate the sampling interval by dividing the total population by the number of clusters. In this example, 24 940 / 30 = 831.
2. Choose a random starting point (x) between 1 and the sampling interval (k, in this example, 831) by using the random number table. For this example, the number 710 is randomly selected.
3. The first cluster will be where the 710th individual is found, based on the cumulative population column, in this example, Mina.
4. Continue to assign clusters by adding 831 cumulatively. For example, the second cluster will be in the village where the value 1541 is located (710 + 831 = 1541), which is Bolama. The third cluster is where the value 2372 is located (1541 + 831 = 2372), and so on. In communities with large populations, more than one cluster will probably be selected.

If two clusters are selected in one community, when the survey is performed the survey team would divide the city into two sections of approximately equal population size and perform a survey in each section. Similarly, if three or more clusters are in a community, the community would be divided into three or more sections of approximately equal population size.

A4.1.2 Selecting households within clusters

Within each cluster there is a need to select households for the assessment. As a general rule, we recommend that, through appropriate sample size calculations, the same number of households be visited in each cluster. In most instances, a sample size between 600 and 900 is sufficient to have a reasonable confidence interval around the coverage estimate. Thus, for example, a 30-cluster survey is desired, and based on sample size calculations, it is found that 20 households are to be visited in each cluster. There are several methods for selecting households within a cluster. In some settings the national census organization may have maps of the areas and census personnel can randomly select the households to

Table 11 Selection of communities in El Saba using the PPS method

NAME	POPULATION	CUMULATIVE POPULATION	CLUSTER	NAME	POPULATION	CUMULATIVE POPULATION	CLUSTER
Utural	600	600		Ban Vinai	400	10 800	13
Mina	700	1 300	1	Puratna	220	11 100	
Bolama	350	1 650	2	Kegalni	140	11 240	
Taluma	680	2 380	3	Hamali-Ura	80	11 320	
War-Yali	430	2 810		Kameni	410	11 730	14
Galey	220	3 030		Kiroya	280	12 010	
Tarum	40	3 70		Yanwela	330	12 340	
Hamtato	150	3 220	4	Bagvi	440	12 780	15
Nayjaff	90	3 310		Atota	320	13 100	
Nuviya	300	3 610		Kogouva	120	13 220	16
Cattical	430	4 040	5	Ahekpa	60	13 280	
Paralai	150	4 190		Yondot	320	13 600	
Egala-Kuru	380	4 570		Nozop	1 780	15 380	17
						18	
Uwarnapol	310	4 880	6	Mapazko	390	15 770	19
Hilandia	2 000	6 880	7				
			8	Lotohah	1 500	17 270	20
Assosa	750	7 630	9	Voattigan	960	18 230	21
						22	
Dimma	250	7 880		Plitok	420	18 650	
Aisha	420	8 300	10	Dopoltan	270	18 900	
Nam Yao	180	8 480		Cococopa	3 500	22 400	23
						24	
						25	
						26	
						27	
Mai Jarim	300	8 780		Famegzi	400	22 820	
Pua	100	8 880		Jigpelay	210	22 840	
Gambela	710	9 590	11	Mewoah	50	22 890	
Fugnido	190	9 880	12	Odigla	350	23 240	28
Degeh Bur	150	10 030		Sanbati	1 440	24 680	29
Mezan	450	10 480		Andidwa	260	24 940	30

be sampled, and provide detailed maps to enumeration teams. In other situations, detailed maps may not be available at the national level and the teams may need to initially spend time at each cluster to perform the household selection themselves. One approach to household selection is to carefully map all households within the cluster and then either randomly or systematically select households to survey. While this approach is ideal, it often requires an additional visit to the cluster, and this can add significantly to the survey cost. Another approach, frequently used in EPI surveys in the past, is to randomly select one household within the cluster and then select subsequent households using the "next nearest household" approach, or selecting households in a specified direction. We do not recommend these approaches, as they may allow some bias in household selection. An alternative recommended method is a segment method. On arrival in the cluster, if the cluster is large, visually divide the cluster into segments. With segmentation, there is an attempt to divide the cluster into approximately equal sample size segments based on roads, rivers, or other geographic demarcations. Each segment should have approximately the same number of households. Once divided, one segment is randomly selected, and a random or systematic selection of households is sampled within that segment.

A4.1.3 Selecting individuals within households

Once households are selected, the response can be taken from the head of the household since the type of salt used in the household likely affects all household members. It is useful to consider collection of urine specimens for urinary iodine assessment in school-age children and pregnant and lactating women, as this will help understand overall iodine intake in these vulnerable groups.

The considerations noted above apply to estimating the proportion of households using iodized salt. Further sample size calculations are needed if additional information is collected, such as urinary iodine or vitamin A supplementation, and this may affect both the number of households to be sampled, and the selection of individuals within the household.

A4.2 School-based surveys

If a school-based survey is to be performed, the Ministry of Education should be contacted to obtain a listing of all schools with children of the appropriate age for the survey. Because the age range for the survey is 6 to 12 years, the grades in which these children are likely to be enrolled should be determined. Ideally, the Ministry of Education will have such a listing.

If one nationwide survey is performed, a listing of schools for the entire nation is needed. If subnational estimates are required, then a listing of the schools for each subnational area is needed. If enrolment information for each school is available, the PPS method should be used for selection. If enrolment information is not available, then systematic sampling can be performed.

A4.2.1 Selecting schools

When performing school-based surveys in a geographical area, the first questions are:

- Is there a list of all schools in the geographic area with the appropriate age range?
- If there is a list of schools, is the number of pupils in each school known?

In most areas, a list of schools and their respective enrolments is available. Ensure that there is the same number of grades/levels in the schools. If a list of schools and enrolments is available, the selection of schools should be performed using the PPS method described for selecting communities. If there is a list of schools but the enrolments are not known, schools can be selected using systematic selection.

Using systematic selection, rather than PPS, complicates analysis somewhat. However, if enrolment information cannot be obtained easily there may be no alternative. If there is an extremely large number of schools in an area, or if a listing of all schools does not exist, another method can be used. This alternative method is described later in these guidelines.

Method 1 – schools when their enrolments are known

In this situation the PPS method for selecting communities, as described earlier in this chapter, should be used. First, generate a list of schools similar to that shown in Table 12. Second, determine the cumulative enrolment. Finally, select schools using the same PPS method as described for selecting communities (see Table 11).

Table 12 Selection of schools using the PPS method

SCHOOL	ENROLMENT	CUMULATIVE ENROLMENT
Utural	600	600
Mina	700	1 300
Bolama	350	1 650
Etc.		

Method 2 – a list of schools is available, but enrolments are not known

When a list of schools is available but the enrolment of each school is not known, the systematic selection method should be employed as follows.

- Obtain a list of the schools and number them from 1 to N (the total number of schools).
- Determine the number of schools to sample (n), usually 30.
- Calculate the "sampling interval" (k) by N/n (always round down to the nearest whole integer).
- Using a random number table, select a number between 1 and k. Use the randomly selected number to refer to the school list, and include that school in the survey.
- Select every kth school after the first selected school.

Example of systematic selection of schools

For illustrative purposes, Table 13 lists 50 schools. The following method would be used to select eight schools:

Step one: There are 50 schools, therefore N = 50.

Step two: The number of schools to sample is eight; therefore n = 8.

Step three: The sampling interval is 50 / 8 = 6.25; round down to the nearest whole integer, which is 6; therefore, k = 6.

Step four: Using a random number table, select a number from 1 to (and including) 6. In this example, suppose the number selected had been 3. Accordingly, the first school to be selected would be the third school on the list, which in this example is Bolama.

Step five: Select every sixth school thereafter; in this example, the selected schools would be the 3rd, 9th, 15th, 21st, 27th, 33rd, 39th, and 45th schools on the list.

In some circumstances, this method might result in the selection of more than the number needed. In the above example, for instance, had the random number chosen in step four been 1 or 2, then nine schools would have been selected rather than eight. This is because the value for k was rounded down from 6.25 to 6.

In this situation, to remove one school so that only eight are selected, again go to the random number table to pick a number. The school that corresponds to that random number is removed from the survey.

To analyse properly the data collected using systematic sampling, additional information needed would include the number of eligible pupils in each school. Note that usually 30 clusters are selected; the eight indicated in Table 13 have been selected in this example for illustrative purposes only.

Table 13 Selection of schools using the systematic selection method

SCHOOL	SELECTED	SCHOOL	SELECTED
1 Utural		26 Ban Vinai	
2 Mina		27 Puratna	Y
3 Bolama	Y	28 Kegalni	
4 Taluma		29 Hamali-Ura	
5 War-Yali		30 Kameni	
6 Galey		31 Kiroya	
7 Tarum		32 Yanwela	
8 Hamtato		33 Bagvi	Y
9 Nayjaff	Y	34 Atota	
10 Nuviya		35 Kogouva	
11 Cattical		36 Ahekpa	
12 Paralai		37 Yondot	
13 Egala-Kuru		38 Nozop	
14 Uwarnapol		39 Mapazko	Y
15 Hilandia	Y	40 Lotohah	
16 Assosa		41 Voattigan	
17 Dimma		42 Plitok	
18 Aisha		43 Dopoltan	
19 Nam Yao		44 Cococopa	
20 Mai Jarim		45 Famegzi	Y
21 Pua	Y	46 Jigpelay	
22 Gambela		47 Mewoah	
23 Fungido		48 Odigla	
24 Degeh Bur		49 Sanbati	
25 Mezan		50 Andidwa	

Method 3 – an extremely large number of schools

In very large populations, it may not be possible or efficient to select schools using either the PPS or the systematic selection method. For example, Szechwan Province in China has a population of approximately 100 million. Even if a list of schools were available at the provincial level, it would take much time and effort to select schools using either of these methods.

Accordingly, another approach may be more appropriate. First, select districts using the PPS method. Develop a listing of the districts, their populations, and cumulative populations similar to the PPS selection

described earlier. Next, determine the number of schools to survey, based on the cumulative population using PPS.

For districts with one or more clusters to be selected, select schools in each district using a random number table. For example, if a district has 200 schools, number them from 1 to 200. Then, randomly select a number from 1 to 200 using the table. If two schools are to be selected, then randomly select two numbers. Finally, and while not technically correct, it would be acceptable to analyse the school-based data as though the schools were selected using PPS methodology.

Selecting students within each selected school

Once the school has been selected, it is usual to select one class or grade, and to sample all students in that class – both male and female. If the schools are large, it may be necessary to divide the class, and pick one of the divisions randomly, sampling all children in the selected portion. If schools are small, it may be necessary to include more than one class.

Other possibilities

In situations where male and female children attend the same school, the selection of schools and pupils would be the same as discussed above. In situations where males and females attend separate schools, when a school of one sex is selected the nearest school of the opposite sex should also be surveyed.

For example, a survey is to be performed in an area where males and females attend separate schools. Thirty schools are to be selected, and 20 pupils sampled in each. When an all-male school is visited, information should be collected on 10 male pupils. Then, the nearest female school is visited, and information collected on 10 female pupils.