

Chapter 4: Sampling

4chsampl, May 15, 2002

Key Messages

- LSMS samples are small in size, generally from 2,000 to 5,000 households, to balance sampling and non-sampling errors.
- LSMS samples are designed to represent the population of the country as a whole, as well as that of certain subgroups of the population, called analytical domains.
- LSMS samples are drawn in two stages. In the first stage, a certain number of area units called *Primary Sampling Units* (or PSUs) are selected. In the second stage, a certain number of households, usually 16, are selected in each of the designated PSUs. Both stages are random selections.
- Two-stage sampling reduces the cost and effort of sampling and of field work compared with single-stage sampling, but at the cost of increasing the sampling error. This is a result of the so-called "cluster effect."
- The first stage of sampling requires developing a sample frame from census files. The second stage requires listing all households in the selected PSUs and then choosing a random sample of those households for the final sample.
- To derive unbiased estimates from the survey, the values observed in the sample may need to be weighted. To compute the needed raising factors, and the correct sampling errors, all stages of sampling must be carefully recorded and made available to the survey analysts, both in written documents and in the survey data sets.

Many of those who work on survey implementation or who use the resulting data never learn the details of how sample designs are chosen and implemented. This chapter tries to dispel some of the mystery. Section A reviews the basics of sample design. It may be skipped by readers who know something about the subject. Section B explains the choices made in the usual LSMS sample design and the reasons for them. All readers should read this section. Section C provides a step-by-step guide on how to carry out the sampling. Readers who will not be involved in sampling may skip it or skim it.

A. Overview of Issues in Sample Design

The main objectives of an LSMS are understanding the determinants of household behavior and the overall distribution of welfare. The sample design should determine the number and location of the households to be observed in a way that best achieves these goals within budgetary and organizational constraints. The following issues must be considered:

To reliably depict the overall situation of the population, the selected sample should contain a sufficient number of households, scattered as much as possible throughout the country. However, to reduce the costs, simplify management and control the quality of the interviews, the sample size and its geographical dispersal must be kept within reasonable limits.

The population of the country may contain certain subgroups, such as urban and rural areas or other aggregates, that deserve to be studied separately. The sample of households should adequately represent each of these subgroups as well as the country as a whole.

Each household in the country should be given a chance to be selected in the sample. To simplify survey design and analysis, this chance should be similar for all households, or at least for all households within the same large domain.

Some insights into how to arbitrate among these objectives and constraints can be obtained from a quick review of four concepts: sampling error, non-sampling error, multi-stage sampling, and analytical domains.

SAMPLING ERROR. Sampling error is the error inherent in making inferences for a whole population from observing only some of its members. Sampling theory studies the behavior of sampling error for different design options. It is usually assumed that one of the variables to be observed is of particular interest, for instance, household income, unemployment, or infant mortality, and that the sample design should maximize the precision of the estimates of this variable, given cost constraints. Several good textbooks explore this complex issue and it does not need to be specified in detail here (see reference list in Annex II). It is important, however, to bear in mind two general conclusions of sampling theory.

First, the law of diminishing returns underlies the relationship between sample size and sampling error. Roughly speaking, and other things being equal, the sampling error is inversely proportional to the square root of the sample size. This means that, even with the best design, to reduce the error of a particular sample by half, the number of households visited must be quadrupled (See Box 4.1).

**Box 4.1: Sampling Error and Sample Size:
A Case of Diminishing Returns**

For a simple illustration of the diminishing returns relationship between sample size and sampling errors, consider the case where a proportion (for instance, the proportion of households with pre-school children) is estimated from a *simple random sample* of n households, extracted from an infinite population. Let p be the value of the proportion in the population. The standard error is:

$$e = \sqrt{\frac{p(1-p)}{n}}$$

The table below gives the values of e for different sample sizes and $p=50\%$:

Table B 4.1.1

Sample size (n):	100	200	500	1000	2000	5000	10000
Standard error (e):	5.00%	3.54%	2.24%	1.58%	1.12%	0.71%	0.50%

Notice that in order to reduce the error from 5.00% to 0.50% (a tenfold reduction), the sample must be increased a hundredfold, from 100 to 10,000 households. (See Cochran (1977) Chapter 3 for more information.)

Second, the sample size needed for a given level of precision is almost independent of the total population. For instance, a 500-household sample would give essentially the same sampling precision whether it is extracted from a population of 10,000 or 1,000,000 households, or indeed, from an infinite population. Some people find it hard to believe that the sample size does not depend very much on the size of the population; they feel the relationship should be more or less proportional. An intuitive grasp of this seemingly striking statistical fact can be obtained by noticing that, in order to test if the soup is salty enough, an army cook does not need to take a larger sip from the regimental pot than a housewife needs to take from the family saucepan (see Box 4.2). This does not *necessarily* mean that the size of an LSMS sample is independent of the size of the country. Large countries generally require larger samples, not because they are large but because large countries tend to demand results for a larger number of internal subdivisions. India, for example, would probably require state-level data from any survey.

Box 4.2: Sample Size and Population Size

The formula in Box 4.1 is valid for simple random sampling from an infinite population. For a finite population of N households, it should be corrected as follows:

$$e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{p(1-p)}{n}}$$

The term:

$$\sqrt{1 - \frac{n}{N}}$$

is called the *finite population correction*, which essentially depends on the *sampling fraction* n/N . Table B 4.2.1 shows the sample size n that is needed to achieve a 5% standard error for a proportion $p=50\%$ and different population sizes N :

Table B 4.2.1

Population Size (N)		500	1000	5000	10000	50000	Infinite
Sample size (n)	83	91	98	99	100	100	
Sampling fraction (n/N)	0.166	0.091	0.020	0.010	0.005	0.000	

Notice how little the required sample size n changes between a population of 5,000 and infinity. In national household surveys the finite population corrections are so small that they are almost always ignored.

NON-SAMPLING ERRORS. Beside sampling errors, data from a household survey are vulnerable to other inaccuracies from causes as diverse as refusals, respondent fatigue, 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. Thus one would expect non-sampling error to increase with sample size.¹

MULTI-STAGE SAMPLING. Samplers usually do not have a single complete list of households from which to draw a random sample. Even if such a list were available, a sample taken from it would entail high travel costs because selected households would be spread thinly over the entire country.

Both of these problems can be diminished by using two or more stages in sampling. In the version of two-stage sampling generally used for LSMS surveys, a certain number of small area units are selected with Probability Proportional to Size (PPS), then a fixed number of households are taken from each selected area, giving to each household in the area the same chance of being chosen.²

The area units are usually the smallest recognizable geographic units in the national census. These are usually *census enumeration areas* (EAs), which are aggregates of 50 to 200 households. Less often, the first stage sampling has used

1. UNHSCP (19XX) "Non-sampling errors in Household Surveys (Assessment and Control)."

2. The size of an area is generally defined as the number of households in the area. Alternative size measures are the number of dwellings and the total population.

administrative units such as wards, sectors, etc. Whatever their nature, these may be called *Primary Sampling Units*, or PSUs. However, in many countries those PSUs that are exceptionally large have been divided into *segments*, one of which is selected per PSU, in order to economize on household listing. The final operational area units are then a mixture of PSUs and segments. To simplify the description it is convenient to continue using the word "PSU" for both PSUs and segments.

The two-stage procedure just described has several advantages. It provides an approximately self-weighted sample (i.e., each household has roughly the same chance of being selected), which simplifies analysis. It also reduces the travel time of the field teams relative to a single-stage sample, because the households to be visited are clumped together in the PSUs rather than spread out evenly over the whole country. An additional advantage of selecting a fixed number of households in each PSU at the second stage is that this makes it easy to distribute the workload among field teams.

A two-stage sample, however, will yield larger errors than a simple random sample with the same number of households because neighboring households tend to have similar characteristics. A sample of households drawn in two stages will therefore reflect less of a population's diversity than a simple random sample of the same size. The influence of two-stage sampling on the precision of the estimates is called the *cluster effect*. As would be expected, the cluster effect **grows** with the number of households selected in each PSU. In other words, for a fixed total sample size, a design with more PSUs and fewer households in each PSU will provide more precise estimates of sample statistics than a design with fewer PSUs and more households in each PSU (see Box 4.3).

Box 4.3: Cluster Effects

If the sample of n households referred to in Box 4.1 is not selected by simple random sampling but in two stages (m households in each of c PSUs, with $n=cm$) and without stratification, the formula for the standard error should be corrected as follows:

4

The term in brackets is called the *design effect* (see Kish, 1965). It 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. ρ is the so-called *intra-cluster correlation coefficient* — a number that measures the tendency of households within the same PSU to behave alike in regards to the variable of interest (for the example in Box 4.1, this would be the tendency of households with pre-school children to be clumped in the same PSUs). 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 variables such as the access of the household to running water. Table B 4.3.1 below gives the design effects due to clustering for various values of ρ and m :

Table B 4.3.1

Number of households per PSU (m)	Intra-cluster correlation (ρ)						
	0.00	0.01	0.02	0.05	0.10	0.20	0.50
	Design Effect						
5	1.00	1.04	1.08	1.20	1.40	1.80	3.00
10	1.00	1.09	1.18	1.45	1.90	2.80	5.50
20	1.00	1.19	1.38	1.95	2.90	4.80	10.50
50	1.00	1.49	1.98	3.45	5.90	10.80	25.50

The field teams will typically spend a large amount of time and thus incur substantial costs in travelling between PSUs. Surveying each PSU also entails certain costs that are independent of the number of households to be visited in each PSU, such as the listing operation explained below. It may therefore be tempting to try to reduce the cost of the survey by increasing the number of households in each PSU and reducing the total number of PSUs accordingly. However, the cluster effect indicates that this may often be a false economy.

ANALYTICAL DOMAINS. For political or policy reasons, some subgroups of the population are so important that the survey is expected to provide separate, reliable results for them. Typical examples include division into urban and rural locations and into major administrative units such as states, but the subgroups do not necessarily have to be geographical aggregates -- for instance, the urban households whose head works for the public sector became an explicit field of interest in certain SDA surveys. The design will then have to ensure a minimum sample size within each of these subgroups, which can then be called analytical domains. For large domains this may occur automatically whereas in other cases it may be necessary to oversample certain

analytical domains and to modify the expansion factors (also called "sampling weights") accordingly. The two stage sampling procedure is applied independently within each of those differently weighted domains.

Analysts would often also like to have sufficient sample sizes in smaller analytical groups, such as rural locations in the irrigated parts of a certain region. They may even want to carry the disaggregation further, for example, to study separately male- and female-headed households in rural irrigated areas. This ideal, however, cannot be fully achieved for all possible analytical domains because it would result in a prohibitively large total sample. Therefore, defining the most significant partitions for a sample entails establishing some priorities at the design stage. Often these will not be dictated by policy relevance alone, but also by local statistical folklore and geopolitical considerations.³

B. Sampling Practice in LSMS Surveys

THE BASIC SAMPLE DESIGN. The sample size for LSMS surveys has usually been small, in the range of 2,000-5,000 households (see Table 4.1). The samples are usually two-stage.⁴ The Primary Sampling Units are area units selected with probability proportionate to size. The second-stage units are households, with a fixed number of households per PSU, normally about 16. When a partition into differently weighted domains has been defined, the two-stage sampling procedure is conducted within each of them; the number of differently weighted domains has generally been kept low, between one and four.

3. The partition of a sample into analytical domains is akin to the concept of "sample stratification." Sample stratification, however, is generally done to improve the overall precision of the sample, rather than to study each partition separately. A stratified design that seeks to reduce the overall error usually entails oversampling the parts of the population with the largest variance. In measuring welfare this would entail oversampling the richer parts of the population.

4. Procedures with more stages are possible and, indeed, are sometimes followed by statistical agencies. In three-stage sampling, for instance, instead of selecting small area units directly, some larger areas (such as provinces) are selected first; smaller areas are then chosen only within the first-stage areas so selected. The effect is that the small area units themselves (and not just the households) become clumped rather than being spread throughout the national territory. The most serious disadvantage of multi-stage sampling is that each additional sampling stage increases the sampling error, sometimes considerably. The one frequently quoted advantage of using more than two stages is that it reduces the amount of travel between survey localities. However, this does not apply to the LSMS because of the way field work is organized: the field teams return to a local headquarters between work in each locality. When they return to the field again it just as easy to go to any one of their assigned localities as to any other. Therefore, we do not recommend using more than two stages of sampling in LSMS surveys.

Table 4.1: Sample Design in Selected LSMS Surveys

Note: Though the Guinea and Mozambique surveys were conducted by the Cornell University Food Security Program, their purpose and methodology are very similar to the World Bank surveys, which makes them interesting examples of LSMS field implementation. **[XX Someone who knows WP tables better than me, please move this note to the bottom of the table, but still inside it].**

<i>Country</i>	<i>Year</i>	<i>Sample Size (HH)</i>	<i>Households per PSU</i>	<i>No. of Differently Weighted Analytical Domains</i>	<i>Partition Criteria</i>
Côte d'Ivoire	1985-88	1600 (per year)	16	1	none
Peru	1986	5120	10 in Lima 16 elsewhere	25	Metro Lima, urban/rural in 12 locations
Ghana	1988	3200	16	1	none
Mauritania	1988	1800	16	4	Nouakchott, other cities, rural in river areas, other rural areas
Pakistan	1991	4800	16	4	Four provinces: Punjab, Sind, Balochistan and NWFP
Tanzania - Kagera Region	1992-93	816	16	3	Groups defined as a function of mortality rates and geographic location
Guinea (Conakry urban area)	1988	1728	8	1	none
Mozambique (Maputo/ Matola urban area)	1991	1840	10	1	none
Nicaragua	1993	4200	10	2	urban/rural
Viet Nam	1992	4800	16	1	none
Nepal	1995	3300	12	4	Mountain, urban hills, rural hills, Terai

Decisions about the sample design for LSMS surveys have been made on a somewhat more qualitative (some would even say *ad hoc*) basis rather than through the application of quantitative sampling formulae for several reasons.

First, one of the overriding objectives of the LSMS was to create very high quality data sets. Thus, great weight has been given to minimizing non-sampling error. Because the questionnaire is complex and fieldwork requires extensive supervision, the consensus has been that non-sampling error could only be kept to the desired standard by using samples in the range of 2,000-5,000 households. As a result, survey planners decided to accept higher sampling error in exchange for lower non-sampling error.

Second, taking advantage of the wealth of information that LSMS surveys provide and addressing the complex behavioral questions that motivate the surveys requires sophisticated multivariate analytical techniques. Thus the precision of estimates of means from simple two- or three-way tables was not deemed of overwhelming importance. Moreover, in designing the LSMS it was judged of much greater analytical interest to have a large amount of information about a relatively small number of households rather than a little information about a larger sample.

Third, given the multiple purposes of an LSMS survey, it is hard to select one single variable for the purpose of minimizing sampling error.

HOUSEHOLDS AND DWELLINGS. The basic analytical unit of LSMS surveys is the **household**. Many surveys define the household as a group of people who share a roof and a cooking pot.⁵ LSMS surveys often also require individuals to have been present for at least three of the past twelve months in order to be considered as household members (though heads of household and newborn infants are considered members even if they have not been present that long).

The second sampling stage almost always requires a field operation called "household listing." Enumerators visit each selected PSU to update the existing maps and prepare the list of all households currently living there. Households to be interviewed are to be selected from this list.

The practical implementation of this operation makes it difficult to preserve the above definition of a household, because that would entail time-consuming interviewing throughout each PSU. In practice, **dwelling**s are listed instead of households. A dwelling is defined as "a group of rooms or a single room occupied or intended for occupancy as separate living quarters by a family or some other group of persons living together, or by a person living alone."⁶ Besides the advantage of being shorter to complete, a listing of dwellings is more permanent than a listing of households.

Strictly speaking, therefore, the LSMS samples are samples of dwellings rather than of households, though the listing operation is still traditionally called "household-listing" rather than "dwelling-listing."⁷ Some dwellings may be unoccupied and some

5. For a discussion of the concept of household and its variants, and details on the operational definitions used by various UN agencies see NHSCP. (1989). "Household Income and Expenditure Surveys: A Technical Study".

6. Kish (1965).

7. The confusion in terms is further complicated by the fact that in regions without street

may be occupied by two households or more, but the large majority of dwellings are occupied by one single household. (The average number of households per dwelling ranges from 0.9 to 1.1 in most countries.) If a dwelling with two households is selected in the sample, both are interviewed separately.

NON-RESPONSE AND HOUSEHOLD REPLACEMENT. Some households selected for the sample will not be interviewed because of one of the following reasons: the interviewer cannot locate the dwelling; the dwelling is uninhabited; the dwelling's residents are away from home and expected to remain so until after the end of the survey period in that area; or the residents refuse to be interviewed.

Non-responding households *cannot* be considered to be a random sample of all households. Non-response rates are always higher in urban than in rural areas and higher in rich households than in poor households. They also have a clear tendency to decrease as the survey proceeds and the field staff becomes more experienced and persuasive. Surprisingly enough, refusal does not seem to be related to the length of the questionnaire but to the unwillingness of certain people to be interviewed at all.⁸

There is a lot of controversy about what should be done about non-response. Some survey practitioners try to achieve the planned sample size by replacing the refusals with other households, whereas other specialists condemn these efforts as sterile and argue that the resulting sample of non-refusals will still be biased by definition. Neither replacing nor failing to replace non-responding households solves the essential problem of bias. Thus everybody agrees that all efforts should be made to keep non-responses to a minimum and that the choice of replacements, if any, should not be given to the interviewers lest a sample of "easy-to-interview" households results.

The solution adopted by LSMS surveys is pragmatic and is based on the principle that interviewers should not be "rewarded" by having to do less work in the case of a non-response. Non-responding households are replaced by other randomly selected households by means of an explicit procedure that is explained in the next section of this chapter. All the details of this process (including the codes of the replaced and the replacing household and the reasons for replacement) are properly documented, both in the questionnaires and in the computer files, to let each analyst decide individually whether or not to include the replacement households in the data sets being analyzed.

The survey managers should carefully monitor all replacements, especially those determined by refusal. Many surveys have demonstrated that refusals rates can be reduced to a minimum, since refusals often depend on the interviewers' attitudes and experience. There is empirical evidence that individual interviewers usually have very different refusal rates. It is useful to stress this to interviewers while monitoring refusal

addresses and house numbers, dwellings are usually identified by the name of the head of the household currently living there.

8. This is worth remembering when it is necessary to defend the riches of the LSMS questionnaire content against those who insist that it is unmanageably long.

rates.

Refusals and replacements have been relatively low in LSMS surveys. In the Mozambique survey, out of the 560 first households visited, only seven were not those originally selected and only three refused to be interviewed, a trifling number that is the more remarkable in a country at war. In Côte d'Ivoire, the non-response rate was 7.8 percent the first year, of which 1.4 percent was refusals. In Peru (1985) the non-response rate was 17.4 percent with a 1.4 percent refusal rate. The overall non-response rate during the first month of the Romania survey was 7 percent, though it reached 18 percent in some neighborhoods in Bucharest.

C. Implementing a Sample Design

Determining the Basic Sample Design Parameters

As explained above, the decisions about the basic sample design parameters (the number of households in total, per PSU, and per analytical domain) are based on qualitative judgements based on past experience and estimates of cost and manageability. The decisions about the basic sample for an LSMS generally follows these steps:

- (1) A preliminary estimate of the total sample size is established. As explained above, the sample rarely exceeds about 5,000 households, but may be much smaller if a single analytical domain is required, or because of constraints on the budget or implementation capacity.
- (2) Using data from the most recent census, this sample is distributed in proportion to the total number of households in the major regions, urban and rural locations, etc. In other words, the option of using a constant sampling fraction throughout the country (i.e. a self-weighted national sample) is taken as a starting point.
- (3) If the sample seems insufficient for some particular analytical domains (fewer than, say, 300 to 400 households)⁹, the sample size may be increased in these domains and decreased in other domains.

While implementing Step (2), parts of the population may be purposely

9. There is no rigorous, quantitative justification for using this particular number. Rather, a wide variety of analyses of different types on different variables have converged upon this as a reasonable ballpark figure. Analysts complain loudly when the numbers get much below this threshold, but are often reasonably content above it. For a variable with a proportion of forty percent (for example, the percent of households with pre-school aged children), ignoring the finite population correction, assuming a typical LSMS take of 16 households per cluster, and an intra-cluster correlation of .05, a 400 household sample gives a 95 percent confidence interval ranging from 33.65 to 46.35 percent. This underscores the need for caution in reporting results for very small subsets of the population.

excluded from the sample, because of their inaccessibility or for security reasons. This happened in Peru, where in 1985 three provinces were controlled by guerrillas and/or drug dealers, and in Pakistan, where the most remote parts of Balochistan were extremely hard to reach.¹⁰ Likewise, the Mauritania survey excluded the nomadic population. The survey in these cases is explicitly designed to represent only the rest of the country.

Step (3) may have to be repeated a few times until a satisfactory partition is achieved. Given that the resources needed to conduct interviews can vary significantly across the territory (interviews are usually more expensive in the rural areas and in the most isolated parts of the country), it is useful and instructive to explore the alternative options with the aid of a spreadsheet to take into account their budgetary and logistical implications.

As a general guideline, we believe it is better to reduce the number of partitions imposed in this way to a minimum and to keep their sampling fractions as close as possible so that the total sample does not differ too much from a self-weighted national sample. While reasonable statisticians and econometricians hold varying opinions of the theoretical virtues of self-weighting, we are much swayed by more pragmatic issues. The more complicated the sample design, the more often the sampler will make mistakes in executing the sampling and the less often others will be able to detect them. There is also a long history of sampling weights being lost, incorrectly calculated, or omitted or misused in analysis. Self-weighted samples are much more robust to this kind of error than more complicated designs.

In a self-weighted sample, the proportions and averages obtained from the sample are unbiased estimates for the proportions and averages in the population. However, when adjustments are made in step (3), the sampling fractions will become different across analytical domains, and the sample will no longer be self-weighted. The households will need to be weighted differently to get unbiased estimates. Calling N_k the total number of households in the population of domain k and n_k the number of households sampled in domain k , the weight w_k to be applied to the values from that domain is

$$w_k = \frac{N_k}{n_k}$$

Note that w_k is the inverse of the selection probability of each household in domain k . As with all sampling information, the basic set of weights (also known as *expansion factors* or *raising factors*) resulting from this step of sample design should

10. The decision to exclude remote areas from the sample has to be considered carefully, though. Often these areas are very vast and tend to be frontier regions that are important to national politics (e.g. the Amazon basin in Brazil or the Chaco region in Paraguay), so that the survey may "look bad" in the eyes of policy makers if they are excluded from the sample. However, these areas tend to be also so sparsely populated that, if included, only a few clusters will be selected for the sample, and the extra cost of visiting them would be manageable.

be carefully documented and made available to the survey analysts.

The number of PSUs to be sampled is determined by the total sample size and the number of households to be interviewed in each PSU. The latter depends on both theoretical and practical considerations. On the one hand, the number of households per PSU affects the precision of the sample, as explained above when discussing cluster effects. On the other hand, the number of households per PSU is a function of the length of the interviews, the number of interviewers in each team, and the time each team spends in the PSU. Typically, each field team visits 20 PSUs per year, spends two weeks in each PSU, and interviews 16 households in each, though in some LSMS surveys as few as 10 or as many as 24 households per PSU have been selected.

Implementation of the First Sampling Stage

THE SAMPLING FRAME. Implementation of the sample begins with the sample frame --the complete list or file of units from which the sample units are selected.¹¹ To develop a sample frame from census data, it is important to obtain a computer-readable list of all PSUs, with a measure of size such as the number of households, the number of dwellings or the population, recorded in each of them.¹² All statistical agencies must eventually process this information in order to obtain the classic census tabulations for larger geographic aggregates, but the preparation of the PSU list as a separate by-product is often forgotten. When the list is not available, the data must be compiled and put into a computer file as quickly as possible. This should not take more than a few weeks, and the list usually fits on one diskette; it does not require that all data from the census be entered or analyzed.

Though only the total number of households or dwellings in each PSU is really needed, the list will probably also include the total population of each PSU, broken down by sex. This information should be entered into a spreadsheet like the one shown in Figure 4.1. If the sample considers differently weighted domains, the procedure described here should be applied independently within each of them (i.e. the sample frame data should be entered in a separate spreadsheet for each domain). The spreadsheet contains one line for each PSU and columns for descriptive information such as the province, district (or whatever administrative hierarchies are used locally), PSU number, population, number of males, number of females, and number of households or dwellings.

11. For an extensive discussion of sample frames, see UNHSCP (1986) *Sampling Frames and Sample Designs for Integrated Household Survey Programmes*.

12. At least minimally sufficient census information has been available in most LSMS surveys. One exception was the Conakry 1988 survey. There the last colonial census had recorded some 50,000 people in the city, which had grown to about 1 million by 1988. This situation was resolved with a special cartographic operation and a subsequent area sampling procedure that does not need to be further described because it is unlikely to be necessary in other countries. The current wave of LSMS surveys will benefit from the 1991-1993 wave of national censuses, which provide census data for most countries.

Figure 4.1: List of First Stage Sampling Units

	A	B	C	D	E	F	G
1	Pro-	Dis-	PSU	Popu-	N° of	N° of	N° of
2	vince	trict		lation	Males	Females	Hholds
3							
4	1	1	1	365	180	185	62
5	1	1	2	262	143	119	43
6	1	1	3	357	172	185	58
7	1	1	4	503	267	236	71
...

After all the data have been entered and before proceeding any further, a series of checks should be carried out to ensure that no PSUs have been omitted from the listing and that all the data are correct. These tests are relatively easy to implement within the spreadsheet, and may include the following: (i) The total population in each PSU should equal the number of males plus the number of females in the PSU. (ii) The masculinity rate (number of males as a percent of the number of females) in each PSU should be within reasonable limits (e.g., between 80 and 120 percent). (iii) The average household size in each PSU should be within reasonable limits (e.g., between 3 and 10 persons per household). (iv) The total number of PSUs and households, as well as the totals by sex in each administrative unit, should be consistent with the other information available from the statistical agency.

Also, the list should be scanned to make sure that the PSUs are not too small. Small PSUs may be too homogeneous (and some of them could even be too small to select the required number of households in the second stage). PSUs smaller than 30 households should be appended to some of the neighboring PSUs, which is facilitated by the fact that statistical agencies generally number the PSUs according to some geographical pattern, so that two PSUs with sequential codes will be neighbors. For example, when the sample frame was being developed for an LSMS being planned for Paraguay, almost all PSUs in urban areas were smaller than 10 households and an ad hoc computer program was written to create larger aggregates.

SELECTING PSUS. After the sample frame has been reviewed, the actual selection of the sample of PSUs to be visited by the survey can proceed. The method for making this random selection with PPS will be explained below. Here we assume that the **number of households** is used as a measure of PSU size; the same method would apply if some other measure of PSU size were used.

Another column must be added to the spreadsheet for the cumulative number of households. This column will contain the total number of households up to and including the corresponding PSU on each line, as in column "H" in Figure 4.2. The last line in column H will contain the total number of households.¹³

Figure 4.2: Cumulative Totals in the List of First Stage Sampling Units

	A	B	C	D	E	F	G	H
1	Pro-	Dis-	PSU	Popu-	N° of	N° of	N° of	Cumulative
2	vince	trict		lation	Males	Females	H Holds	N° of HHs
3								
4	1	1	1	365	180	185	62	62
5	1	1	2	262	143	119	43	105
6	1	1	3	357	172	185	58	163
7	1	1	4	503	267	236	71	234
..

The complete spreadsheet should be printed and kept for reference. Selecting PSUs with PPS can be done manually on the printout or automatically with the spreadsheet. For the sake of simplicity the manual procedure is described here.

First, divide the total number of households by the number of PSUs to be selected and round it to the nearest whole number. Call this number "SI" (the sampling interval).

$$SI = \frac{\text{Number of households}}{\text{Number of PSUs to be selected}}$$

For instance, if the number of households is 200,000, and 184 PSUs are to be selected, then $SI = 200,000 / 184 = 1,087$.

Second, using a table of random numbers or a scientific pocket calculator, obtain a random number between 1 and SI (if a calculator is used, obtain a random number between 0 and 1, multiply it by SI, add 1, and drop the decimals). Call this number "RS" (the random start). Assume, for instance, that RS turns out to be 127.

13. Column H can easily be calculated within the spreadsheet with a simple formula. Continuing with the example in Figure 3.1, the formula G4+H3 would be entered in cell H4, and then copied all the way down column H.

Third, write a sequence of the 184 numbers obtained by starting with RS, and repeatedly adding SI. With the above values of RS and SI, this sequence would start like this:

$$\begin{aligned}
 &127 \\
 127 + 1087 &= 1214 \\
 1214 + 1087 &= 2301 \\
 2301 + 1087 &= 3388 \\
 \dots &\quad \dots
 \end{aligned}$$

Fourth, starting with the first number in the sequence, scan the printout of the PSU list for the first PSU where the "Cumulative N° of Households" is equal to or larger than this number. This PSU is selected for the sample.

Continuing with the example above, the first number in the sequence is 127. Scanning the PSU list, the first and second PSUs should be skipped, because the respective cumulative numbers of households are 62 and 105, which are less than 127. However, the cumulative number of households for the third PSU is 163, which is greater than 127. PSU Number 3 in District 1 of Province 1 would therefore become the first PSU selected in the sample (see Figure 4.3).

Figure 4.3: Selecting the First Stage Sampling Units

	A	B	C	D	E	F	G	H
1	Pro-	Dis-	PSU	Popu-	N° of	N° of	N° of	Cumulative
2	vince	trict		lation	Males	Females	Hholds	N° of Hhs
3								
4	1	1	1	365	180	185	62	62
5	1	1	2	262	143	119	43	105
6	1	1	3	357	172	185	58	163
7	1	1	4	503	267	236	71	234
..

Finally, repeat the above procedure for the remaining 183 numbers in the sequence and create a separate list of the province, district, and numbers of the PSUs thus selected.¹⁴

SORTING THE SAMPLE FRAME. The selection procedure described above will almost certainly result in a sample of households that conserves the overall characteristics of the sample frame. In other words, the proportion of urban households

14. This method is known as "systematic sampling with PPS". Alternative methods for PPS selection are possible but seldom used in practice.

in the sample, the distribution of the sample by province, and so forth, will all be statistically similar to those in the general population. However, since the selection is random some slight deviations may occur. For instance, by sheer bad luck the sample may contain a larger proportion of northern households than the sample frame.

There is, however, a simple way of making sure that one particular distributional criterion of the households is reproduced in the sample in the best possible way. All that is needed is to sort the PSUs in the sample frame according to that criterion (north to south, for instance) before the selection.¹⁵ In many cases, the "natural" order of the sample frame -- according to encoding of administrative units -- will be adequate and no further sorting will be necessary.

SEGMENTING LARGE PSUS. The household listing operation becomes too burdensome in PSUs larger than 300 households. This problem is aggravated by the PPS procedure, which tends to bring disproportionately many of the larger PSUs into the sample. One possible solution is just to accept that the household listing operation will be harder and longer than usual in those cases, but if they are very large or if many of them are selected in the sample, it may become necessary to split them into smaller units, called *segments*. This need only be done for the large PSUs actually selected in the sample. Segmentation consists of dividing the area of the PSU into pieces, only one of which is selected in the sample. Segments should have clearly defined boundaries, and a rough estimate of the number of households in each segment should be made, either using recent maps or aerial photographs or by means of a "quick count" of dwellings in the field. The original PSU in the list is replaced by the segments (each with their size measures adding up to the original). Only the segment that is selected would have to be listed.

PLANNING THE FIELD WORK. To distribute the selected PSUs among field teams, their locations should first be plotted on a map of the country. They can then be grouped into regions of approximately equal size while trying to spread the workloads evenly and reduce travel time as much as possible. As a by-product of this process, the optimal locations of the teams' base stations are determined.

The next step is to establish the work schedule for each team, that is, to determine in advance when each PSU will be visited. In the standard LSMS surveys household interviews are conducted throughout a 12-month period. To even out the effects of seasonality, the order in which each team visits the PSUs assigned to it should be random.¹⁶

15. Sorting of the sample frame prior to systematic selection is sometimes referred to as "implicit stratification." This method is simpler and more reliable than forcibly allocating the number of PSUs to be selected to certain categories. The latter approach is prone to subjective decisions that unnecessarily sacrifice the self-weighted character of the sample or its domains. All too often these decisions are undocumented, or the documentation is lost, so that the required corrective weights cannot be used.

16. It is sometimes argued that such a random arrangement is too expensive because it forces the teams to move back and forth across their territories during the year rather than visiting the PSUs in a more orderly fashion. The latter option, however, entails the danger of confusing time and

For the Nepal LSMS, this was done by giving a serial number to each of the 275 PSUs selected in the first sampling stage. The numbers 001 to 275 were given to the PSUs at random. After the 275 PSUs were distributed among the 12 field teams (unevenly in this case, given the differences in accessibility inside the country), a simple sort by the PSU serial number produced a work schedule for each team.

Most programming languages and other software have built-in random number generators, but applying them to assign serial numbers to a group of objects in a random order (a problem technically known as "random permutation", or colloquially as "shuffling") is not as easy as it seems. A short algorithm to produce a random permutation of the first N integers is given in Basic in Figure 4.4. The algorithm can easily be implemented in other languages.

Figure 4.4: An Algorithm to Produce a Random Permutation of the Integers 1 to N.

```

+-----+
|                                             |
|                                             |
|               randomize timer                |
|               input N                        |
|               dim P(N)                       |
|               for I=1 to N                    |
|                   P(I)=I                     |
|                   K=1+int(I*rnd)              |
|                   swap P(I),P(K)              |
|               next                            |
|                                             |
|                                             |
|   The statement "dim P(N)" initializes an array P with N elements. In the |
|   subsequent "for ... next" loop, the array elements are successively given |
|   the values 1, 2, 3, ..., I, ..., N, and element I is interchanged with one |
|   of the elements already present in the array (K), selected at random. The |
|   initial values are given in the "P(I)=I" statement and the interchange is |
|   done with the "swap P(I),P(K)" statement". The statement "K=1+int(I*rnd)" |
|   produces a random integer K, from 1 to I ("rnd" generates a random real |
|   number between 0 and 1 and "int" takes the integer part of a number). |
|                                             |
+-----+

```

Implementation of the Second Sampling Stage

space at the analytical stage. In other words, if all PSUs in an area are visited in the same months, it may be unclear if a certain constant condition is due to seasonality or to some geographic characteristic. An "orderly" arrangement of the PSUs is also unlikely to be more economic in any case because LSMS surveys are devised so that field teams come back to their headquarters between field visits — a feature that will be explained later, in Chapter 5.

HOUSEHOLD LISTING. A list of all dwellings in each selected PSU is needed to determine which dwellings on the list will be visited in the survey. Usually this list will have to be created or updated for the survey, though in some cases it can be borrowed from a census or from another survey. This option should be examined critically, however, to ensure that the existing lists are recent, complete, and have good addresses. In particular, demographic mobility makes it dangerous to use lists that will be more than one or two years old by the time of the actual field work. The standard for completeness is difficult to set, but under-enumeration in the census of five percent would be worrisome and the standard could well be stricter. The information on the list should make it easy to locate the households once they are selected. In areas with a good street address system, addresses may be sufficient. Alternately, grid codes on census maps may be used, or references to landmarks and the name of the household head.

Household listing can be carried out either as a separate field operation conducted in all PSUs before the survey starts or by the survey teams themselves when they first arrive in each PSU. The first option is more expensive but more reliable. The expense is incurred because each locality must be visited twice, once during the listing and then again during the survey. It may also entail some difficulty in locating the selected dwellings during the survey because of the time that will pass between the listing and the survey itself.

Listing as a separate exercise is more reliable than listing as part of field work because staff that are specifically trained and devoted to listing are less likely to bias the sample by excluding the dwellings that are harder to reach. (These dwellings are usually inhabited by poorer households who have arrived in the area recently). The survey teams, working under pressure to start interviewing quickly, are more prone to make mistakes in this regard. Also, with separate listing the dwellings to be surveyed can be randomly selected from lists in a single central location using reliable and uniform procedures.

The two most important characteristics of the list are that all dwellings in each PSU be included on it and that it allows the selected dwellings to be located easily.¹⁷ Some practical guidelines can help attain these objectives:

17. The importance of listing procedures is underscored by the experience of the Côte d'Ivoire LSMS surveys. The mean household size observed in the survey dropped from 8.31 to 6.33 persons between 1985 and 1988. Close investigation of this striking phenomenon suggests that it was probably caused by a change in the listing method (see Coulombe and Demery, 1993 and Demery and Grootaert, 1993.) In 1985 and 1986 shortcut procedures were used, rather than the recommended full listing of the households in reasonably sized PSUs. The implications for policy analysis of the apparently inaccurate sampling in the early years were considerable. Demery and Grootaert, 1993, calculated weights to try to correct for the change in sampling procedures. They then calculated mean consumption, poverty, and a series of other important indicators using the weighted and unweighted data and found substantial differences. For example, the head count estimate of poverty in 1986 fell by 14 percent when corrective weights were applied. The bias differed widely among socioeconomic groups and regions. The time series analysis of poverty was also affected. The unweighted data apparently underestimated the increase in poverty between 1985 and 1987.

- Field work should always start with a cartographic reconnaissance. The maps do not need to be very precise in terms of scale or the locations of the dwellings, but they should show the PSU boundaries and the landmarks used to split it into smaller areas. This helps to organize the daily work of the different enumerators.
- Each enumerator should scan the assigned area in an orderly fashion, striving to keep neighboring dwellings close to each other in the list.
- As a rule of thumb, the time needed to list a PSU can be estimated from a standard daily yield of 80 dwellings per enumerator in urban areas to 50 in rural areas.
- The lists should reflect the proper concepts of dwellings and households. Enumerators should be trained to tell the difference between the two.
- Dwellings should be clearly listed with appropriate addresses so that interviewers can find them easily during the survey. Designers should use some imagination to achieve this goal where street names and house numbers are not well established. In many surveys dwellings are numbered as a part of the listing operation, either by affixing a numbered sticker to the outside of the home or by painting a number on the wall or door. At the time this is being written (July 1995) the possibility of using Global Positioning Systems (GPSs) to support field work of future LSMSs is being considered. GPSs are battery-operated devices the size of a pocket calculator, currently commercially available for about \$500, that use satellite signals to pinpoint the user's position with remarkable accuracy (within 10 meters or so in the three dimensions: latitude, longitude and altitude). Enumerators could use GPSs to record the dwelling locations during the listing operation; interviewers would use them later to locate those selected for the sample.
- The complete list should always be recorded in a standard form with one line per dwelling. The list can be several pages long, depending on the size of the PSU and the number of enumerators engaged in the operation. Though the precise layout of such a form depends on local conditions, a typical list form is shown in Figure 4.5.
- Supervision of the listing operation is crucial. Listers have an obvious incentive not to be too diligent in locating hard-to-find or remote dwellings. Since there is no criterion to tell how diligent they are being that can be easily monitored in the office, the field supervision will be key. Supervisory staff (or other listers) must re-visit a subset of listed areas, especially the difficult parts of them, to verify the listing.¹⁸ An option that might be feasible in some settings is to use lists from

18. It can be especially useful to do this around dusk, when lights or smoke from cooking fires may help locate dwellings. Carrying binoculars may be useful for finding dwellings across ravines, or down roads marked no trespassing.

other sources to help in this process. For example, if the PSUs can be identified with electoral areas, voting lists might be used. Although not every resident of the PSU will be on the voting list, any address on the voting list should be listed in the PSU.

Figure 4.5: Typical Listing Form

```

+-----+
Region: _____ Province: _____ Locality: _____ PSU Code: |  |
+-----+
Date of the listing: _____ Enumerator: _____ Page: | / |
+-----+

+-----++-----++-----++-----++-----++-----++-----++-----++-----++
|Serial||Address of the dwelling |Head of the household      |+- Household size |
|Number||                      |                          || M | F |total|
+-----++-----++-----++-----++-----++-----++-----++-----++
|  01  ||                      |                          ||   |   |   |   |
+-----++-----++-----++-----++-----++-----++-----++-----++
|  02  ||                      |                          ||   |   |   |   |
+-----++-----++-----++-----++-----++-----++-----++-----++
|  ..  ||                      |                          || ... | ... | ... |
+-----++-----++-----++-----++-----++-----++-----++-----++
|  nn  ||                      |                          ||   |   |   |   |
+-----++-----++-----++-----++-----++-----++-----++-----++

```

Columns may be added to this model for key landmarks, the occupations of the head of the household, or whatever other information could help in finding the dwelling. It may also be useful to have the enumerators fill in separate lines for buildings that are not dwellings, such as shops and offices; in that case, a special check column should be added so that the real dwellings can be told from the other buildings. However, only essential information needed to identify the dwelling should be recorded. Including too much data slows the field process and risks shifting the enumerators' interest from listing to interviewing.

So far this discussion has assumed that the maps from the most recent census are available, so that the listing focusses on updating the listing of dwellings within well defined boundaries. In fact, it is often the case that some, or even all, of the maps have been lost in the intervening years.¹⁹ In such cases, it is sometimes possible to reconstruct the maps. This would happen, for example, when only the occasional map has been lost and other maps for the contiguous sampling units still exist.

Another means of reconstructing the maps may be possible when the sampling units correspond to some administrative unit that the populace or officials will recognize. This is often the case, especially in rural areas. For example, a sampling unit might correspond to a ward or village. In this case there is a special detail to watch for. Say

19. The sampling chapter in Delaine *et al.*, treats the allied problem of what to do when the boundaries were poorly defined in the original maps.

that PSU 348 was labeled Alama, which is the name of the ward that it corresponds to.

It would seem a straightforward matter to send listers to the ward of Alama and have them establish the boundaries of the ward and start listing. But Alama may have grown a good deal in the several years since the census and the area subdivided into new wards. The central area will still be called Alama, but the new wards will have other names, say Bendicion, Caceres, Durango and Esperanza. In this case if the lister goes to Alama and asks where its boundaries are, s/he will be told about the new boundaries that cover only a fraction of the area of the original Alama. All the area covered by Bendicion, Caceres, Durango and Esperanza would be omitted and would not be listed.

The population of these areas would effectively be excluded from the sample. The solution to these problems lies in trying to verify from the appropriate authorities (the ministry of local government, ward officials, etc.) whether the boundaries and names have been constant since the last mapping. This should be done both by the statistical agency's central office for the country as a whole, and verified by individual listers.

ADJUSTING FOR DIFFERENCES IN PSU SIZE. Differences are sure to be found between the "census" size of each PSU (the size that was used for PPS selection in the first sampling stage) and the "observed" size (from the listing operation). For instance, the listing operation of the Nepal LSMS --conducted in mid-1994, two years after the 1992 census-- showed that XXX.

These differences are partly due to imperfections in the census and partly due to demographic mobility. Whatever the reason, the differences alter the self-weighted character of the sample in each analytical domain, which makes it necessary to correct the sampling weights in order to obtain unbiased point estimates from the survey. Assuming for simplicity that the number of households was used as a measure of PSU size in the first sampling stage, and calling C_i and O_i the census and observed number of households in PSU i (belonging to weighted domain k) the expansion factor w_i for the households in that PSU should be

$$w_i = w_k \frac{O_i}{C_i}$$

where $w_k = N_k/n_k$ is the basic sampling weight of domain k , defined before (see Section *Determining the Basic Sample Design Parameters*). The formula would be slightly different if some other measure of size (such as the population or the number of dwellings) had been used in the first sampling stage.

It goes without saying that the complete list of weights w_i for all PSUs (and better still, the list of all C_i 's and O_i 's) should be carefully kept and made available to analysts as a part of the survey documentation and data sets.

SELECTING DWELLINGS. The dwellings to be visited are selected by systematic sampling from the PSU listings. A few extra dwellings are also selected to be used if replacements are needed in the field.

The selection procedure, though generally well known to statistical officers everywhere, is illustrated below in Figure 4.6. This example assumes that 16 dwellings

are to be interviewed and that 4 extra dwellings are to be selected in each PSU as replacements. The exercise is to select those 20 dwellings, based on information contained on a typical listing form such as that shown in Figure 4.5.

First, count the total number of dwellings in the PSU and record it in the space on top of the form. Assume, for example, that there are 86 dwellings in the PSU.

Second, divide the total number of dwellings by the number of dwellings to be selected and keep the first decimal place. The result is called the sampling interval (SI) and is also recorded on top of the form. In this example, if the number of dwellings to be selected is 20, SI would be 4.3 (because $86 / 20 = 4.3$).

Third, select a one-decimal random number less than the sampling interval (in the example, this would be a number from 0.0 to 4.2; it can be obtained by selecting a random integer from 00 to 42 and inserting a decimal point before the last digit). Add 1 to that random number. The result is called the "random start" (RS) and is also recorded on top of the form. Assume, for instance, that RS turns out to be 3.2. Write the 20 numbers obtained by starting with RS and repeatedly adding SI. With the above values of RS and SI, the 20 numbers would be:

3.2	$20.4 + 4.3 = 24.7$	$41.9 + 4.3 = 46.2$	$63.4 + 4.3 = 67.7$
$3.2 + 4.3 = 7.5$	$24.7 + 4.3 = 29.0$	$46.2 + 4.3 = 50.5$	$67.7 + 4.3 = 72.0$
$7.5 + 4.3 = 11.8$	$29.0 + 4.3 = 33.3$	$50.5 + 4.3 = 54.8$	$72.0 + 4.3 = 76.3$
$11.8 + 4.3 = 16.1$	$33.3 + 4.3 = 37.6$	$54.8 + 4.3 = 59.1$	$76.3 + 4.3 = 80.6$
$16.1 + 4.3 = 20.4$	$37.6 + 4.3 = 41.9$	$59.1 + 4.3 = 63.4$	$80.6 + 4.3 = 84.9$

Finally, take the integer part of each number. The 20 numbers obtained in this way (3, 7, 11, 16, 20, 24, 29, 33, 37, 41, 46, 50, 54, 59, 63, 67, 72, 76, 80 and 84), are the sequence numbers of the dwellings to be visited in the survey. The corresponding lines in the listing should be transferred to another form, called the *List of Selected Dwellings* (see Figure 4.6).

The households to be visited during the survey are those listed on the sixteen unshaded lines in the form. The dwellings on the shaded lines are kept as reserve for possible replacements.

Both the full listing form with all dwellings and the list of selected dwellings will be needed by the field team responsible for the PSU during the survey (the former will help them locate the selected dwellings in the field, by referring to their neighbors). As this operational requirement entails the risk of losing these documents, it is highly recommended to provide the field teams with photocopies, and to file the original lists securely for at least five years. The lists constitute a precious material for central supervision, and may even be required long after the end of the original project, for panel or follow-up surveys, or even as base material for different surveys conducted by the statistical agency.

REPLACING HOUSEHOLDS. The above selection procedure implicitly assumes that it may be impossible to interview the households in some of the selected dwellings and that a standard procedure for replacing them has to be implemented. The most frequent reasons for replacement are:

The dwelling is unoccupied and is likely to remain unoccupied for the full survey period.

The dwelling has disappeared or is not being used for housing.

The dwelling cannot be located because the information in the listing is bad or insufficient (for example, illegible names or addresses).

The household refuses to be interviewed.

These cases should be carefully studied by the team supervisor. Only when the supervisor is convinced that the interview is impossible should the dwelling be replaced with the one on the nearest shaded line in the form.²⁰

If the dwelling is occupied by a household different from the one recorded during the listing operation, the new household is interviewed without more ado. As said before, the LSMS samples are actually samples of dwellings, and such cases should not be counted as non-responses.

20. Notice that the shaded lines in the form are evenly interspaced between the unshaded lines. The idea is to replace households by near neighbors, which are likely to have similar socioeconomic characteristics. Shading every fifth line allows for replacement of up to 4 out of the 16 dwellings selected (a 25 percent non-response rate). A smaller proportion of replacements could be insufficient in certain worst-case PSUs; shading a much larger proportion of lines could be interpreted by some field supervisors as an invitation to replace with abandon.

Selecting Random Persons in a Household¹

To reduce interview time, the LSMS questionnaire is sometimes designed so that certain modules are applied to one randomly selected person in the household. The Côte d'Ivoire LSMS, for instance, collected fertility information from one woman 15 years or older.

As opposed to the other random selections described so far, which are most reliably carried out at central offices, the choice of a person at random in each household must be performed by the interviewer in the field. A simple procedure must be devised for this that gives each eligible person the same chance of selection and is verifiable so that the work of the interviewer can be tested for accuracy (the latter precludes the use of dice or other "truly random" methods).

Instead of the traditional Kish tables (Kish, 1965), LSMS surveys have opted for an original, alternative method.²² As explained in the chapter on questionnaire design, each household member is assigned an identification code, generally from 01 to 20, in the household roster of the questionnaire. An adhesive label with a different random permutation of these numbers is affixed to each questionnaire. To select the person, the interviewer scans the list of identification codes on the label until the code of an individual who meets the defined eligibility criteria is reached. Figure 4.7 shows one of these labels.

Figure 4.7: Sticker Used for Selecting a Random Individual Within the Household

```

+-----+
|       |
| 03 06 07 08 11 12 10 17 04 02 |
|       |
| 16 15 05 18 19 01 13 20 09 14 |
|       |
+-----+

```

The procedure is simple but requires careful training of the interviewers. They should scan the list of ID codes one line after the other, always from left to right, crossing out all the numbers that are rejected and circling the number of the first person who qualifies. This was not made clear at the beginning in Côte d'Ivoire, where at least one of the interviewers always searched for code 02 (usually the head's wife), and circled it without considering other women's IDs.

21. This section has been adapted from LSMS Working Paper No. 26, *The Côte d'Ivoire Living Standards Survey: Design and Implementation*, by Martha Ainsworth and Juan Muñoz (1986).

22. Kish tables do not always give exactly the same chance of selection to every eligible individual. A more serious disadvantage of Kish tables is that they require the eligible individuals to be given a serial number prior to the selection, in addition to their standard ID codes. The co-existence of two different numbering systems for the same person is potentially confusing to the interviewer.

The process is verifiable by the supervisor, who can repeat the procedure with the label stuck to each questionnaire. It can also be checked by the data entry program.

The labels for all questionnaires in a survey can be quickly generated with a personal computer. A complete program is not given here because it needs to be adapted to specific circumstances as well as to the number of ID codes needed. The production of a different random permutation for each sticker is done with the algorithm presented in Figure 4.4.